Seeing is Believing: Interpreting Behavioral Changes in Audio Deepfake Detectors Arising from Data Augmentation

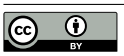
Boo Fullwood
Georgia Institute of Technology
Atlanta, USA
boo@gatech.edu

Fabian Monrose
Georgia Institute of Technology
Atlanta, USA
fabian@ece.gatech.edu

ACM Reference Format:
Boo Fullwood and Fabian Monrose. 2025. Seeing is Believing: Interpreting Behavioral Changes in Audio Deepfake Detectors Arising from Data Augmentation. In *Proceedings of the 2025 Workshop on Artificial Intelligence and Security (AISec '25), October 13–17, 2025, Taipei, Taiwan.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3733799.3762979

Today, creating audio deepfakes is easier than ever, and the proliferation of high-fidelity text-to-speech and voice conversion tools has underscored the need for technologies that can quickly differentiate between real and spoofed audio. In lieu of such techniques, spoofed audio poses a serious epistemic threat to public trust in audio validity. The push for more competent detectors to combat this threat has led to the adoption of increasingly powerful detector architectures and the development of novel data augmentation techniques to better adapt to out-of-distribution data.

However, these architectures generally lack the interpretability of simpler models, preventing researchers and end-users from fully understanding model behavior and the impact of augmentation. To address this, we demonstrate an occlusion-based explainability analysis technique, enabling the identification of specific changes in model behavior induced by data augmentation. We show that these differences can be identified even between highly similar augmentations and observe that common augmentation techniques, namely random input masking, produce counter-intuitive and potentially undesirable behavioral characteristics and may fail to improve model robustness. We further demonstrate the utility of behavior visualization by identifying undesirable behavior in response to encoded audio and developing a corresponding augmentation that recovers a majority (51.6%) of the lost performance. The developed augmentation shows higher generalization across other classes of distorted audio on our model than the "general purpose" augmentations. Our explainability technique is enabled by the use of a spectrogram-based detector. We specifically select the Audio Spectrogram Transformer, which has seen limited use in the field compared to similar, less explainable alternatives. Alongside improved explainability, AST shows performance matching or exceeding the existing state-of-the-art (Equal Error Rate 0.001) on large deepfake datasets.



AISec '25, Taipei, Taiwan

ACM ISBN 979-8-4007-1895-3/25/10
https://doi.org/10.1145/3733799.3762979

1 Introduction

The widespread deployment and adoption of generative AI systems offer significant benefits and pose significant risks to society. On the latter, fake content is now everywhere. Creating realistic synthesized audio (colloquially referred to as audio deepfakes) [36] no longer requires expert knowledge or access to large training datasets, with some systems claiming to offer "zero-shot cross-lingual voice cloning", free of charge, in minutes [52]. Voice cloning tools like these are not only readily available [16], but also easy to use [43], and their ubiquity is fueling an unprecedented wave of malicious incidents. While some of the more damaging incidents have been reported in the popular press (e.g., fraudsters cloning the voice of a company director to dupe a bank manager into authorizing transfers of $35 million)[1], Hutiri et al. [25] recently detailed over 35 incidents where specific harms (including coercion, deception, and laundering) were traced back to speech generation tasks. The startling realism offered by these tools has raised concerns about the looming challenges to democracy and national security, leading to executive orders [5] in the United States on the trustworthy development and use of Artificial Intelligence, and the adoption of laws governing the creation and use of deepfakes [17] worldwide.

Audio deepfakes often fool listeners because humans rely more on visual information than any other form of sensory information [51], making it difficult to notice indicators of forgery in isolation. In the absence of widespread deployment of automated detectors in popular online platforms, observers are left to make judgments about whether the information they encounter is real or fake based on mental shortcuts for establishing veracity [24, 44] — but, all too often, they cannot do this well. For instance, with video deepfakes, visual inconsistencies (e.g., incorrect lip-syncing or unnatural facial expressions) [16] are commonly relied upon as cues for determining content veracity, but similar perceptual cues [12, 34] in audio-only media are less clear [33, 44]. Indeed, several studies [33, 68] have shown that humans are highly susceptible to deceptive audio and that we are approaching the point where our perceptual capabilities no longer offer an effective defense against AI-generated forgeries. Today, many scholars consider deepfakes to be epistemically harmful [22, 23] because they undermine trust in our senses, inevitably leading people down a path of insulating themselves against the threat of deception by restricting the channels where they access information.

The expectation from the security community is that Audio Deepfake Detection (ADD) technologies will address this impending threat by picking up on subtle artifacts of forgery that are imperceptible, or at least not typically perceived, by human observers. Along

[1] See *"Fraudsters Clone Company Director's Voice in $3.5 Million Heist, Police Find*, Forbes Magazine, Oct 14, 2021.

those lines, recent research [1, 72] has demonstrated the effectiveness of a variety of detector architectures on closed-world detection tasks. These detectors report near-perfect classification [60, 67] of spoofed audio within a single target dataset. However, because real-world deepfakes can span a breadth of generator architectures, spoof types, and acquisition environments [72], the same detectors have been shown to perform poorly when exposed to new information without retraining [15, 46].

In response, data augmentation is widely used [29, 30, 40, 49] to help improve performance on out-of-distribution data, typically by perturbing training inputs. While several of these augmentations are well justified in theory, oftentimes, the actual impact on model behavior is not systematically investigated because doing so requires a model that allows for the detailed localization of attention or otherwise determining regions of importance in the input sample. Ideally, a model will allow localization of model attention both in terms of time (*what points during the audio are important?*) and frequency (*at those points, what frequency ranges are important?*). Unfortunately, many detectors operate on raw audio [60] or on intermediate representations [61] that only permit interpretability along the time axis, making it difficult to identify what characteristics [58] of the audio at the identified time points are important for classification.

To address those needs, this paper presents a reliable visualization of model behavior, projected onto a spectrogram [54] representation of the input audio, along with a novel quantification of model attention distribution. This approach is enabled by the Audio Spectrogram Transformer (AST), which achieves state-of-the-art performance while maintaining the explainability of both time and frequency features, allowing for detailed localization of model attention.

Our contributions include:

- We demonstrate the visualization and identification of changes in model behavior in response to several common augmentation techniques [49] (e.g., time masking, frequency masking), which is not possible with competing architectures.
- We showcase the development of a bespoke augmentation informed by behavior visualization, dramatically improving performance on audio encoded with high-quality audio codecs (G.711, OPUS), and attaining modest performance gains on other common classes of audio distortion.
- The application of the AST architecture to the domain of audio deepfake detection and evaluation on three large datasets, including spoofed audio from 27 high-fidelity generators and collections of real-world deepfakes, demonstrating state-of-the-art performance (an equal error rate of 0.001) matching or exceeding widely adopted detectors (Wav2Vec2.0 [2], Whisper [53], RawNet2 [60]).

The remainder of the paper is structured as follows. Section 2 offers essential background information, including key terminology and concepts fundamental to our approach and analyses. Section 3 reviews related work, followed by a detailed explanation of our approach for behavioral visualization in Section 4. In Section 5, we investigate the behavioral impact of common spectrogram augmentations. In Section 6, we investigate how interpretability can be used to develop and validate novel augmentations and evaluate the developed augmentations against several important types of audio distortion. and the description of our experimental setup in Section 7. Section 7 discusses our performance evaluations and results under a standardized setting. We discuss limitations and future work in Section 8 and conclude in Section 9.

2 Relevant Background

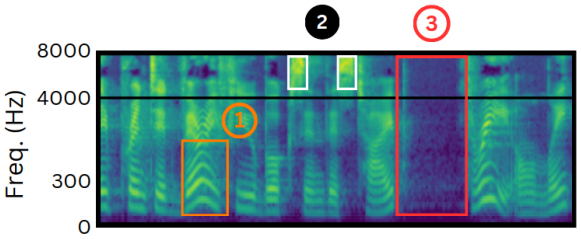


Figure 1: A spectrogram with quasi-periodic signal ①, aperiodic signal ② and silence ③ highlighted. Note the logarithmic scale.

The spectrogram representation of speech allows for changes in frequency properties of the speech signal to be observed over time. Traditionally, a spectrogram is created by separating the source signal into discrete time windows, performing a Fourier transform on each window, and concatenating the resulting transformations to produce a time-domain representation. The windowing process is necessary given that a typical speech signal is not stationary, that is, the intensity of the different frequency components are expected to vary over the duration of the sample. However, most speech signals are approximately stationary over a short duration.

For speech signals, the frequency scale is often converted to a Mel-scale [62], a logarithmic scale corresponding to human-perceived frequency distances, such that a doubling in Mel frequency corresponds to a doubling in perceived frequency. This accounts for the reduction in perceived frequency distance at higher frequencies, e.g. a change from 400Hz to 800Hz is perceived to be larger than a change from 3600Hz to 4000Hz.

Broadly, speech signals can be classified as voiced (active speech when the vocal cords are vibrating), unvoiced (active speech with no vocal chord vibration), and silent. An example Mel-spectrogram is shown in Figure 1, with highlighted examples of characteristics of voiced and unvoiced speech. In this paper, we rely heavily on the spectrogram representation for modeling and visualizing model attention. We refer interested readers to the seminal textbook by Reetz and Jongman [54] on acoustic production and perception.

3 Related Work

Audio deepfake detection is an active research topic [7, 30, 38, 40, 46, 58, 60] with numerous surveys [1, 31, 37, 72] dedicated to the subject alone. Yi et al. [72], for example, provides a comprehensive look at differences across types of deepfake audio, widely used features, datasets, and evaluations of state-of-the-art architectures. As might be expected, the application of transformer architectures to deepfake detection is not new. In particular, in the domain of audio deepfakes, the Wav2Vec2.0 architecture [2] achieved strong

classification performance (i.e., equal error rate of < 0.04) [38, 61] on a popular challenge dataset. Our use of an Audio Spectrogram Transformer for the problem of deepfake detection achieves performance on par with, or exceeding, previous approaches, and more importantly, operates on a more interpretable intermediate representation. Existing applications of AST to audio deepfake detection focus on alternative training paradigms, namely online/continuous learning [35] and contrastive learning [20], and neglect the explainability utility of the model.

To improve generalizability [46], architecture selection along with data augmentation techniques [38, 61, 65] are commonly employed to boost model performance on both intra-dataset evaluations and out-of-distribution data. These approaches employ global noise addition [38, 61], feature masking/warping [49], encoding and compression of the source audio [38], and ablation-based entropy addition [65]. However, unlike the direction we take, these solutions only evaluate augmentations in terms of the change in model detection performance. While proposed augmentations are often offered alongside reasonable assumptions of their impact on model behavior, this impact is not validated, and a performative augmentation is assumed to be working as designed. We evaluate all augmentations both on performance and on validated behavior.

Alongside augmentation, recent work on model explainability [3, 6, 73] has sought to offer insights into discriminatory features for audio deepfake detection. However, the focus has been on analyzing the behavior of static models. We expand this analysis by showing changes in behavior as a result of augmentation and extend the analysis of model explanations beyond basic region-feature correspondence.

Recently, Maltby et al. [40] observed that spectral differences between real and spoofed speech were increased in high-frequency regions, and showed that increasing the intensity of spectral features in these regions can improve detector performance. Unfortunately, as we show later, these are the same regions that are less likely to be preserved through normal audio transformations, like transmission over VoIP. Likewise, Kawa et al. [30] show that the performance of spectrogram-based detectors improved when using spectrograms generated by the Whisper [53] feature encoder. That encoder uses uneven, learned frequency bin intervals to optimize classification performance on the Whisper transformer architecture. While it is conceivable that incorporating Whisper spectrograms into our framework could be beneficial, these adjusted spectrograms are incompatible with the pretrained AST features.

Lastly, given the current interest in deepfakes, it is not surprising that a number of adversarial attacks against audio deepfake detectors and speaker verification systems [8, 29, 66] have emerged. Kassis and Hengartner [29], Chen et al. [8], and Wang et al. [66] present optimization-based adversarial attacks and show that these attacks can circumvent most or all tested speaker verification systems. All of these attacks require the ability to repeatedly test adversarial samples against the detector in question, but can produce distortion-minimal modifications to the audio to produce misclassifications. In this paper, we focus on transforms that are likely to be applied in the general course of recording, storage, and transmission, which are likely to be encountered by all deepfakes. However, for completeness, we demonstrate that a black box adversarial attack [9] remains effective against the augmented model, though the number of optimization steps needed to produce satisfactory adversarial examples increases substantially. Further details are given in Appendix 10.2.

4 Approach

Our approach for empowering model designers with the information needed to understand behavioral changes introduced by data augmentation strategies takes advantage of the Audio Spectrogram Transformer (AST) [21] that was originally designed for the task of general audio event classification (e.g., distinguishing between speech commands, environmental sounds, music, animal sounds, etc.). The spectrogram representation used by AST is a natural choice for audio classification as it preserves both temporal and frequency features. More importantly, this architecture allows for easy visualization of model attention compared to convolutional feature extractors used by other transformer models such as Wav2Vec [2]. Convolutional feature extractors produce an intermediate representation that is not visually related to any physical qualities of the source audio, complicating any visualization of model attention projected onto them. This additional interpretability is independent of the detection capabilities of the model.

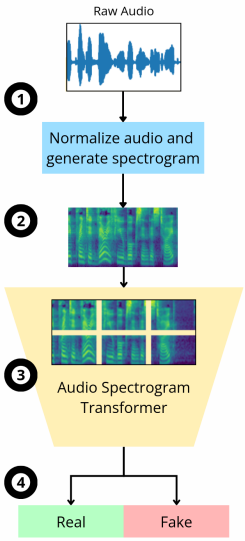


Figure 2: Model pipeline during standard detection tasks. Steps ① and ② preprocess audio and produce the input Mel spectrogram. The fine-tuned AST model performs inference in step ③ and reports the predicted class (step ④) along with confidence scores for each class.

For audio deepfake detection we use a traditional pipeline among transformer architectures, as shown in Figure 2. In step ①, the input audio is normalized to match the sample rate and amplitude of AST's pretraining data to ensure compatibility with pretrained features. The normalized audio is then converted to a Mel-spectrogram (in step ②). Generated spectrograms are padded or truncated to a length of 1024 time bins, corresponding to an audio length of about 10s. The internal Vision Transformer (ViT) [13] model (step ③) splits the spectrogram into smaller patches that are then flattened

and dimensionally reduced through linear projection. The resulting embeddings are combined with positional tokens to encode each patch's location within the larger spectrogram. The embeddings then pass through a standard transformer encoder [63] architecture that computes the attention between each embedding and all other embeddings. The attention maps that govern this process are the primary trainable parameters of the encoder architecture. The attention values are then summed with the original embedding, and the resulting feature vectors are passed to a dense classifier. The model outputs classifications (step ④) as a confidence probability for each class.

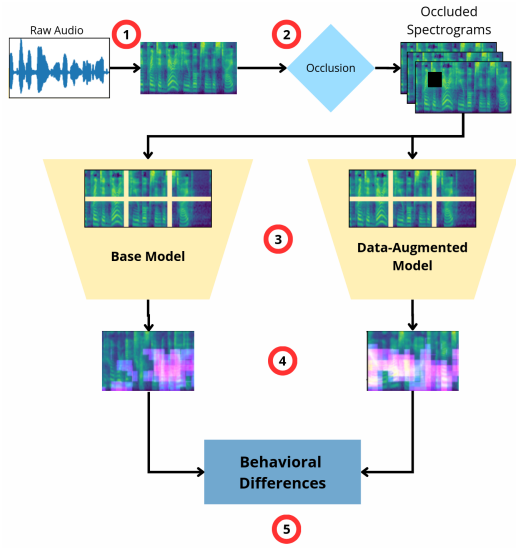


Figure 3: Model pipeline during behavioral analysis. Step ① encompasses preprocessing and spectrogram generation. Each spectrogram is repeatedly occluded (step ②) and then run through the detection pipeline for the base model and the model trained on augmented data (step ③). The confidences for the occluded spectrograms are aggregated into heatmaps (step ④) and then compared to find behavioral changes (step ⑤).

When performing behavioral analysis, we incorporate several steps that allow localization of model attention, as shown in Figure 3. The initial spectrogram is generated in the same way as for normal inference (step ①). A sliding window is then used to mask out small regions of the spectrogram (shown as a black box in step ②). Each occluded spectrogram is then classified by both the base model and the data-augmented model (step ③) and the change in model confidence (relative to the original spectrogram) is recorded. The confidence differential is mapped onto the occluded region of the spectrogram, producing a heatmap (step ④) of attention for each model. These heatmaps are then compared (step ⑤) across aggregated groups to identify changes in behavior induced by the augmentation. We provide a detailed discussion of the visualization and comparison processes in Section 5. Additionally, as the relevance of the visualizations and statistics produced by our approach are partially dependent on the model's classification performance to ensure that changes in model predictions align well with the underlying data. We address the performance capabilities of AST in Section 7.

It is important to note that AST is pretrained for audio classification tasks from a large dataset, called AudioSet, containing 2.1 million audio samples across 527 classes. Of these, 1 million samples contain speech. Pretraining allows high-parameter models to be used without access to the large datasets needed to train them from a random initialization. Leveraging this pretrained model, however, requires that parameter differences between fine-tuning data and pretraining data be minimized. Thus, all data are resampled to a 16000Hz sampling rate and normalized to the same mean/variance as the AudioSet pretraining data.

5 Model Augmentation and Visualization of Model Behavior

A seminal work in the area of augmentation techniques for automatic speech recognition is that of Park et al. [49]. In that work, three augmentation techniques are suggested, namely time masking, frequency masking, and time warping. Time and frequency masking are conceptually simple, removing information in a random range along the specified axis (time or frequency) of the spectrogram. Time warping dilates or contracts the spectrogram along the time axis about a randomly selected point. The transform maintains the overall dimensions of the spectrogram but alters the relative scale of the features. The augmentations were initially developed to improve automatic speech recognition on two datasets, LibriSpeech 960H [48] and Switchboard 300H [19]. The two augmentation policies for LibriSpeech are denoted as LibriSpeech Basic (LB) and LibriSpeech Double (LD), with LD masking twice as many segments as LB. The Switchboard augmentations, Switchboard Mild (SM) and Switchboard Strong (SS) both mask two frequency and two time segments, but vary in the width of the masked frequency segments with SS allowing larger masked regions. We do not implement time warping as Park et al. [49] concluded that the performance improvement associated with time warping was small relative to the other augmentations. In fact, models we trained with warped spectrograms showed subpar performance compared to other policies. The parameters for these augmentations are shown in Table 1 and correspond to the policies recommended in the original paper.

Policy	F	m_F	T	p	m_T
None	0	-	0	-	-
LibriSpeech Basic (LB)	27	1	100	1.0	1
LibriSpeech Double (LD)	27	2	100	1.0	2
Switchboard Mild (SM)	15	2	70	0.2	2
Switchboard Strong (SS)	27	2	70	0.2	2

Table 1: Parameters for SpecAugment augmentations, controlling the count (m_F, m_T) and width (F, T) for time and frequency masks, resp. The parameter p sets a limit on the width of time masking segments proportional to the length of the audio sample.

The parameters m_T and m_F determine the number of masked sections for time and frequency, respectively. Similarly, the width

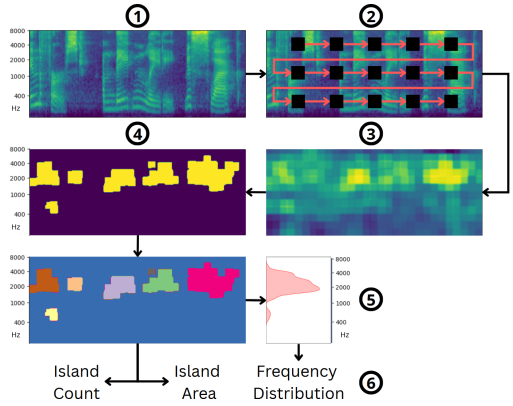


Figure 4: Diagram of Explainability Pipeline. The Mel Spectrogram ① is repeatedly occluded ② to generate an importance heatmap ③. This heatmap is thresholded and segmented ④ to identify separate islands of high attention ⑤. Island statistics and distribution ⑥ are computed and compared between models.

of each masked section is selected uniformly from the range $[0, F)$ and $[0, T)$. Time masking has a unique parameter p which caps the maximum width of a masked time section to $p \cdot t_{max}$ where t_{max} is the number of time steps in the spectrogram. Thus, for a policy with a p value of 0.2, no masked time segment can be wider than 20% of the total spectrogram duration. This prevents over-masking of short-duration samples.

5.1 Visualizing Model Behavior

The premise behind occlusion-based analysis is simple: if a region of the input is important to the model's classification process, removing (occluding) the information in that region should reduce the model's prediction confidence. This reduction in confidence is then assumed to be proportional to the importance of the region. By systematically occluding all regions of the input, we produce a map of input importance. This process is shown in Figure 4 (steps ① - ③). This method of visualization does require care in the selection of parameters to ensure that the results are both robust and meaningful.

First, the size of each occluded area must be selected to provide a balance between heatmap resolution (smaller window is desirable), confidence change magnitude (larger window is desirable), and computational cost (larger window is desirable). We select a window of 21x21, primarily driven by the desired change in prediction confidence. Smaller windows caused minimal impact, indicating that the window was not fully occluding features of interest. Second, we select a baseline value (that is, the value with which occluded regions are replaced) of -1.27…, which corresponds to minimal spectral energy in our normalized spectrograms. The de facto value of 0 would instead be interpreted as average spectral energy, which we found induced spurious attention in silence regions where it was interpreted as new information.

We evaluated several alternatives to occlusion, including Grad-CAM [57], Transformer Input Sampling (TIS) [14], and LIME [56]. We found that the gradient-based techniques (Grad-CAM, TIS) worked poorly in the pretraining/fine-tuning ecosystem as the transformer weights that are used by these methods are mostly stationary during the fine-tuning process, and are therefore mostly determined by the original pretraining. While this is a desirable characteristic for applying large models to small datasets, these techniques fail to capture the domain-specific learning that happens predominantly in the final classification head. LIME, which is also occlusion-based, segments the input image prior to occlusion, and occludes each segmented region. The default quick shift segmentation did not produce meaningful regions when applied to spectrogram images (as opposed to traditional, pictorial images), and spectrogram-specific segmentation approaches resulted in coarse segmentations, which impacted the usability of the resulting heatmaps. Finally, our approach avoids recent concerns over complications with interpreting explanations generated by attention rollout and token masking approaches [27, 69] and ensures that the explanations are as interpretable as is possible.

5.2 Identifying Changes in Behavior

Given a base model and a model trained on augmented data, we wish to identify shifts in model behavior when given the same inputs. By applying the occlusion method to each model and input, we produce pairs of attention heatmaps that highlight regions of importance for each model. While comparing single heatmaps can be illustrative, it is necessary to aggregate many heatmaps to identify larger trends in model behavior.

To identify changes in the overall distribution of attention, we first identify the number and size of regions of high attention using an island-finding algorithm as shown in Figure 4. While this is not a new method of region mapping, its application in this space is unique and enables robust aggregation of behavior patterns across samples. Each heatmap is normalized between 0 and 1 and then thresholded such that areas of high attention are preserved and all other regions are set to 0 (Step ④). We define an island as a contiguous region of non-zero attention in the heatmap. We iterate over each point in the heatmap and initiate a breadth-first search from each unvisited, non-zero point. The breadth-first search aggregates all surrounding non-zero points, returning a list of points belonging to the current island (Step ⑤).

When comparing heatmaps, we can then look at both the total count of islands and the average area of islands for each model (Step ⑥). Together, these provide a good measure of attention uniformity. We apply either a Student's t-test or a Welch's t-test as appropriate based on population variance to determine whether there is a significant difference in the distribution of attention regions between the base model and each data-augmented model. We use a threshold of $p < 0.05$ for each T-test.

In addition to the count and area of islands in a heatmap, we compute the centroid of each island and compare the distribution of islands along the frequency axis (Step ⑥). This is motivated by two key characteristics of speech in general, and deepfake speech in particular. First, lower frequency components of speech are more important for intelligibility [54] and tolerate less distortion before perceived speech quality is impacted. As such, many encoding schemes, including the G.711 codec discussed in Section 6, do not

Augmentation	Island Count	Island Area
None	4.66	23.24
LB	5.44	44.95
LD	3.83	17.86
SM	6.38	72.26
SS	4.83	18.72

Table 2: Island finding statistics for each augmentation, and hypothesis test results compared to baseline model. Count indicates the number of discrete islands, while area indicates the average area of islands. Non-statistically different values are grayed out. The baseline (unaugmented) values are highlighted in yellow.

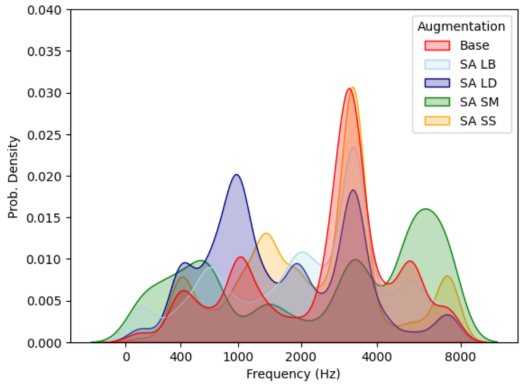


Figure 5: Probability distribution function of islands over the frequency axis for each augmentation. The distribution is calculated by computing the centroid of each island, and then computing a histogram of the frequency coordinate of each centroid. The PDF is then calculated using a kernel density estimator.

preserve high-frequency information, instead prioritizing maintaining as much low-frequency information as possible for a given bandwidth. Therefore, a robust detector must necessarily be able to operate in the absence of high-frequency components. Secondly, Maltby et al. [40] determined that differences between real and deepfake speech are greater in high-frequency regions. This introduces a conflicting desire with our first point: low-frequency attention may be more robust against manipulation, but high-frequency attention may capture more usable features for classification.

5.3 Results

The statistics for the distribution of attention are shown in Table 2. We can see that all augmentations induced significant changes in model behavior, as indicated by the diverging distribution of attention islands. Counterintuitively, the SpecAugment-derived augmentations do not produce similar changes in behavior. Given that these augmentations apply the same type of input perturbation and vary only in the scale of the perturbations, we would expect that the induced behavior change would be similar in type if not in magnitude. However, we see that while both the *LD* and *SS* policies produced similar reductions in average island size, the *LB* and *SM* policies resulted in a dramatic increase in island size. Similarly, the *LB* and *SM* policies produced a greater number of high-attention regions which, when combined with the larger average island size, indicates a significantly more broad distribution of attention compared to the other augmentation types. This behavior is more in line with the expected impact of masking augmentations which punish dependence on small artifacts or highly specific regions. The overall reduction in attention area associated with the other SpecAugment policies likely contradicts the behavioral change expected by most users. Based on these statistics, the SpecAugment augmentations broadly sort into two groups based on behavior (*LD* and *SS* reducing attention diffusion and *LB* and *SM* increasing diffusion). However, these groups do not fall along either the original dataset separation (LibriSpeech vs. Switchboard) or parameter differences (LB having only one masked segment per axis) as might be expected. This contradicts the intuitive assumption that similarity in augmentation mechanism or magnitude implies similarity in augmentation impact.

Figure 5 shows the distribution of islands along the frequency axis for each augmentation. The most obvious feature across augmentations is a strong concentration of attention at 3600Hz. All SpecAugment augmentations, except for SM, and the base model exhibit this behavior. Interestingly, this peak is at the same frequency identified by Maltby et al. [40] as an inflection point between well-reproduced (low frequency) and poorly reproduced (high-frequency) deepfake speech. Frequency trends between LB and SS augmentations are highly similar to the behavior of the base model, with a strong attention peak at 3600Hz. The LD policy shows a bimodal distribution of attention with a reduced peak at 3600Hz and a second peak at 1000Hz. The SM policy is also strongly bimodal with a primary peak at 7000Hz and a secondary peak at 600Hz. Overall, only the *LD* and *SM* policies induced strong changes in frequency distribution.

Takeaway: The results presented here show that the differences in model behavior induced by data augmentation are readily visible in our pipeline. We observe behavior that is both expected (diffusion of attention after masking) and counter-intuitive (sparse attention after *LD, SS* augmentations). We note that the frequency distribution of attention regions aligns well with expectations based on existing literature. The variation in both attention region diffusion and frequency distribution between augmentations indicates that the similarity of augmentations does not imply similarity in the final model behavior — underscoring the practical value of the approach we present.

6 Audio Augmentation for Improved Resilience Against Common Audio Encodings

In addition to evaluating existing augmentation techniques, behavior visualization can aid in the development and validation of augmentations where specific behaviors are desired or are known to be undesirable. To illustrate this, we demonstrate the development of an augmentation to reduce performance degradation on destructively encoded audio. The encoding system could be any number

of lossy compression algorithms, transmission across traditional or VoIP telephony systems, or other audio transformations that do not fully preserve the original audio content. Encoded audio more accurately represents deepfakes as they may be found in the wild, as opposed to clean, high-fidelity lab-recorded/generated samples and is motivated by several examples of real-world AI-enabled telephone fraud [45, 47] and attacks in the academic literature [8, 29, 68].

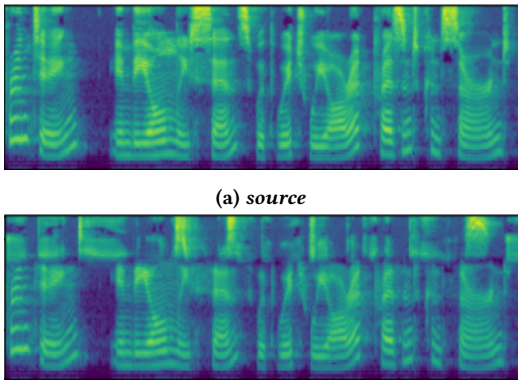
We simulate a telephony system using the G.711 audio codec, a high-quality, narrowband codec in wide use throughout the Public Switched Telephone Network which encompasses most global traditional and VoIP telephony [28]. Therefore, any audio, fake or otherwise, that is transmitted through the telephony system has a high chance of being encoded to the G.711 standard. This audio codec is designed to preserve information in the frequency region that is most associated with speech intelligibility (300Hz-3400Hz) and maintains a relatively high perceived quality of speech. However, the codec's 8000Hz sample rate leads to the loss of all speech information above 4000Hz. Additionally, the source audio is compressed to an 8-bit depth through μ-law encoding, which provides higher quality reproduction of lower amplitude samples at the cost of reduced dynamic range and distortion of high amplitude samples.

Given that the G.711 codec primarily impacts information above 4000Hz, we can label all model attention above this point as undesirable. The first step is to determine whether our model exhibits such behavior. To do so, we can look at the frequency distribution for the base model given in Figure 5, which shows a moderate amount of attention explicitly above 4000Hz. Additionally, given that the large attention peak at 3600Hz represents the center of those islands, we can expect that some regions centered on this peak will have attention above 4000Hz. Knowing that the base model exhibits moderate attention in the unpreserved region, we explicitly define two goals: *reduce or eliminate attention in and adjacent to the unpreserved region and induce more diffuse attention in general*, which is associated with more robust performance.

Our augmentation uses an alternative form of frequency masking, randomizing all information above 4000Hz, but maintaining the speech-like characteristics. Rather than replace the masked region with a constant value, we randomly select another sample and replace the high-frequency information in the original sample with the high-frequency information in the random sample. This is accomplished by low-pass filtering the original sample, high-pass filtering the random sample at the same cutoff frequency, and summing the two signals. The result is that the masked region contains random data that is not useful for classification, but still maintains the expected structure of speech data in that region. An example spectrogram before and after augmentation is shown in Figure 6.

When evaluating the performance of augmentations, we must consider performance on both the original, undistorted audio and the target distorted or out-of-distribution audio. In addition to the unencoded and the G.711 encoded data, we evaluate augmentation performance on two other classes of distortion: a more modern, high-fidelity, and similarly widely used encoding, OPUS, and generic noise addition in which Gaussian noise is added to each sample to produce a low but perceptible noise floor.

To determine the impact on perceptual quality for each distortion class, we employ the ViSQOL metric proposed by Chinen et al. [11] which gives an objective measure of perceptual change relative to an unperturbed sample as a Mean Opinion Score (MOS). MOS ranges from 5 (high quality/low distortion) to 1 (poor quality/high distortion). For reference, an MOS of 3-3.5 is generally considered sufficient for usable commercial VoIP service while an MOS of 4.3-4.5 is considered excellent or high-fidelity VoIP. ViSQOL MOS metrics are summarized in Table 3. The average ViSQOL score for OPUS-encoded samples is 4.38, G.711 is 3.32, and noisy samples score 3.08.



(a) *source*

(b) *spliced*

Figure 6: Spectrograms for samples before and after augmentation showing the replacement of the high frequency region with random speech.

Distortion	ViSQOL MOS
None (Baseline)	4.52
G.711 Codec	3.32
OPUS Codec	4.43
Noise Add.	3.08

Table 3: ViSQOL Mean Opinion Scores (MOS) for each tested audio distortion. Scores between 3 and 4 are considered acceptable for commercial VoIP while scores between 4 and 5 are considered high-fidelity.

Augmentation	Base	G.711	Noise	Opus
None	0.997	0.751	0.308	0.115
Splice	0.936	**0.878**	**0.600**	**0.773**
SA LB	0.998	0.552	0.030	0.051
SA LD	**1.000**	0.716	0.276	0.223
SA SM	**1.000**	0.608	0.084	0.111
SA SS	**1.000**	0.702	0.167	0.200

Table 4: Performance of base and data-augmented models on baseline and distorted audio. Performance is reported by MCC. Top-performing entries are highlighted and bold.

6.1 Augmentation Results

The performance results for each augmentation are shown in Table 4. The Splice augmentation provided the best performance on all distorted audio. It outperforms the SpecAugment augmentations by an average of 36% on the target G.711 encoded audio. Although the raw performance of the Splice augmentation is lower on the other distortions than on the target G.711 distortion, the relative improvement over competing augmentations is significantly greater, rising to a 330% improvement on Noise data and a 429% improvement on Opus encoded data. The SpecAugment augmentations do achieve perfect classification on the undistorted audio, a small but significant improvement over the base model. However, the dramatically better performance of the Splice augmentation on distorted audio, achieved while maintaining acceptable performance on the undistorted samples, makes it a more robust choice than the alternatives.

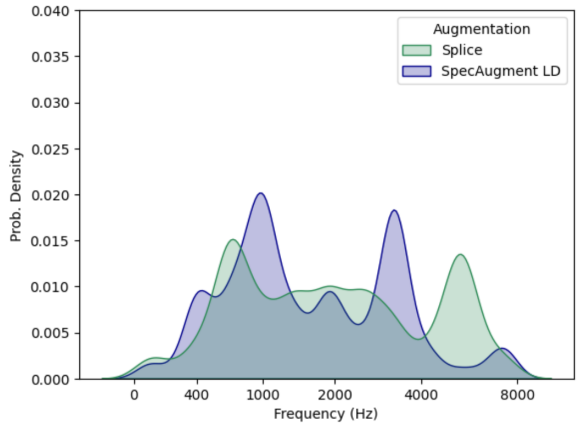


Figure 7: Probability distribution function of attention islands for the Splice augmentation and the LD SpecAugment policy, the highest performing SpecAugment policy. LD better eliminates high-frequency attention but has less uniform attention overall.

When we view the distribution of attention for the Splice augmentation (Figure 7), we see that it exhibits a more uniform behavior across frequency ranges. The distribution does have two small peaks: one at 700Hz and the other at 5500Hz. The high-frequency peak closely matches a small peak in the base model distribution (Figure 5), indicating that, despite our augmentation's high performance, we did not entirely eliminate high-frequency attention. The Splice distribution does show that the major peak at 3600Hz has been entirely eliminated, with most attention now distributed between 600Hz and 2500Hz.

Takeaway: Using the proposed approach, we can conclude that our augmentation successfully encourages diffuse attention and partially eliminates high-frequency attention. While the raw performance characteristics of this augmentation would be a compelling justification for its use, confirmation that the induced behavior matches the rationale for the performance increase greatly increases user confidence in the augmentation. Such introspection would not have been possible without our deliberate choice to utilize a spectrogram-based architecture.

7 Performance Evaluation

Our approach, as with other class-based visualization approaches (Grad-CAM, TIS), requires that the model be an effective classifier to ensure that the change in confidence from spectrogram occlusion is strongly related to the relevant features in the target audio. We validate AST's performance on several modern deepfake datasets to ensure that it is suitable for complex detection tasks. Additionally, given the substantial number of audio classification approaches being applied to audio deepfake detection and the limited usage of AST in the same space, we demonstrate that AST matches or exceeds the performance of three commonly used, high-performing classifiers, further motivating its wider adoption.

7.1 Datasets

The first dataset, Wavefake [18] is a lab-generated dataset. Wavefake's composition is unique in that it contains wide variability in deepfake generators (7 generators), but no speaker variation, with all real samples taken from a single female speaker. In The Wild [46] provides both higher speaker variation (58 speakers) and a representative selection of generators as they occur in real-world deepfakes. Both of these datasets are widely used in existing literature. We also incorporate data from the lesser-known Political Deepfakes Incidents Database (PDID) as an additional source of "deepfakes appearing in the wild" [64].

Structurally, each dataset consists of real and deepfake audio, divided into training, validation, and testing splits in a 7:2:1 ratio. The audio samples for each dataset are stored as .wav files with a sampling rate of 16000Hz and a bit depth of 16 bits.

Wavefake [18] is derived from the LJ Speech [26] dataset which consists of 13,100 recordings of a single speaker reading non-fiction passages. Wavefake augments this with several voice conversion models that attempt to reproduce the original audio from a transcript and training examples of the speaker. The architectures used to generate spoofed audio were selected based on high-performing models that include MelGAN Large, Full Band MelGAN, WaveGlow, HiFiGAN, and Parallel WaveGAN. The fake audio data contains an equal number of samples from each generator.

In The Wild [46] is a collection of *in vivo*-captured deepfakes of 58 celebrity voices, along with real samples for all speakers, totaling 38 hours of audio. We remove samples with durations less than 2.0s as some excessively short files do not contain recognizable speech. This results in a final dataset size of 31,779 samples. While the number and types of generators are unknown, the varied, real-world selection makes this a valuable performance comparison.

The Political Deepfakes Incidents Database (PDID) [64] is a curated list of deepfakes of prominent political figures. The database includes video samples collected from social media and news outlets. We extracted audio from the videos and excluded samples that contained a mixture of real and spoofed audio, or those labeled as "cheap fakes" (e.g., speeding, slowing, or otherwise re-contextualized footage). Each sample was manually vetted by the authors, removing unrelated audio (e.g., commentary at the beginning), and we separated multi-speaker samples into individual

Dataset \ Metric	Real Samples	Fake Samples	No. Spkrs	No. Gens.	Avg Length
Wavefake	13,100	91,700	1	7	6.6s
In The Wild	11,816	19,963	58	-	4.3s
PDID	none*	41	-	-	15.9s

Table 5: Composition information. *We augmented the PDID dataset with real samples from the Fake-or-Real [55] corpus. The '-' symbol denotes unknown.

files. As this dataset does not contain any bonafide audio, real audio from a different speech corpus [55] was incorporated to serve as training data. This corpus contains both male and female samples from a variety of speech corpora. We calculate the distribution of sample lengths in this real audio, split the PDID samples to match that distribution, and normalize the amplitude of each sample to reduce the potential for introducing extraneous distributional differences.

7.2 Selecting Competitive Detectors

We compare AST to three advanced detectors, namely Wav2Vec2.0 [2], Whisper [53], and RawNet2 [60].

The RawNet2 architecture is a convolution-based deep learning approach that operates directly on the raw audio waveform rather than computing audio statistics like Mel Frequency Cepstral Coefficients or Linear Frequency Cepstral Coefficients prior to classification. The architecture is one of the baseline models associated with the ASVspoofdataset and is widely used as a point of comparison for novel detection approaches. Additionally, it serves as a fundamental baseline for non-transformer performance. We trained RawNet2 from scratch for each performance evaluation, following the training approach suggested in the original paper [60].

The Wav2Vec2.0 detector was selected given its current status as the highest-performing approach on the ASVspoof dataset and its broad adoption [41, 61, 70] across audio classification and automatic speech recognition fields. Wav2Vec2.0 represents a strong point of comparison for all other transformer-based approaches.

Finally, Whisper is another high-performance transformer architecture that emphasizes robust automatic speech recognition. It has seen moderate adoption in the audio deepfake detection field [37], often as part of ensemble models [39, 50] or as a trainable preprocessor [30, 50].

For the transformer-based models, we utilize pre-trained models provided by the HuggingFace repository and fine-tune them for each performance evaluation. We use the `facebook/wav2vec2-base` and `openai/whisper-large` checkpoints for Wav2Vec2.0 and Whisper, respectively. Together, the selected models provide coverage of both the current state-of-the-art detector architectures and significant milestones in audio deepfake detection.

Training Parameters: Each model is trained for 10 epochs, using the Adam [32] gradient optimizer and the cross-entropy loss function. The learning rate for AST models was set at 1e-5. A warmup ratio of 0.05 was used to ramp up the learning rate from 0 to its set point over the first 5% of training steps. Doing so limits training misbehavior while the model output is largely random. A plateau learning rate scheduler is used to ramp down the learning rate once validation performance improvements stabilize or begin to degrade. This strategy limits the overfitting of models that may reach a plateau earlier than 10 epochs. More information about hyperparameters is given in Appendix 10.1.

7.3 Metrics

To gauge the efficacy of our approach, we conduct evaluations using Accuracy, Equal Error Rate, and Matthews Correlation Coefficient metrics. In the context of these metrics, spoofed samples are considered as *positive*. False positive rate, therefore, refers to the rate at which real samples are misclassified.

7.3.1 Accuracy (Acc). The ratio of correct predictions to total predictions. It is easily interpretable, but it can be misleading as strong performance on an overrepresented class can lead to overly optimistic performance.

$$Acc = \frac{TP + TN}{TP + TN + FP + FN}$$

7.3.2 Equal Error Rate (EER). The average error rate of a model whose decision threshold is set such that the model's false positive rate (FPR) and false negative rate (FNR) are equal. The decision threshold is determined by a linear search, evaluating the FPR and FNR at each point. The threshold with the minimum difference between FPR and FNR is selected, and the EER is reported as:

$$EER = FPR + FNR$$

Unfortunately, even though EER is no longer recommended as a performance metric for authentication systems [59], it is the de facto standard in deepfake detection papers. Thus, we include it here primarily to facilitate comparisons with prior work.

7.3.3 Matthews Correlation Coefficient (MCC). For binary classification tasks, the Area Under the Curve of the Receiver Operating Characteristic curve is commonly used. However, if not carefully applied, that metric can provide inflated results [10]. A more appropriate choice is the Matthews correlation coefficient [4, 42] that better captures when base metrics (recall, precision, specificity, and negative predictive value) simultaneously achieve high scores. MCC ranges between $[-1, 1]$ and a value of 0 indicates a random classifier. Negative values indicate inverse predictions, i.e. the model is more likely to generate an incorrect prediction than correct.

$$MCC = \frac{TP * TN - FP * FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

Dataset	Model	Acc	EER	MCC
Wavefake	AST	0.99	0.001	**0.99**
	Wav2Vec	0.99	0.006	0.97
	Whisper	0.94	0.022	0.89
	RawNet2	0.98	0.01	0.96
In The Wild	AST	1.00	0.00	**1.00**
	Wav2Vec	1.00	0.00	**1.00**
	Whisper	0.97	0.03	0.95
	RawNet2	0.91	0.04	0.84
PDID	AST	1.00	0.00	**1.00**
	Wav2Vec	1.00	0.00	**1.00**
	Whisper	1.00	0.00	**1.00**
	RawNet2	0.92	0.04	0.86

Table 6: Model performance all datasets. Top-performing entries are highlighted in bold.

We include MCC scores because they offer a comprehensive evaluation of model performance in a single metric [71].

7.4 Results from Standard Testing Scenarios

The results in Table 6 show that our application of the Audio Spectrogram Transformer model for deepfake detection performs exceedingly well, achieving MCC scores of 0.99, 1.00, and 1.00 on the Wavefake, In The Wild, and PDID datasets, respectively. In particular, it outperforms all competing architectures on the Wavefake dataset, with the other state-of-the-art detectors, Wav2Vec2 and Whisper, achieving MCCs of 0.97 and 0.89, respectively. On the In The Wild, both AST and Wav2Vec achieve perfect classification, while RawNet dataset, AST performance is slightly lower than Wav2Vec, though the EER of 0.02 puts it within the top models on this dataset. All transformer models (AST, Wav2Vec, Whisper) achieve perfect classification on the PDID dataset, despite a total training data duration of only 25.4 minutes. This high performance may be due to distributional differences between the PDID fake samples and the introduced real samples used for training. Though steps were taken to minimize potential differences, this risk is implicit for datasets in which real and fake samples are not collected from the same environment.

Interestingly, performance on the Wavefake dataset is generally worse than for either real-world dataset, with AST and Wav2Vec showing small, but noticeable, reductions (1% and 3% MCC reduction) and Whisper showing a larger drop of 7.3%. This is despite the simplified testing scenario presented by the Wavefake dataset, where speaker variation is minimized. These results may indicate that the quality of lab-produced deepfakes is higher than that of the typical *in vivo* deepfake. Alternatively, as noted above, real-world datasets must contend with strong distributional differences between classes for features unrelated to the primary classification task (variation in file source compression, leading and trailing silence duration, etc.). While steps have been taken to minimize the risk of confounding dataset characteristics in these evaluations, as noted above, the difficulty in preparing diverse audio datasets remains.

Takeaway: The findings indicate that the AST architecture is well suited for challenging audio deepfake detection tasks, on par with state-of-the-art detectors. The high level of performance attained — coupled with the explainability inherent in the spectrogram representation — makes it a strong candidate for detector-based analysis of deepfake generators and ensures its utility on detection tasks where feature localization may be non-trivial.

8 Future Work

Although both G.711 and OPUS are widely used codecs, they only represent a subset of the types of transforms in the broader class of encoding schemes. Further research is needed to identify shared characteristics between the class as a whole and to identify methods to help design detectors that are resilient to a wider spectrum of encoding schemes, as well as to identify other sources of systematic distortion that are likely present in audio transmission channels.

9 Conclusions

The increase in spread and quality of audio deepfakes necessitates a deep understanding of the behavior of deepfake detectors and how this behavior can be modified to achieve more robust detection. We demonstrate that 'common sense' expectations of the impact of data augmentation often translate poorly to the actual change in model behavior, even when the augmentations improve model performance. This revelation is enabled by adopting classifier architectures that provide easily interpretable behavior, which can be achieved without sacrificing detection performance. We further show that this interpretability can be used to not only grant insight into the impact of existing augmentations, revealing both expected and counterintuitive behaviors, but can also be leveraged by researchers to better inform the augmentation development process. We illustrate this utility by applying our framework to the development of a novel augmentation based on a priori assumptions of desirable behavior and demonstrate significant improvements in detector resilience to common audio distortions over existing general augmentations. We validate that the behavior induced by this augmentation aligns with our expectations.

We hope that the approach we present for improvements in detector interpretability will help other researchers design more robust models and augmentations that offer effective classification of high-quality deepfakes in real-world environments, thus decreasing the potential for societal harm that audio deepfakes currently pose.

10 Code and Data Availability

To encourage further research in this area, we have shared all code used to train models, visualize and compare model attention, and generate augmented audio samples at https://github.com/funkshun/ASTExplainability. Additionally, we provide the trained models for our evaluations and code used to compute the performance metrics in Sections 6.1 and 7.4.

References

[1] 2022. Deep learning for deepfakes creation and detection: A survey. *Computer Vision and Image Understanding* 223 (2022), 103525.

[2] Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. arXiv:2006.11477 [cs.CL]

[3] Muhammad Khurram Zahur Bajwa, Aniello Castiglione, and Chiara Pero. 2025. Mel Spectrogram-Based CNN Framework for Explainable Audio Deepfake Detection. In *Advanced Information Networking and Applications*, Leonard Barolli (Ed.). Springer Nature Switzerland, Cham, 407–416.

[4] Pierre Baldi, Søren Brunak, Yves Chauvin, Claus AF Andersen, and Henrik Nielsen. 2000. Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics* 16, 5 (2000), 412–424.

[5] Joseph R Biden. 2023. Executive order on the safe, secure, and trustworthy development and use of artificial intelligence. (2023).

[6] Carmen Bisogni, Vincenzo Loia, Michele Nappi, and Chiara Pero. 2024. Acoustic features analysis for explainable machine learning-based audio spoofing detection. *Computer Vision and Image Understanding* 249 (2024), 104145. https://doi.org/10.1016/j.cviu.2024.104145

[7] Logan Blue, Kevin Warren, Hadi Abdullah, Cassidy Gibson, Luis Vargas, Jessica O'Dell, Kevin Butler, and Patrick Traynor. 2022. Who are you (I really wanna know)? Detecting audio DeepFakes through vocal tract reconstruction. In *USENIX Security Symposium*. 2691–2708.

[8] Guangke Chen, Sen Chen, Lingling Fan, Xiaoning Du, Zhe Zhao, Fu Song, and Yang Liu. 2019. Who is Real Bob? Adversarial Attacks on Speaker Recognition Systems. *IEEE Symposium on Security and Privacy* (2019), 694–711.

[9] Minhao Cheng, Simranjit Singh, Patrick Chen, Pin-Yu Chen, Sijia Liu, and Cho-Jui Hsieh. 2020. Sign-OPT: A Query-Efficient Hard-label Adversarial Attack. arXiv:1909.10773 [cs.LG]

[10] Davide Chicco and Giuseppe Jurman. 2023. The Matthews correlation coefficient (MCC) should replace the ROC AUC as the standard metric for assessing binary classification. *BioData Mining* (2023).

[11] Michael Chinen, Felicia S. C. Lim, Jan Skoglund, Nikita Gureev, Feargus O'Gorman, and Andrew Hines. 2020. ViSQOL v3: An Open Source Production Ready Objective Speech and Audio Metric. arXiv:2004.09584 [eess.AS]

[12] Thien-Phuc Doan, Long Nguyen-Vu, Souhwan Jung, and Kihun Hong. 2023. BTS-E: Audio Deepfake Detection Using Breathing-Talking-Silence Encoder. *International Conference on Acoustics, Speech and Signal Processing* (2023), 1–5.

[13] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. arXiv:2010.11929 [cs.CV]

[14] Alexandre Englebert, Sédrick Stassin, Géraldin Nanfack, Sidi Ahmed Mahmoudi, Xavier Siebert, Olivier Cornu, and Christophe De Vleeschouwer. 2023. Explaining through Transformer Input Sampling. In *International Conference on Computer Vision*. 806–815.

[15] Youngsik Eom, Yeonghyeon Lee, Ji Sub Um, and Hoi Rin Kim. 2022. Anti-Spoofing Using Transfer Learning with Variational Information Bottleneck. In *Interspeech*.

[16] Hany Farid. 2022. Creating, using, misusing, and detecting deep fakes. *Journal of Online Trust and Safety* 1, 4 (2022).

[17] FCC. 2024. FCC Makes AI-Generated Voices in Robocalls Illegal. https://www.fcc.gov/document/fcc-makes-ai-generated-voices-robocalls-illegal. [Accessed 29-04-2024].

[18] Joel Frank and Lea Schönherr. 2021. WaveFake: A dataset to facilitate audio DeepFake detection.

[19] John J. Godfrey and Edward Holliman. 1993. Switchboard-1 Release 2 LDC97S62.

[20] Chirag Goel, Surya Koppisetti, Ben Colman, Ali Shahriyari, and Gaurav Bharaj. 2023. Towards Attention-based Contrastive Learning for Audio Spoof Detection. In *INTERSPEECH 2023*. ISCA, 2758–2762. https://doi.org/10.21437/interspeech.2023-245

[21] Yuan Gong, Yu-An Chung, and James Glass. 2021. AST: Audio Spectrogram Transformer. arXiv:2104.01778 [cs.SD]

[22] Joshua Habgood-Coote. 2023. Deepfakes and the epistemic apocalypse. *Synthese* 201, 3 (2023), 103.

[23] Keith Raymond Harris. 2022. Real fakes: The epistemology of online misinformation. *Philosophy & Technology* 35, 3 (2022), 83.

[24] Keith Raymond Harris. 2024. AI or Your Lying Eyes: Some Shortcomings of Artificially Intelligent Deepfake Detectors. *Philosophy & Technology* 37, 1 (2024), 7.

[25] Wiebke Hutiri, Oresiti Papakyriakopoulos, and Alice Xiang. 2024. Not My Voice! A Taxonomy of Ethical and Safety Harms of Speech Generators. *arXiv preprint arXiv:2402.01708* (2024).

[26] Keith Ito and Linda Johnson. 2017. The LJ Speech Dataset. https://keithito.com/LJ-Speech-Dataset/.

[27] Sarthak Jain and Byron C. Wallace. 2019. Attention is not Explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Jill Burstein, Christy Doran, and Thamar Solorio (Eds.). Association for Computational Linguistics, Minneapolis, Minnesota, 3543–3556. https://doi.org/10.18653/v1/N19-1357

[28] Stylianos Karapantazis and Fotini-Niovi Pavlidou. 2009. VoIP: A comprehensive survey on a promising technology. *Computer Networks* 53, 12 (2009), 2050–2090.

[29] Andre Kassis and Urs Hengartner. 2023. Breaking Security-Critical Voice Authentication. In *IEEE Symposium on Security and Privacy (SP)*. IEEE, 951–968.

[30] Piotr Kawa, Marcin Plata, Michał Czuba, Piotr Szymański, and Piotr Syga. 2023. Improved DeepFake Detection Using Whisper Features. (2023).

[31] Zahra Khanjani, Gabrielle Watson, and Vandana P. Janeja. 2021. How Deep Are the Fakes? Focusing on Audio Deepfake: A Survey. arXiv:2111.14203 [cs.SD]

[32] Diederik P. Kingma and Jimmy Ba. 2017. Adam: A Method for Stochastic Optimization. arXiv:1412.6980 [cs.LG]

[33] Nils C Köbis, Barbora Doležalová, and Ivan Soraperra. 2021. Fooled twice: People cannot detect deepfakes but think they can. *Iscience* 24, 11 (2021).

[34] Seth Layton, Thiago De Andrade, Daniel Olszewski, Kevin Warren, Carrie Gates, Kevin Butler, and Patrick Traynor. 2024. Every Breath You Don't Take: Deepfake Speech Detection Using Breath. *arXiv preprint arXiv:2404.15143* (2024).

[35] Tuan Duy Nguyen Le, Kah Kuan Teh, and Huy Dat Tran. 2024. Continuous Learning of Transformer-based Audio Deepfake Detection. arXiv:2409.05924 [cs.SD] https://arxiv.org/abs/2409.05924

[36] Lauren Leffer. 2024. AI Audio Deepfakes Are Quickly Outpacing Detection — scientificamerican.com. https://www.scientificamerican.com/article/ai-audio-deepfakes-are-quickly-outpacing-detection/. [Accessed 29-04-2024].

[37] Menglu Li, Yasaman Ahmadiadli, and Xiao-Ping Zhang. 2024. Audio Anti-Spoofing Detection: A Survey. arXiv:2404.13914 [cs.SD]

[38] Yuang Li, Min Zhang, Mengxin Ren, Miaomiao Ma, Daimeng Wei, and Hao Yang. 2024. Cross-Domain Audio Deepfake Detection: Dataset and Analysis. arXiv:2404.04904 [cs.SD]

[39] Qian Luo and Kalyani Vinayagam Sivasundari. 2024. Whisper+ AASIST for DeepFake Audio Detection. In *International Conference on Human-Computer Interaction*. Springer, 121–133.

[40] Harry Maltby, Julie Wall, Cornelius Glackin, Mansour Moniri, Nigel Cannings, and Iwa Salami. 2024. A Frequency Bin Analysis of Distinctive Ranges Between Human and Deepfake Generated Voices. In *International Joint Conference on Neural Networks (IJCNN) - Neural Networks Models*.

[41] Juan M. Martín-Doñas and Aitor Álvarez. 2022. The Vicomtech Audio Deepfake Detection System Based on Wav2vec2 for the 2022 ADD Challenge. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 9241–9245.

[42] Brian W Matthews. 1975. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochimica et Biophysica Acta (BBA)-Protein Structure* 405, 2 (1975), 442–451.

[43] Pulak Mehta, Gauri Jagatap, Kevin Gallagher, Brian Timmerman, Progga Deb, Siddharth Garg, Rachel Greenstadt, and Brendan Dolan-Gavitt. 2023. Can Deepfakes be created on a whim?. In *ACM Web Conference*.

[44] Nicolas M. Müller, Karla Pizzi, and Jennifer Williams. 2022. Human Perception of Audio Deepfakes. 85–91.

[45] Shannon Murphy. 2024. Deepfake CFO Video Calls Result in $25M in Damages — trendmicro.com. Trend Micro.

[46] Nicolas M. Müller, Pavel Czempin, Franziska Dieckmann, Adam Froghyar, and Konstantin Böttinger. 2022. Does Audio Deepfake Detection Generalize? arXiv:2203.16263 [cs.SD]

[47] Security News. 2019. Unusual CEO Fraud via Deepfake Audio Steals US $243,000 From UK Company.

[48] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: An ASR corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 5206–5210.

[49] Daniel S. Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D. Cubuk, and Quoc V. Le. 2019. SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition. In *Interspeech*. ISCA.

[50] Lam Pham, Phat Lam, Truong Nguyen, Huyen Nguyen, and Alexander Schindler. 2024. Deepfake Audio Detection Using Spectrogram-based Feature and Ensemble of Deep Learning Models. In *2024 IEEE 5th International Symposium on the Internet of Sounds (IS2)*. IEEE, 1–5.

[51] Michael I Posner, Mary J Nissen, and Raymond M Klein. 1976. Visual dominance: an information-processing account of its origins and significance. *Psychological review* 83, 2 (1976), 157.

[52] Zengyi Qin, Wenliang Zhao, Xumin Yu, and Xin Sun. 2024. OpenVoice: Versatile Instant Voice Cloning. arXiv:2312.01479 [cs.SD]

[53] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust Speech Recognition via Large-Scale Weak Supervision. arXiv:2212.04356 [eess.AS]

[54] Henning Reetz and Allard Jongman. 2020. *Phonetics: Transcription, production, acoustics, and perception*. John Wiley & Sons.

[55] Ricardo Reimao and Vassilios Tzerpos. 2019. FoR: A Dataset for Synthetic Speech Detection. In *International Conference on Speech Technology and Human-Computer Dialogue*. 1–10.

[56] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*. 1135–1144.
[57] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2019. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. *International Journal of Computer Vision* 128, 2 (Oct. 2019), 336–359.
[58] Tsu-Hsien Shih, Chin-Yuan Yeh, and Ming-Syan Chen. 2024. Does Audio Deepfake Detection Rely on Artifacts? *IEEE Conference on Acoustics, Speech and Signal Processing* (2024).
[59] Shridatt Sugrim, Can Liu, Meghan McLean, and Janne Lindqvist. 2019. Robust performance metrics for authentication systems. In *Network and Distributed Systems Security Symposium*.
[60] Hemlata Tak, Jose Patino, Massimiliano Todisco, Andreas Nautsch, Nicholas Evans, and Anthony Larcher. 2021. End-to-end anti-spoofing with RawNet2. arXiv:2011.01108 [eess.AS]
[61] Hemlata Tak, Massimiliano Todisco, Xin Wang, Jee weon Jung, Junichi Yamagishi, and Nicholas Evans. 2022. Automatic speaker verification spoofing and deepfake detection using wav2vec 2.0 and data augmentation. arXiv:2202.12233 [eess.AS]
[62] S. Umesh, L. Cohen, and D. Nelson. 1999. Fitting the Mel scale. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vol. 1. 217–220 vol.1.
[63] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2023. Attention Is All You Need. arXiv:1706.03762
[64] Christina P. Walker, Daniel S. Schiff, and Kaylyn Jackson Schiff. 2024. Merging AI Incidents Research with Political Misinformation Research: Introducing the Political Deepfakes Incidents Database. *AAAI Conference on Artificial Intelligence* 38, 21 (Mar. 2024), 23053–23058.
[65] Chenglong Wang, Jiayi He, Jiangyan Yi, Jianhua Tao, Chu Yuan Zhang, and Xiaohui Zhang. 2024. Multi-Scale Permutation Entropy for Audio Deepfake Detection. In *International Conference on Acoustics, Speech and Signal Processing*. 1406–1410.
[66] Shu Wang, Kun Sun, and Qi Li. 2023. Compensating Removed Frequency Components: Thwarting Voice Spectrum Reduction Attacks. *arXiv preprint arXiv:2308.09546* (2023).
[67] Xin Wang and Junich Yamagishi. 2021. A Comparative Study on Recent Neural Spoofing Countermeasures for Synthetic Speech Detection. arXiv:2103.11326 [eess.AS]
[68] Emily Wenger, Max Bronckers, Christian Cianfarani, Jenna Cryan, Angela Sha, Haitao Zheng, and Ben Y Zhao. 2021. "Hello, It's Me": Deep Learning-based Speech Synthesis Attacks in the Real World. In *Proceedings of the ACM SIGSAC Conference on Computer and Communications Security*. 235–251.
[69] Sarah Wiegreffe and Yuval Pinter. 2019. Attention is not not Explanation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (Eds.). Association for Computational Linguistics, Hong Kong, China, 11–20.
[70] Yuankun Xie, Haonan Cheng, Yutian Wang, and Long Ye. 2023. Learning A Self-Supervised Domain-Invariant Feature Representation for Generalized Audio Deepfake Detection. In *Proc. INTERSPEECH*, Vol. 2023. 2808–2812.
[71] Jingxiu Yao and Martin Shepperd. 2020. Assessing software defection prediction performance: why using the Matthews correlation coefficient matters. In *International Conference on Evaluation and Assessment in Software Engineering*. 120–129.
[72] Jiangyan Yi, Chenglong Wang, Jianhua Tao, Xiaohui Zhang, Chu Yuan Zhang, and Yan Zhao. 2023. Audio Deepfake Detection: A Survey. arXiv:2308.14970 [cs.SD]
[73] Ning Yu, Long Chen, Tao Leng, Zigang Chen, and Xiaoyin Yi. 2024. An explainable deepfake of speech detection method with spectrograms and waveforms. *Journal of Information Security and Applications* 81 (2024), 103720. https://doi.org/10.1016/j.jisa.2024.103720

10.1 Training Hyperparameters

Training was conducted using the hyperparameters given in Table 7 on an nVidia A30 GPU. The hyperparameters were taken from recommended values distributed with each of the models. The batch size was selected based on available GPU memory.

Parameter	Value
Epochs	10
Learning Rate	1×10^{-6}
Batch Size	8
Gradient Accumulation Steps	4
Warmup Ratio	0.05
LR Scheduler Factor	0.3
LR Scheduler Patience	1.0

Table 7: Training Hyperparameters

10.2 Adversarial Attacks

In this paper, we do not consider optimized distortions introduced by adversarial example generators, instead focusing on distortion types that are likely to be introduced in non-adversarial settings, but which still pose a significant risk to detector efficacy. The analyzed augmentation techniques are intended to provide improvements in these areas, but only provide static defenses, i.e., the decision boundary of the detector does not move once the model is deployed, which is insufficient to prevent adversarial optimization. Therefore, given a sufficient number of model queries, these generators can produce a minimal distortion modification to any audio to cause it to be misclassified, regardless of data augmentation. Nevertheless, for completeness, we adapt the Sign-OPT black-box attack [9] from the image classification space to apply to the generation of adversarial audio perturbations. The Sign-OPT attack requires significantly fewer queries to optimize than competing approaches. We tested the attack against all three transformer architectures in our evaluation and found that, as expected, all three were susceptible to the adversarial examples. The samples had an average ViSQOL score of 4.448, which is very close to the baseline average of 4.52, demonstrating the minimal distortion characteristics of these attacks. On average, Sign-OPT required approximately 20,000 queries to produce an adversarial example against AST regardless of augmentation. This was higher than Wav2Vec and Whisper (approx. 16000 queries), but we still consider it a successful attack. This class of adversarial approach remains a significant threat to detector robustness despite significant work in this area.