What You Read Isn't What You Hear: Linguistic Sensitivity in Deepfake Speech Detection

Anonymous ACL submission

Abstract

Recent advances in text-to-speech technologies have enabled realistic voice generation, fueling audio-based deepfake attacks such as fraud and impersonation. While audio anti-spoofing systems are critical for detecting such threats, prior work has predominantly focused on acoustic-level perturbations, leaving the impact of linguistic variation largely unexplored. In this paper, we investigate the linguistic sensitivity of both open-source and commercial anti-spoofing detectors by introducing transcript-level adversarial attacks. Our extensive evaluation reveals that even minor linguistic perturbations can significantly degrade detection accuracy: attack success rates surpass 60% on several open-source detector–voice pairs, and notably one commercial detection accuracy drops from 100% on synthetic audio to just 32%. Through a comprehensive feature attribution analysis, we identify that both linguistic complexity and model-level audio embedding similarity contribute strongly to detector vulnerability. We further demonstrate the real-world risk via a case study replicating the Brad Pitt audio deepfake scam, using transcript adversarial attacks to completely bypass commercial detectors. These results highlight the need to move beyond purely acoustic defenses and account for linguistic variation in the design of robust anti-spoofing systems. All source code will be publicly available.

1 Introduction

Recent advances in text-to-speech (TTS) technology have enabled natural speech synthesis in over 7,000 languages (Lux et al., 2024) and high-fidelity audio from just a single sample of a target voice (Chen et al., 2024). However, these innovations have also made it easier for attackers to create deepfake audio for identity fraud, evident in a surge by more than 2,000% over the past three years in deepfake fraud (Da Silva, 2024) and the recent notorious case of Brad Pitt impersonation scam of over $800K (Signicat, 2024).

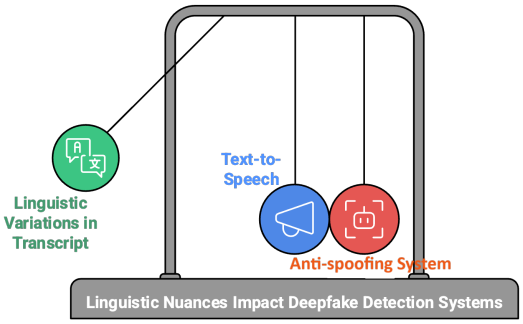

Figure 1: Linguistic variation of the transcript can swing the confidence of audio anti-spoofing system.

To counter deepfake audio, audio anti-spoofing systems (AASs) have been developed to distinguish genuine–i.e., human-spoken speech, from spoofed one–i.e., machine-synthesized speech for identify falsification (Jung et al., 2021; Tak et al., 2021; Wu et al., 2024; Tak et al., 2022). However, AAS are known to be vulnerable to *acoustic-level manipulations* such as injection of small noise, volume modification, and even deliberate attacks or often so-called *adversarial manipulations* (Wu et al., 2024, 2020; Müller et al., 2023). Such the vulnerability is also analogous to *text-level manipulations* targeting deepfake text detectors where synonym replacement of only a few words or small variations in word choice can significantly alter their detection probabilities (Uchendu et al., 2023).

Since any speech fed into an AAS is either synthesized by TTS or spoken by humans from an input transcript written in human languages, it is natural to hypothesize that AASs, although only accept audio inputs, might be also *indirectly influenced by text-level, linguistic manipulations* on the audio' transcripts (Fig. 1). However, research questions such as *"whether such an effect of such linguistic variations, such as word choice or dialect, on anti-spoofing performances of AAS exist?*

or *"when and how much linguistic variations in transcripts influence the effectiveness of audio anti-spoofing systems?"* are under-explored.

Such questions are also intuitive and relevant to how humans perceive information from auditory speech. Particularly, linguistic differences can influence how human listeners perceive and evaluate speech, sometimes resulting in bias or negative responses (Peters, 2024). Such factors may also potentially introduce vulnerabilities or biases into AASs, especially from adversarial machine learning perspectives when they are often trained on human-curated data. For instance, terms like "illegal" versus "undocumented", although semantic neighbors, can affect a speaker's perceived credibility (Lim et al., 2018). If such linguistic sensitivity indeed exists in current AASs, they can be manipulated by malicious actors to carefully craft spoof audio that is much more challenging for AASs to accurately detect as fake.

To examine the linguistic sensitivity of AASs, this work takes an initial step toward evaluating the central hypothesis: *subtle linguistic variations within a transcript can propagate through a text-to-speech (TTS) pipeline and significantly impact the predictions of AASs* (Fig. 1). To this end, we formulate our investigation as an adversarial attack scenario, wherein a malicious actor strategically introduces minimal perturbations to an audio transcript, while preserving its original meaning, prior to its conversion to audio via a TTS pipeline, with the goal of evading detection by state-of-the-art (SOTA) AASs. Our empirical validation demonstrates that both research and production-grade detectors are significantly vulnerable to such subtle linguistic manipulations, with attack success rates exceeding 60% across both open-source and commercial AASs. Moreover, we show that linguistic nuances correspond to translated acoustic qualities in the spoofed audio, ultimately affecting AASs' accuracy. **Our contributions** can be summarized as follows.

1. To the best of our knowledge, our work is the first that formulates and examines linguistic sensitivity in automatic audio anti-spoofing systems.
2. We develop a transcript-level adversarial attack pipeline that generates semantically valid perturbations and demonstrates how subtle linguistic changes can degrade detection accuracy, in many cases from over 90% to just below 20%, in both open-source and commercial detectors.
3. We perform feature attribution analysis of over 14 linguistic, acoustic, and model-level features and analyze how they correlate with such linguistic vulnerability, offering insights for more robust audio anti-spoofing systems.

2 Motivation

2.1 Related Work

Most adversarial attack work in audio anti-spoofing focuses on signal or acoustic-level attacks, such as noise injection or frequency masking, to expose vulnerabilities in spoof detection models (Attorresi et al., 2022; Ba et al., 2023). However, little attention has been paid to the role of linguistic variation. As a result, efforts to improve robustness have mainly addressed acoustic distortions and cross-dataset challenges via domain adaptation and knowledge distillation (Arora et al., 2022). In contrast, specific impacts of transcript manipulations, such as what types of text perturbations and what are their effectiveness, remain underexplored in audio anti-spoofing. Whereas text-based adversarial attacks in NLP have demonstrated that small semantic changes can fool classifiers (Jin et al., 2020; Le et al., 2022), it is still unknown how such linguistic perturbations, when indirectly propagated through TTS synthesis (Fig. 1), would affect downstream audio spoofing detection.

2.2 Preliminary Analysis

We first carry out a preliminary analysis is to examine whether linguistic variations in transcripts might affect the robustness of AASs. To do this, we randomly substitute one word in each of 1439 transcripts with a synonym, synthesized audio for both versions with a TTS model, and test them with the high-performing open-source audio spoofing AASIST-2 detector. Surprisingly, even minimal, one-word changes cause AASIST-2 detector to misclassify up to 5.7% of samples, and bona-fide detection probabilities drop by as much as 67.9% in some cases (Table 1). Moreover, most open-source AAS are trained on the ASVSpoof-2019 LA dataset, which displays significant linguistic disparities between spoofed and bona-fide samples (Table A1). Statistical tests confirm that spoofed transcripts are statistically more complex in terms of token perplexity and readability than bona-fide cases (Table 2). Such disparities in training data can introduce linguistic bias into the trained anti-spoofing models. Motivated by these observations,

Transcript	Bona-fide %
She is a ~~successful~~**good** actor with . . .	3.6 → 80.5
The trust was unable to ~~pay~~**reward** the. . .	3.7 → 34.9
. . . for a ~~man~~**guy** or woman of letters.	3.7 → 50.2

Table 1: A few examples of preliminary transcript-level adversarial attacks on anti-spoofing detector AASIST-2.

we will then systematically evaluate and quantify what degree AASs are sensitive to small changes in audio' transcripts through a comprehensive, algorithmic approach.

3 Problem Formulation

Let $\mathcal{F}{:}X{\to}Y$ be an AAS, which maps the audio input X to the bona-fide label output Y. Given a set of N transcripts $\mathcal{T}{=}\{T_1, T_2, \ldots, T_N\}$ with $T_i{\in}\mathbb{R}^M$, and a TTS model $\mathcal{G}{:}T{\to}X$, we synthesize a collection of N audio $\mathcal{X}{=}\{X_1, X_2, \ldots, X_N\}$ with $X_i{\in}\mathbb{R}^L$, by entering each transcript in $\mathcal{T}$ to $\mathcal{G}(\cdot)$. Moreover, $\mathcal{Y}{=}\{Y_1, Y_2, \ldots, Y_N\}$ is the ground truth of $\mathcal{X}$, where $Y_i{\to}0$ indicates a spoofing label, and $Y_i{\to}1$ means a bona-fide label. In our setting, Y_i is a spoofing label–i.e., $Y_i{\leftarrow}0\ \forall i$, since X_i is a machine-synthesized audio. M is the number of words in a transcript, and L is the wavelength of an audio.

We define an AAS $\mathcal{F}(\cdot)$ as **linguistically sensitive to a specific TTS model** $\mathcal{G}(\cdot)$ if its prediction of audio synthesized via $\mathcal{G}(\cdot)$ is flipped *solely by making small changes to its underlying transcript while preserving its original semantic meaning*, without modifying the audio synthesis system, the speaker profile, the acoustic characteristics directly. Formally, given a transcript T, $\mathcal{F}(\cdot)$ exhibits linguistic sensitivity to $\mathcal{G}(\cdot)$ if there exists a *perturbed* transcript $\tilde{T}$ such that:

$$\text{SIM}(T, \tilde{T}){=}{=}1,\ \mathcal{F}(\mathcal{G}(\tilde{T})){=}{=}\tilde{Y}, \text{and } \tilde{Y}{\neq}Y, \tag{1}$$

where the boolean indicator $\text{SIM}(T, \tilde{T})$ ensures the adversarial transcript $\tilde{T}$ remains faithful to the original meaning or purpose (e..g, intend to transfer money) and also the original structure or syntax of T, such that the changes are subtle enough that a human cannot easily detect an intent to fool the system.

To systematically find such perturbed $\tilde{T}$ or to find out if $\mathcal{F}(\cdot)$ is linguistically sensitive to $\mathcal{G}(\cdot)$, therefore, we formulate this task as an optimization problem with an objective function as follows.

Metric	Δ	t	p(**one-sided**)
Tokens	1.59	30.70	0.0000
Phonemes	6.26	26.86	0.0000
Readability	0.17	2.12	0.0172
Token PPL	27.43	6.46	0.0000
Phoneme PPL	0.00	0.17	0.4345

Table 2: Results of independent two-sample t-tests comparing spoof and bona-fide items on ASVSpoof 2019 training data statistics. PPL is the perplexity.

Objective Function

Given a transcript $T{=}\{w_1, w_2, \ldots, w_M\}$, a target AAS $\mathcal{F}$, and a TTS model $\mathcal{G}$, our goal is to find an alternative transcript $\tilde{T}$ by *minimally perturbing* T, or:

$$\tilde{T}^* = \arg\min_{\tilde{T}} \text{distance}(T, \tilde{T}^*) \text{ s.t.}$$
$$\text{SIM}(T, \tilde{T}){=}{=}1,\ \mathcal{F}(\mathcal{G}(\tilde{T})){=}{=}\tilde{Y}, \text{and } \tilde{Y}{\neq}Y$$

4 Method

To solve the introduced optimization problem, we adapt the adversarial attack framework in adversarial NLP literature to design an adversarial transcript perturbation framework that exploits the linguistic sensitivity of the AASs. The overall algorithm is formalized in Alg. 1 (Appendix).

Step 1: Finding Important Words. First, we want to measure how sensitive $\mathcal{F}(\cdot)$'s prediction is to each word, allowing us to prioritize perturbations to the most influential locations in the transcript. To do this, given a transcript $T{=}\{w_1, w_2, \ldots, w_m\}$, we estimate the impact of each word w_i on the anti-spoofing prediction (Lines 3-5). For each word position i, we synthesize audio without w_i in the transcript and then compute the bona-fide probability $p_i{=}\mathcal{F}(\mathcal{G}(T_{\setminus w_i}))$. Then, we aggregate all masked-transcript candidates $\mathcal{W}{=}\{T_{\setminus w_1}, ..., T_{\setminus w_m}\}$ and sort them by the descending impact scores p_i (Line 6, Alg. 1).

Step 2: Greedy Word Perturbations. Next, our algorithm iteratively attempts to find effective word substitutions beginning with the positions with the largest impact, potentially minimizing the number of words needed to perturb. For each candidate position, we use a *Search*$(\cdot)$ to propose a set of replacement candidates (Line 9). To search for the replacement $\tilde{w}_i$ for w_i, *Search*$(\cdot)$ first finds a list of replacement candidates by utilizing either WordNet to find synonym replacements (Ren et al., 2019),

or a Masked Language Model (Devlin et al., 2019) that masks the word w_i as the token $[MASK]$ and performs the token predictions (Jin et al., 2020). For each substitution candidate c_k, we then construct a new transcript T' by replacing w_i in $\tilde{T}$ with c_k (Line 10). The new transcript is then synthesized into speech using $\mathcal{G}(T')$ and evaluated by the anti-spoofing model to obtain an updated bona-find prediction score p' (Line 11). The best replacement candidate c_k is one that best maximizes p' while satisfying all constraints listed in Eq. (3) (Line 12-15, Alg. 1).

Step 3: Semantic and Syntax Preservation. We factorize the boolean check $\text{SIM}(T, \tilde{T})$ (Eq. 3) to (1) syntactic and (2) semantic preservation. To check for syntactic preservation, we only accept a replacement $\tilde{w}_i$ only if its part-of-speech (POS) function in $\tilde{T}$ preserves that of w_i in the original transcript T. To check for semantic preservation, we ensure that the cosine similarity, denoted as $\cos(\cdot)$, between semantic vectorized representations of T and the current $\tilde{T}$ embed using the popular Universal Sentence Encoder (USE) (Cer et al., 2018), denoted as $f_{\text{embed}}(\cdot)$, is at least δ threshold:

$$\cos\left(f_{\text{embed}}(T), f_{\text{embed}}(\tilde{T})\right) \geq \delta \qquad (2)$$

Overall, our framework is *model-agnostic* and does *not* require access to internal parameters or gradients of the target AAS, making it applicable in practical black-box settings where we only have query access to the AAS. By leveraging linguistic variability at the transcript level while enforcing functional constraints, the proposed method is able to systematically probe and exploit the linguistic sensitivity of end-to-end audio anti-spoofing detectors.

5 Experiments

5.1 Set-up

Datasets. We evaluate our algorithm on a total of 1,439 test transcripts from the deepfake speech VoiceWukong dataset (Yan et al., 2024), statistics of which is provided in Table A2. VoiceWukong is constructed based on the English VCTK dataset (Yamagishi et al., 2019) and is released under the Creative Commons Attribution-NonCommercial 4.0 International Public License, with which we comply by using the data exclusively for research purposes. To better reflect real-world attack scenarios, we restrict our evaluation to transcripts having at least 10 tokens.

Anti-spoofing Detectors. We use three open-source anti-spoofing models: AASIST-2 (Tak et al., 2022), CLAD (Wu et al., 2024), RawNet-2 (Tak et al., 2021), and two commercially available deepfake speech detection APIs, which will be refereed using pseudonyms API-A and API-B.

Table A3 presents the precision and recall scores for bona-fide and spoof prediction across all models. Notably, AASIST-2 achieves the most balanced and robust detection, with high precision and recalls. In contrast, CLAD and RawNet-2 show comparatively lower and more variable performance. Commercial detectors exhibit much lower bona-fide recall, indicating a tendency to misclassify legitimate speech as spoofed.

Text-to-Speech Models. We employ Kokoro TTS, a lightweight high-quality, and community-known model, capable of generating $10K$ audio in only 832 seconds on an NVIDIA A100 GPU. For voice cloning TTS, we use Coqui TTS (having over 39K stars on GitHub) to replicate the voices of four well-known individuals. Additionally, we evaluate F5 TTS (Chen et al., 2024), a recently proposed SOTA model with best-in-class generation quality. For commercial TTS, we employ OpenAI TTS due to its popularity and low cost.

Word Perturbation Methods. We adapt four word perturbation strategies for Step 2 of Alg. 1 and TextAttack framework (Morris et al., 2020), including PWWS (Ren et al., 2019), which swaps words using WordNet; TextFooler (Jin et al., 2020), which substitutes words based on contextual word embeddings while respecting part-of-speech and filtering stop words; BAE (Garg and Ramakrishnan, 2020), which leverages BERT to propose plausible replacements; and BERTAttack (Li et al., 2020), which also uses BERT to generate adversarial substitutions. These strategies cover most of the word perturbation methods in adversarial NLP literature.

Metrics. We report the Original Accuracy (OC), Accuracy Under Attack (AUA) of the target AAS on the synthesized spoof samples. We also measure Attack Success Rate (ASR) or the percentage of spoof audio out of the tested transcripts that were able to flip the original correct spoof predictions of the target AAS detector. We also report the semantic preservation score (COS) calculated via Eq. (2) and standardized to 0-100% scale. Intuitively, the higher the ASR, the lower AUA, and the higher the COS, the better an attack is able to preserve the original transcripts' meaning.

Text-to-Speech	AASIST-2				CLAD				RawNet-2			
	OC↑	AUA↓	ASR↑	COS↑	OC↑	AUA↓	ASR↑	COS↑	OC↑	AUA↓	ASR↑	COS↑
Kokoro (British Male)	95.1%	39.9%	**58.1%**	90.5%	96.9%	43.9%	54.7%	91.7%	100.0%	98.9%	1.1%	89.5%
Kokoro (British Female)	92.4%	30.4%	**67.1%**	91.3%	97.7%	62.4%	36.2%	90.4%	92.4%	56.9%	38.4%	90.4%
Kokoro (American Male)	90.4%	37.9%	**58.0%**	91.1%	83.3%	43.3%	56.6%	91.5%	100.0%	100.0%	0.0%	88.2%
Kokoro (American Female)	92.2%	32.6%	64.6%	91.0%	98.3%	32.6%	66.9%	92.1%	83.2%	20.3%	**75.6%**	93.3%
Coqui (Donald Trump)	85.5%	25.7%	**69.9%**	93.2%	98.9%	62.0%	37.3%	92.3%	99.8%	88.4%	11.4%	90.2%
Coqui (Elon Musk)	98.0%	75.9%	22.5%	90.9%	98.4%	62.1%	**36.9%**	91.1%	100.0%	99.9%	0.1%	89.7%
Coqui (Taylor Swift)	94.4%	17.0%	**82.0%**	92.9%	95.1%	20.0%	79.0%	92.8%	99.7%	88.2%	11.5%	89.5%
Coqui (Oprah Winfrey)	98.9%	79.0%	**20.1%**	91.5%	99.7%	99.7%	0.0%	90.5%	95.8%	86.2%	9.9%	91.0%
F5 (Male)	88.5%	33.4%	62.2%	92.8%	93.2%	7.9%	**91.6%**	94.6%	99.4%	78.0%	21.5%	90.7%

Table 3: Open-source model results. Bold values indicate the TTS-voice pair that is most effective at attacking (ASR) each detector model.

We refer the readers to the Appendix for additional implementation details.

5.2 Results

5.2.1 Attacking Open-Source Detectors

Table 3 summarizes the average performance of three open-source anti-spoofing detectors: AASIST-2, CLAD, and RawNet-2 under adversarial attacks on synthetic speech generated from a variety of TTS models across four word perturbation strategies (PWWS, TextFooler, BAE, and BERTAttack), totaling 108 experiments.

Overall Linguistic Sensitivity. All three detectors show a marked reduction in AUA, suggesting that adversarially perturbed transcripts can noticeably degrade anti-spoofing performance while consistently maintaining high semantic preservation close to or higher than 90%. Consequently, ASR is often substantial, reaching as high as 82%, specially for certain voice profiles.

Voice Gender Effect. Overall, female voices exhibit a higher ASR than male voices across both detectors and TTS systems. For example, for Kokoro TTS voices, British Female and American Female identities consistently yield higher ASRs than their male counterparts, often accompanied by sharply lower AUA. This implies that spoof female voices are more prone to become undetected under linguistic adversarial manipulations.

Notable Exceptions. Some voice profiles are notably resistant to attack. For instance, Coqui TTS voice for Oprah Winfrey shows almost zero ASR on both CLAD (0.02%), but this phenomenon does not repeat with other detectors. Similarly, the RawNet-2 detector demonstrates strong robustness to some male voice profiles, such as Kokoro TTS (British Male and American Male) and Coqui Elon Musk voice cloning, where the ASR only reaches (1.06% and 0.00%) and 0.14%, respectively, indicating that linguistic sensitivity of an AAS is TTS-specific and *some detector-voice combinations are far less susceptible to transcript-based attacks*. This also validates our AAS-TTS pair linguistic sensitivity formulation in Sec. 3. We later show that these voice-detector combinations have nearly perfect Audio Encoder Similarity (Fig. 2), meaning that audio encoders of the TTS and the detector are more or less encoding similar information.

	OC↑	AUA↓	ASR↑	COS↑
API-A - Coqui	100.0%	98.0%	2.0%	85.7%
API-A - F5	99.0%	70.0%	29.3%	86.2%
API-A - Kokoro	100.0%	74.0%	26.0%	84.1%
API-A - OpenAI	95.0%	32.0%	66.3%	89.3%
API-B - Coqui	100.0%	100.0%	0.0%	87.0%
API-B - F5	100.0%	100.0%	0.0%	87.3%
API-B - Kokoro	100.0%	100.0%	0.0%	80.8%
API-B - OpenAI	100.0%	96.0%	4.0%	86.3%
CLAD - OpenAI	86.0%	4.0%	95.3%	93.4%

Table 4: Commercial Anti-spoofing Detectors Results

5.2.2 Attacking Commercial Detectors

Table 4 presents the attack results on commercial AASs when paired with both commercial and non-commercial TTS models. To conserve API usage and cost, each experiment applies only the strongest attack method identified in prior experiments (TextFooler), and evaluates 100 items that were randomly sampled to maintain the same length distribution as the main test set. For each TTS-detector pair, we attack the voice profile with the highest original accuracy (OC) to demonstrate the lower bound effectiveness in the hardest-case scenario. For OpenAI's TTS, we choose CLAD which has the highest original accuracy among the open-source models.

For API-A, we observe a substantial drop in de-

tection accuracy under attack when pairing with most TTS models except Coqui. While its OC is nearly perfect across all voices, adversarial attack reduces AUA to as low as 32% when paired with OpenAI TTS, resulting in a high attack success rate (ASR) of 66.3%. Notably, API-A is vulnerable when tested with realistic, high-quality TTS synthesis. API-B, in contrast, retains perfect detection (AUA = 100%) for Coqui, F5, and Kokoro TTS, and only exhibits a minor decrease (AUA = 96%, ASR = 4%) with OpenAI TTS. However, Table A3 reveals this robustness is partly due to a strong bias toward labeling all samples as spoof, with poor bona-fide recall and moderate spoof precision. For the open-source CLAD model evaluated on OpenAI TTS, adversarial attack drops the accuracy from 86% to just 4%, yielding the highest ASR (95.3%) among all tested scenarios. These findings highlight the concerning vulnerability to linguistic sensitivity of commercial detectors faced when with high-fidelity synthetic speech.

6 Feature Analysis

Beyond providing empirical validation on our initial hypothesis that audio anti-spoof detectors are sensitive to small linguistic changes in the audio' underlying transcripts, this section aims to investigate and analyze *what factors and how much they associate with varying anti-spoofing detectors' decisions under adversarial attacks*. Particularly, we extract **linguistic, acoustic, and model-level features** from 108 open-source attack experiments, utilize logistic regression analysis, and train predictive models to estimate the bona-fide probability of perturbed inputs. Formulations of all features are provided in the Appendix.

6.1 Feature Engineering

We first seek to understand how **linguistic features (LF)** at transcript-level shift under adversarial attacks.

LF1. Perturbed Percentage measures the fraction of modified words in a transcript; higher values indicate more extensive lexical changes.

LF2. Readability Difference quantifies the change in reading comprehension difficulty between the original and perturbed transcripts using the Dale-Chall Readability Score.

LF3. Semantic Similarity assesses the similarity in meaning between the original and perturbed transcripts using Universal Sentence Encoder embeddings, or the COS evaluation metric.

	coef	std err	z	P>|z|
Perturbed Percentage	-0.3507	0.009	-39.31	0.000
Δ Readability	0.0352	0.007	4.69	0.000
Δ PPL	0.0758	0.009	8.48	0.000
Δ Tree Depth	0.0125	0.007	1.76	0.077
Δ Duration	0.0748	0.008	8.89	0.000
DTW Distance	0.2063	0.008	26.73	0.000
Δ Phoneme PPL	-0.0112	0.007	-1.54	0.122
Δ Content Enjoyment	0.0536	0.013	4.18	0.000
Δ Content Usefulness	0.0679	0.021	3.17	0.001
Δ Production Complexity	0.0110	0.010	1.12	0.264
Δ Production Quality	0.0307	0.017	1.82	0.069
Audio Encoder Similarity	-0.9013	0.011	-81.60	0.000
Spoof F1	3.1254	0.159	19.60	0.000
Bona-fide F1	-2.5911	0.159	-16.25	0.000

Table 5: Logistic regression feature analysis for bona-fide detection on adversarial samples. Δ is the difference and Semantic Similarity feature is removed due to the high Variance Inflation Factor to avoid multicollinearity.

LF4. Perplexity Difference measures the change in perplexity between the original and perturbed transcripts.

LF5. Syntactic Complexity Difference measures the change in maximum syntactic tree depth between the original and perturbed transcripts.

Given that text-level changes can propagate to measurable differences at the acoustic level, we further investigate how variations in several **acoustic features (AF)** contribute to the performance of anti-spoofing detectors.

AF1. DTW Distance utilizes Dynamic Time Warping to measure the alignment cost between the mel spectrograms of the original and perturbed audio.

AF2. Duration Difference captures the difference in audio length.

AF3. Phoneme Perplexity Difference measures the corresponding change in phoneme sequence perplexity, calculated via the CharsiuG2P (Zhu et al., 2022) T5-based model.

AF4. Aesthetics Difference measures the shifting aesthetics computed by Meta Audiobox Aesthetics (Tjandra et al., 2025) which includes four automatic quality assessment measures: Content Enjoyment (CE), Content Usefulness (CU), Production Complexity (PC), Production Quality (PQ).

For **model-level features (MF)**, we propose:

MF1. Audio Encoder Similarity(AES) metric quantifies how closely synthesized audios of the same voice cluster in the detector's representation

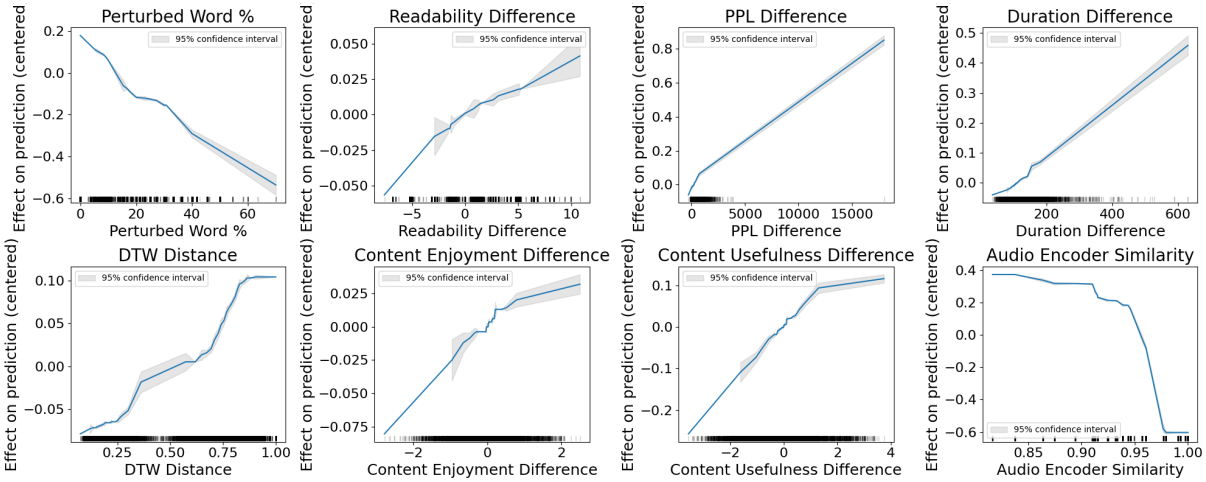

Figure 2: Feature impact on bona-fide probability prediction. A positive effect means the feature increases the likelihood of a perturbed item being classified as bona-fide.

space. A high AES score indicates that the detector perceives all TTS-generated audio for a given voice as acoustically similar, or being able to capture them as originally from the same voice profile, which may enhance robustness against transcript-level adversarial attacks.

MF2, MF3. Spoof and bona-fide F1. Additionally, we include spoofed and bona-fide F1 scores (Table A3) as model-level features to analyze how pre-existing biases influence behavior under adversarial attacks. Notably, if these features can predict attack outcomes, they are especially useful because they can be computed **before** any adversarial perturbation, guiding the selection or development of more robust anti-spoofing models.

6.2 Analysis Results

Table 5 presents the results of a logistic regression analysis predicting the bona-fide probability for adversarial audio samples using the engineered features.

Linguistic Features Impact. Several features display significant associations with the detector's response to adversarial perturbations with statistical significance. Notably, the proportion of perturbed words in a transcript is negatively correlated with bona-fide detection, indicating that increasing lexical modifications decreases the likelihood that the detector classifies the input as bona-fide. Syntactic complexity differences are less significant, indicating that deep syntactic restructurings are less impactful than surface-level wording and fluency changes. PPL difference and readability difference are both positively correlated with bona-fide probability. The trends in Fig. 2 suggests that when the perturbed transcript exhibits greater linguistic complexity than the original, the adversarial sample is more likely to be classified as bona-fide. This leads to an assumption that the disparities in linguistic features between spoofed and bona-fide training samples (Table A1) might have introduced linguistic vulnerabilities that can be exploited by adversarial attack algorithms.

Acoustic Features Impact. The DTW distance between mel-spectrograms and the duration difference indicate that greater spectral or temporal deviations between original and perturbed audio samples are associated with higher bona-fide probabilities (trends in Fig. 2). In contrast, the effect of phoneme perplexity difference is not statistically significant, suggesting that changes in phoneme-level predictability are less associated with variations in acoustic realization, such as spectral and durational differences. The positive correlations observed for Content Enjoyment and Content Usefulness suggest that enhanced emotional and artistic qualities in perturbed audio may increase its likelihood of deceiving anti-spoofing detectors.

Model-Level Features Impact. AES provides the strongest predictive signals for susceptibility to adversarial attacks. AES is negatively associated with bona-fide prediction, implying that models that produce highly clustered audio embeddings for a given TTS and voice are less likely to recognize perturbed inputs as bona-fide. Notably, as shown in Fig. 2, when AES approaches 100%, there is a significant reduction in the likelihood of attack success. Additionally, the original detector F1 scores on

	Transcript	Bona-fide
Fraud:	Anne, I need to ~~be~~**become** direct. . . . I need your help ~~immediately~~**succinctly**.	0.2 → **69.7**
Victim:	Brad? What is it? You sound serious.	N/A
Fraud:	I'm in the ~~hospital~~**convalescent**. It's serious. ~~Kidney~~**Liver** cancer. They ~~need~~**crucial** to . . .	0.2 → **62.4**
Victim:	Cancer? Oh no, Brad, I'm so sorry to hear that. What kind of problem with funds? Don't you?	N/A
Fraud:	My accounts are frozen. . . . the ~~courts~~**judiciary** have tied up everything. . . ~~hospital~~**outpatient** bill.	0.4 → **90.3**
Fraud:	The doctors need ~~payment~~**reimbursement** now to proceed with this vital step. . . .	0.3 → **73.1**
Victim:	Me? But I... I'm not a millionaire, Brad. How much do they need?	N/A
Fraud:	. . . It's 830,000 euros. I ~~know~~**understand** it's a huge ask, Anne, but my life could depend on this.	0.7 → **58.2**
Victim:	830,000 euros?! That's an enormous sum. But how would I even? And where would I send it?	N/A

Table 6: Illustration of how adversarial transcript attacks on Brad Pitt voice cloning scam enable the attacker to significantly undermine a commercial audio anti-spoof detector.

Model	Precision	Recall	F1 Score
Logistic Regression	66.82%	71.54%	69.10%
XGBoost	74.70%	78.39%	76.50%
Random Forest	64.28%	76.88%	70.02%
SVM (poly kernel)	46.83%	97.00%	63.16%
LightGBM	72.98%	77.00%	74.94%
AdaBoost	70.30%	77.82%	73.87%

Table 7: Performance of approximating anti-spoof detector's bona-fide prediction of various predictive models built on our engineered features.

bona-fide and spoofed samples have the highest-magnitude coefficients, indicating that initial model bias in spoof discrimination translates directly into vulnerability or robustness under adversarial conditions.

Table 7 further shows how predictive models built on our engineered features can effectively approximate the outcomes of the detector's decisions. Notably, XGBoost achieves an F1 score of up to 76.50%, with several other models performing in the 69% to 74% range. This shows that these models can enable the development of grey-box or black-box adversarial attacks where attacker access to the actual detector is restricted or limited. By optimizing transcript modifications based on proxy predictions from the predictive models, adversaries can effectively attack audio anti-spoofing systems even without full transparency of the detector, underscoring the urgent need for more robust and resilient defenses.

7 Case Study: Deepfake Voice Cloning

Table 6 presents a simulated case study adapting from the recent, notorious Brad Pitt impersonation scam where adversarial perturbations are applied to the fraudster's dialogue generated by ChatGPT and synthesized using the SOTA F5 TTS voice-cloned model. The transcripts illustrate the original and perturbed words, with corresponding bona-fide detection probabilities reported from API-A commercial anti-spoofing detector. In all tested exchanges, the unperturbed fraud utterances are assigned extremely low bona-fide probabilities ($< 1\%$), suggesting effective spoof detection. However, after targeted adversarial perturbation of key lexical items (e.g., "be"→"become", "hospital"→"convalescent"), the bona-fide probability rises dramatically, with post-attack scores ranging from 58.2% to as high as 90.3%. Notably, even minimal lexical changes can evade commercial detectors, flipping the label from clear spoof to likely bona-fide.

These findings demonstrate the concerning real-world risks of transcript-level adversarial attacks in voice cloning scenarios, highlighting the urgency for developing more robust anti-spoofing mechanisms that can withstand subtle semantic and lexical manipulations.

8 Conclusion

This work demonstrate that SOTA audio anti-spoofing systems are vulnerable to transcript-level linguistic nuances. By systematically applying semantic preserving perturbations to transcripts, we show that even subtle linguistic changes can significantly degrade detection accuracy in both open-source and commercial deepfake detectors. Our experiments and feature analyses reveal that these vulnerabilities are driven by both linguistic complexity and characteristics of the model's learned audio representations. This underscores the need for anti-spoofing systems to consider linguistic variation, not just acoustics. For future work, we plan to further investigate the interplay between model architecture, training data, and linguistic features that contribute to adversarial susceptibility, with the goal of guiding more comprehensive and resilient detection strategies.

Limitation

One limitation of our work is the lack of experimentation on false positive cases, such as those involving non-native speakers who may stutter or use incorrect wording during conversations. These effects can act as natural adversarial attacks on the transcript and potentially reduce bona-fide detection accuracy. Additionally, vocalizations such as laughter, giggling, and chuckling, which may enhance the enjoyment and naturalness of generated audio, are not addressed in this study; these elements could also serve as another modality for transcript-based adversarial attacks.

Our experiments are primarily limited to English-language data, leaving open the question of how linguistic attacks generalize to other languages and multilingual TTS systems. Diverse syntactic and morphological structures in non-English languages may uniquely impact anti-spoofing system robustness, which remains unexplored in this work.

Furthermore, although the linguistic perturbations in our experiments are constrained to retain semantic meaning, we do not measure their detectability by humans or plausibility in real conversational contexts. User studies are needed to assess whether such adversarial modified transcripts sound unnatural or prompt suspicion among human listeners.

Broader Impacts and Potential Risks

By uncovering vulnerabilities related to linguistic perturbations, our findings encourage audio anti-spoofing research to move beyond acoustic analysis and incorporate linguistic robustness into system design and evaluation. This insight can directly inform the development of safer, more resilient voice authentication and verification products.

Our methodology highlights the importance of adversarial testing and "red teaming" in the responsible development of AI security systems. This proactive approach enables the community to identify and mitigate attacks before they are exploited in practice, ultimately safeguarding critical voice-driven infrastructure.

This research is conducted to advance audio security and raise awareness of vulnerabilities in current anti-spoofing systems. The authors are committed to promoting social good and responsible AI development, with no intention of enabling any malicious or unethical applications of these findings.

However, by publicly revealing specific attack strategies and demonstrating their effectiveness, our work could inadvertently lower the barrier for malicious actors to evade anti-spoofing systems. Additionally, making our adversarial techniques and code openly available—while essential for reproducibility and further research—introduces the risk that these methods might be misused for fraudulent purposes.

References

Rohit Arora, Anmol Arora, and Rohit Singh Rathore. 2022. Impact of channel variation on one-class learning for spoof detection. *Preprint*, arXiv:2109.14900.

Luigi Attorresi, Davide Salvi, Clara Borrelli, Paolo Bestagini, and Stefano Tubaro. 2022. Combining automatic speaker verification and prosody analysis for synthetic speech detection. *Preprint*, arXiv:2210.17222.

Zhongjie Ba, Qing Wen, Peng Cheng, Yuwei Wang, Feng Lin, Li Lu, and Zhenguang Liu. 2023. Transferring audio deepfake detection capability across languages. In *Proceedings of the ACM Web Conference 2023*, WWW '23, page 2033–2044, New York, NY, USA. Association for Computing Machinery.

Daniel Cer, Yinfei Yang, Sheng yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. 2018. Universal sentence encoder. *Preprint*, arXiv:1803.11175.

Yushen Chen, Zhikang Niu, Ziyang Ma, Keqi Deng, Chunhui Wang, Jian Zhao, Kai Yu, and Xie Chen. 2024. F5-tts: A fairytaler that fakes fluent and faithful speech with flow matching. *Preprint*, arXiv:2410.06885.

Chantal Da Silva. 2024. French woman falls victim to online romance scam by ai brad pitt, report says. https://www.nbcnews.com/news/world/ai-brad-pitt-woman-romance-scam-france-tf1-rcna187745. NBC News, Accessed: 2024-06-13.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *Preprint*, arXiv:1810.04805.

Siddhant Garg and Goutham Ramakrishnan. 2020. BAE: BERT-based adversarial examples for text classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6174–6181, Online. Association for Computational Linguistics.

Di Jin, Zhijing Jin, Joey Tianyi Zhou, and Peter Szolovits. 2020. Is bert really robust? a strong baseline for natural language attack on text classification and entailment. *Preprint*, arXiv:1907.11932.

Jee-weon Jung Jung, Hee-Soo Heo, Hemlata Tak, Hye jin Shim, Joon Son Chung, Bong-Jin Lee, Ha-Jin Yu, and Nicholas Evans. 2021. Aasist: Audio anti-spoofing using integrated spectro-temporal graph attention networks. *Preprint*, arXiv:2110.01200.

Thai Le, Jooyoung Lee, Kevin Yen, Yifan Hu, and Dongwon Lee. 2022. Perturbations in the wild: Leveraging human-written text perturbations for realistic adversarial attack and defense. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2953–2965, Dublin, Ireland. Association for Computational Linguistics.

Linyang Li, Ruotian Ma, Qipeng Guo, Xiangyang Xue, and Xipeng Qiu. 2020. BERT-ATTACK: Adversarial attack against BERT using BERT. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6193–6202, Online. Association for Computational Linguistics.

So-jeong Lim Lim, Adam Jatowt, and Masatoshi Yoshikawa. 2018. Understanding characteristics of biased sentences in news articles. In *CIKM Workshops*.

Florian Lux, Sarina Meyer, Lyonel Behringer, Frank Zalkow, Phat Do, Matt Coler, Emanuël A. P. Habets, and Ngoc Thang Vu. 2024. Meta learning text-to-speech synthesis in over 7000 languages. *Preprint*, arXiv:2406.06403.

John Morris, Eli Lifland, Jin Yong Yoo, Jake Grigsby, Di Jin, and Yanjun Qi. 2020. Textattack: A framework for adversarial attacks, data augmentation, and adversarial training in nlp. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 119–126.

Nicolas M. Müller, Philip Sperl, and Konstantin Böttinger. 2023. Complex-valued neural networks for voice anti-spoofing. *Preprint*, arXiv:2308.11800.

Uwe Peters Peters. 2024. The philosophical debate on linguistic bias: A critical perspective. *Philosophical Psychology*, 37(6):1513–1538.

Shuhuai Ren, Yihe Deng, Kun He, and Wanxiang Che. 2019. Generating natural language adversarial examples through probability weighted word saliency. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1085–1097, Florence, Italy. Association for Computational Linguistics.

Gil Shomron and Uri Weiser. 2020. Post-training batch-norm recalibration. *Preprint*, arXiv:2010.05625.

Signicat. 2024. Fraud attempts with deepfakes have increased by 2137% over the last three years. https://www.signicat.com/press-releases/fraud-attempts-with-deepfakes-have-increased-by-2137-over-the-last-three-year. Signicat, Accessed: 2024-06-13.

Hemlata Tak, Jose Patino, Massimiliano Todisco, Andreas Nautsch, Nicholas Evans, and Anthony Larcher. 2021. End-to-end anti-spoofing with rawnet2. *Preprint*, arXiv:2011.01108.

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. *Preprint*, arXiv:2202.12233.

Andros Tjandra, Yi-Chiao Wu, Baishan Guo, John Hoffman, Brian Ellis, Apoorv Vyas, Bowen Shi, Sanyuan Chen, Matt Le, Nick Zacharov, Carleigh Wood, Ann Lee, and Wei-Ning Hsu. 2025. Meta audiobox aesthetics: Unified automatic quality assessment for speech, music, and sound. *Preprint*, arXiv:2502.05139.

Adaku Uchendu, Thai Le, and Dongwon Lee. 2023. Attribution and obfuscation of neural text authorship: A data mining perspective. *ACM SIGKDD Explorations Newsletter*, 25(1):1–18.

Haibin Wu, Songxiang Liu, Helen Meng, and Hung yi Lee. 2020. Defense against adversarial attacks on spoofing countermeasures of asv. *Preprint*, arXiv:2003.03065.

Haolin Wu, Jing Chen, Ruiying Du, Cong Wu, Kun He, Xingcan Shang, Hao Ren, and Guowen Xu. 2024. Clad: Robust audio deepfake detection against manipulation attacks with contrastive learning. *Preprint*, arXiv:2404.15854.

Junichi Yamagishi, Christophe Veaux, and Kirsten MacDonald. 2019. CSTR VCTK Corpus: English Multi-speaker Corpus for CSTR Voice Cloning Toolkit (version 0.92), [sound]. `https://doi.org/10.7488/ds/2645`. University of Edinburgh. The Centre for Speech Technology Research (CSTR).

Ziwei Yan, Yanjie Zhao, and Haoyu Wang. 2024. Voicewukong: Benchmarking deepfake voice detection. *Preprint*, arXiv:2409.06348.

Jian Zhu, Cong Zhang, and David Jurgens. 2022. Byt5 model for massively multilingual grapheme-to-phoneme conversion. *Preprint*, arXiv:2204.03067.

	Male Voice		Female Voice	
	Spoof	**Bona-fide**	**Spoof**	**Bona-fide**
$\mathbb{E}[Tokens]$	7.85	7.08	9.09	6.92
$\mathbb{E}[Phonemes]$	30.09	27.26	35.56	26.89
$\mathbb{E}[Readability]$	6.27	6.26	6.89	6.60
$\mathbb{E}[TokenPPL]$	100.12	96.18	94.58	96.20
$\mathbb{E}[PhonePPL]$	1.0461	1.0458	1.0460	1.0460

Table A1: ASVSpoof 2019 LA statistics. All values is statistically significant ($p - value < 0.001$). $\mathbb{E}[\cdot]$ is the average value of that metric, and PPL is perplexity.

Feature	**Value**
Average Tokens	11.05
Average Phonemes	42.46
Average Readability	7.28
Token PPL	43.86
Phoneme PPL	1.0497

Table A2: VoiceWukong dataset features such as readability and perplexity.

A Dataset statistics

A.1 ASVSpoof 2019 statistics

Table A1 summarizes the key linguistic and structural statistics of the ASVSpoof 2019 LA training dataset, segmented by speaker gender (male and female) and ground-truth label (spoof vs. bona-fide). For each group, we report the average number of tokens and phonemes per utterance, the average readability score (which reflects the linguistic complexity of the transcripts), and perplexity values computed at both the token and phoneme levels. Notably, all reported values are statistically significant (p-value < 0.001), suggesting consistent differences between spoof and bona-fide samples across these linguistic features. These statistics provide critical insight into the dataset composition, which may influence both the performance and the generalization capacity of anti-spoofing models during training and evaluation.

A.2 VoiceWukong dataset statistics

Table A2 presents key statistical features of the VoiceWukong dataset used in our experiments. In addition to average transcript length in tokens and phonemes, we report average readability scores, which provide an estimate of the linguistic complexity of the dataset's transcripts, as well as token-level and phoneme-level perplexity (PPL), which serve as measures of sequence unpredictability. These features offer a comprehensive overview of the structural and linguistic properties of the evaluation data, and help contextualize the challenges involved for both TTS synthesis and anti-spoofing detection.

B Equations and Results

B.1 Implementation Details

For open-source detectors, instead of fine-tuning models for each specific TTS-generated voice, we adapt them using batch normalization calibration (Shomron and Weiser, 2020) on a small set that does not overlap the evaluation data, which shifts the mean and variance of the feature distributions to better match those of the current TTS system until detection accuracy exceeds 90%. We argue that retraining on every possible voice is infeasible, given the potentially over 8 billion unique speakers worldwide; however, these voices are likely to share similar statistical properties in their acoustic features.

B.2 Anti-spoof detection performance on VoiceWukong dataset

Table A3 compares the performance of various anti-spoofing detectors on the VoiceWukong dataset, reporting both precision and recall for bona-fide and spoofed audio. We evaluate two commercial APIs (API-A and API-B) alongside several state-of-the-art open-source models: RawNet-2, CLAD, and AASIST-2. The results reveal considerable variation in performance across different systems. Notably, AASIST-2 achieves the highest and most balanced precision and recall scores for both bona-fide and spoofed classes, suggesting superior generalization capability. In contrast, the commercial detectors—especially API-B—exhibit strong bias, with high spoof recall but low bona-fide recall, indicating a tendency to label most samples as spoofed. These findings highlight the strengths and limitations of existing anti-spoofing solutions on challenging synthetic datasets, and motivate the need for further robustness against linguistic and generative variation.

B.3 Transcript-level Adversarial Attack Algorithm

Our transcript-level adversarial attack Algorithm 1 identifies the most influential words in a target transcript and greedily substitutes them—using synonym replacement or masked language models—with alternatives that maximize the chance

Detector	Bona-fide		Spoof	
	Precision	Recall	Precision	Recall
API-A	80.9%	63.0%	58.2%	77.5%
API-B	90.0%	28.1%	46.8%	95.3%
RawNet-2	89.1%	63.0%	61.3%	88.3%
CLAD	88.4%	66.7%	63.4%	86.8%
AASIST-2	90.3%	90.9%	86.1%	85.4%

Table A3: Anti-spoof detection performance comparison.

Algorithm 1 Adversarial Transcript Generation

Input: A transcript $T = \{w_1, w_2, \ldots, w_m\}$, the audio anti-spoofing detection $\mathcal{F}(\cdot)$, a Text-to-Speech model $\mathcal{G}(\cdot)$, $SeachMethod$ and $Constraints$
Output: Adversarial transcript $\tilde{T}$
Identify the impact p_i of a word w_i
for w_i in $w_1, w_2, \ldots, w_m$ **do**
    $p_i \leftarrow \mathcal{F}(\mathcal{G}(T_{\setminus w_i}))$
end for
$\mathcal{W} \leftarrow \{T_{\setminus w_1}, T_{\setminus w_2}, \ldots, T_{\setminus w_m}\}$, sorted by descending values of p_i
$\tilde{T} \leftarrow T$, $\tilde{p} \leftarrow \mathcal{F}(\mathcal{G}(\tilde{T}))$
for $T_{\setminus w_i}$ in $\mathcal{W}$ **do**
    Candidates $\leftarrow$ Search$(T_{\setminus w_i})$
    for c_k in Candidates **do**
        $T' \leftarrow$ Replace c_k with w_i in $\tilde{T}$
        $p' \leftarrow \mathcal{F}(\mathcal{G}(T'))$
        if SIM$(T, \tilde{T})$ **AND** $p' > \tilde{p}$ **then**
            $\tilde{T} \leftarrow T'$, $\tilde{p} \leftarrow p'$
        end if
    end for
end for
return $\tilde{T}$

of misclassification by the anti-spoofing system, while ensuring both semantic and syntactic fidelity through embedding similarity and part-of-speech checks. This model-agnostic, black-box framework exploits the linguistic sensitivity of audio anti-spoofing systems without requiring access to internal model parameters, demonstrating the practical risks posed by transcript-level adversarial perturbations.

B.4 Additional Results

We provide experimental results for AASIST-2 in Table A4, CLAD in Table A5, and Rawnet-2 in Table A6.

Voice	Method	OC	AUA	ASR	COS
Donald Trump	BAE	85.5	27.3	68.0	93.3
Donald Trump	BertAttack	85.5	21.6	74.7	93.7
Donald Trump	PWWS	85.5	35.7	58.2	94.4
Donald Trump	TextFooler	85.5	18.3	78.6	91.3
Elon Musk	BAE	98.0	78.3	20.1	91.7
Elon Musk	BertAttack	98.0	76.3	21.4	92.3
Elon Musk	PWWS	98.0	79.8	19.1	93.6
Elon Musk	TextFooler	98.0	70.3	28.3	87.5
Oprah Winfrey	BAE	98.9	79.0	20.1	91.8
Oprah Winfrey	BertAttack	98.9	77.1	22.4	92.5
Oprah Winfrey	PWWS	98.9	86.2	12.9	94.1
Oprah Winfrey	TextFooler	98.9	71.9	27.3	88.6
Taylor Swift	BAE	94.4	21.5	77.2	92.7
Taylor Swift	BertAttack	94.4	6.2	93.4	94.2
Taylor Swift	PWWS	94.4	32.9	65.2	93.8
Taylor Swift	TextFooler	94.4	7.5	92.0	91.0
F5 Male	BAE	88.5	32.9	62.8	92.8
F5 Male	BertAttack	88.5	27.4	69.0	93.6
F5 Male	PWWS	88.5	41.5	53.1	94.4
F5 Male	TextFooler	88.5	31.9	64.0	90.3
American Female	BAE	92.2	37.3	59.5	91.7
American Female	BertAttack	92.2	26.9	70.8	92.3
American Female	PWWS	92.2	52.5	43.1	92.6
American Female	TextFooler	92.2	13.9	84.9	87.4
American Male	BAE	90.4	40.3	55.4	92.3
American Male	BertAttack	90.4	35.7	60.5	92.0
American Male	PWWS	90.4	58.6	35.2	93.1
American Male	TextFooler	90.4	17.2	81.0	87.0
British Female	BAE	92.4	39.2	57.6	92.2
British Female	BertAttack	92.4	22.4	75.7	92.1
British Female	PWWS	92.4	47.5	48.5	93.2
British Female	TextFooler	92.4	12.5	86.5	87.8
British Male	BAE	95.1	41.8	56.0	91.5
British Male	BertAttack	95.1	33.8	64.4	91.4
British Male	PWWS	95.1	62.1	34.8	92.3
British Male	TextFooler	95.1	21.9	77.0	86.9

Table A4: Complete experimental results on AASIST-2 detector

B.5 Linguistic Feature Equations

Equation 3: $\rho_{\text{perturbed}}$ quantifies the ratio of words that have been perturbed in the transcript.

$$\rho_{\text{perturbed}} = \frac{\text{N}_{\text{perturbed}}}{\text{N}_{\text{words}}} \tag{3}$$

Equation 4: Δ_{read} measures the change in transcript readability after perturbation.

$$\Delta_{read} = \text{read}_{\text{perturbed}} - \text{read}_{\text{original}} \tag{4}$$

Equation 5: $\text{sim}_{\text{semantic}}$ computes the cosine similarity between the semantic embeddings of the perturbed and original transcripts.

$$\text{sim}_{\text{semantic}} = \text{cosine}(Emb_{\text{perturbed}}, Emb_{\text{original}}) \tag{5}$$

Equation 6: Δ_{PPL} captures the difference in language model perplexity before and after perturbation, as measured by Llama 3.

$$\Delta_{PPL} = \text{PPL}_{\text{perturbed}} - \text{PPL}_{\text{original}} \tag{6}$$

Voice	Method	OC	AUA	ASR	COS
Donald Trump	BAE	98.8	63.7	35.6	91.7
Donald Trump	BertAttack	98.8	48.2	51.3	92.1
Donald Trump	PWWS	99.0	74.2	25.1	93.0
Donald Trump	TextFooler	99.0	46.2	53.6	91.7
Elon Musk	BAE	98.5	69.8	29.1	91.9
Elon Musk	BertAttack	98.1	47.5	51.6	91.4
Elon Musk	PWWS	98.5	80.4	18.4	93.7
Elon Musk	TextFooler	98.5	50.7	48.5	87.5
Oprah Winfrey	BAE	99.7	99.7	0.0	91.3
Oprah Winfrey	BertAttack	99.7	99.7	0.0	91.0
Oprah Winfrey	PWWS	99.7	99.7	0.1	93.3
Oprah Winfrey	TextFooler	99.7	99.7	0.0	86.4
Taylor Swift	BAE	95.1	24.3	74.4	92.9
Taylor Swift	BertAttack	95.1	14.4	84.8	93.6
Taylor Swift	PWWS	95.1	34.5	63.7	93.8
Taylor Swift	TextFooler	95.1	6.6	93.1	91.1
F5 Male	BAE	93.2	9.8	89.5	94.5
F5 Male	BertAttack	93.2	2.9	96.9	95.4
F5 Male	PWWS	93.2	16.6	82.2	94.5
F5 Male	TextFooler	93.2	2.1	97.7	93.9
American Female	BAE	98.3	37.9	61.4	92.3
American Female	BertAttack	98.3	27.3	72.2	92.8
American Female	PWWS	98.3	48.0	51.2	93.8
American Female	TextFooler	98.3	17.1	82.6	89.7
American Male	BAE	99.8	64.5	35.4	91.1
American Male	BertAttack	99.8	58.9	41.0	90.9
American Male	PWWS	34.0	0.0	100.0	98.8
American Male	TextFooler	99.8	49.8	50.1	85.2
British Female	BAE	97.7	67.3	31.1	91.1
British Female	BertAttack	97.7	57.8	40.8	91.0
British Female	PWWS	97.7	77.7	20.5	93.4
British Female	TextFooler	97.7	46.6	52.3	86.0
British Male	BAE	96.9	48.4	50.0	92.2
British Male	BertAttack	96.9	39.2	59.5	91.8
British Male	PWWS	96.9	59.1	39.0	93.6
British Male	TextFooler	96.9	28.9	70.2	89.2

Table A5: Complete experimental results on CLAD detector

Voice	Method	OC	AUA	ASR	COS
Donald Trump	BAE	99.8	88.6	11.2	91.2
Donald Trump	BertAttack	99.8	87.5	12.7	89.8
Donald Trump	PWWS	99.8	91.2	8.6	93.0
Donald Trump	TextFooler	99.8	85.3	14.5	86.3
Elon Musk	BAE	100.0	99.9	0.1	90.8
Elon Musk	BertAttack	100.0	99.9	0.1	91.5
Elon Musk	PWWS	100.0	99.9	0.1	93.1
Elon Musk	TextFooler	100.0	99.8	0.2	85.2
Oprah Winfrey	BAE	95.8	86.7	9.4	91.9
Oprah Winfrey	BertAttack	95.8	84.1	12.2	91.4
Oprah Winfrey	PWWS	95.8	90.8	5.2	93.7
Oprah Winfrey	TextFooler	95.8	83.4	12.9	87.0
Taylor Swift	BAE	99.7	93.3	7.4	92.3
Taylor Swift	BertAttack	99.7	82.1	18.3	87.5
Taylor Swift	PWWS	99.7	94.8	4.9	93.2
Taylor Swift	TextFooler	99.7	81.7	18.1	85.8
F5 Male	BAE	99.4	79.1	20.5	91.3
F5 Male	BertAttack	99.4	69.3	30.3	91.3
F5 Male	PWWS	99.4	86.0	13.6	93.6
F5 Male	TextFooler	99.4	77.7	21.9	86.3
American Female	BAE	83.2	21.9	73.7	93.9
American Female	BertAttack	83.2	16.1	80.6	94.1
American Female	PWWS	83.2	36.1	56.7	94.4
American Female	TextFooler	83.2	7.0	91.6	90.7
American Male	BAE	100.0	100.0	0.0	88.6
American Male	BertAttack	100.0	100.0	0.0	90.6
American Male	PWWS	100.0	100.0	0.0	92.9
American Male	TextFooler	100.0	100.0	0.0	83.5
British Female	BAE	92.4	62.2	32.7	91.8
British Female	BertAttack	92.4	45.4	50.9	91.0
British Female	PWWS	92.4	75.0	18.7	93.2
British Female	TextFooler	92.4	45.1	51.2	85.7
British Male	BAE	100.0	99.4	0.6	90.6
British Male	BertAttack	100.0	96.8	3.2	90.1
British Male	PWWS	100.0	99.9	0.1	92.9
British Male	TextFooler	100.0	99.6	0.4	84.5

Table A6: Complete experimental results on Rawnet-2 detector

Equation 7: $\Delta_{syntactic}$ reflects the change in syntactic tree depth between the perturbed and original transcripts.

$$\Delta_{syntactic} = \text{Depth}_{\text{perturbed}} - \text{Depth}_{\text{original}} \quad (7)$$

B.6 Acoustic Feature Equations

Equation 8: $\text{dtw}_{\text{distance}}$ calculates the dynamic time warping (DTW) distance between the mel-spectrograms of the perturbed and original audio.

$$\text{dtw}_{\text{distance}} = \text{DTW}(mel_{\text{perturbed}}, mel_{\text{original}}) \quad (8)$$

Equation 9: $\Delta_{duration}$ shows the change in duration between the perturbed and original synthesized speech.

$$\Delta_{duration} = D_{\text{perturbed}} - D_{\text{original}} \quad (9)$$

Equation 10: $\Delta_{PhonePPL}$ measures the change in phoneme-level perplexity after transcript perturbation.

$$\Delta_{PhonePPL} = \text{PhonePPL}_{\text{perturbed}} - \text{PhonePPL}_{\text{original}} \quad (10)$$

Equation 11: Δ_{CE}, Δ_{CU}, Δ_{PC}, and Δ_{PQ} represent the changes in various audio aesthetics metrics—clarity, continuity, pronunciation correctness, and prosody quality—due to the perturbation.

$$\begin{aligned} \Delta_{CE} &= \text{CE}_{\text{perturbed}} - \text{CE}_{\text{original}} \\ \Delta_{CU} &= \text{CU}_{\text{perturbed}} - \text{CU}_{\text{original}} \\ \Delta_{PC} &= \text{PC}_{\text{perturbed}} - \text{PC}_{\text{original}} \\ \Delta_{PQ} &= \text{PQ}_{\text{perturbed}} - \text{PQ}_{\text{original}} \end{aligned} \quad (11)$$

B.7 Audio Encoder Similarity Equation

We first extract acoustic embeddings for all original transcripts by the detector. We then compute

the centroid embedding by averaging these embeddings and normalizing the result to unit length:

$$\mathbf{c} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{e}_i, \qquad \hat{\mathbf{c}} = \frac{\mathbf{c}}{|\mathbf{c}|} \tag{12}$$

where $\mathbf{e}_i$ denotes the embedding for the i-th sample and N is the total number of samples in the group.

The Audio Encoder Similarity for the group is then defined as the average cosine similarity between the centroid $\hat{\mathbf{c}}$ and each sample embedding:

$$\text{AES} = \frac{1}{N}\sum_{i=1}^{N}\cos(\hat{\mathbf{e}}_i, \hat{\mathbf{c}}) \tag{13}$$