Diffusion-based Adversarial Attack to Automatic Speech Recognition

Ying Wang WANGYING@NUDT.EDU.CN
Yuchuan Luo LUOYUCHUAN09@NUDT.EDU.CN
Zhenyu Qiu QIUZHENYU22@NUDT.EDU.CN
Lin Liu LIULIN16@NUDT.EDU.CN
Shaojing Fu FUSHAOJING@NUDT.EDU.CN
College of Computing, National University of Defense Technology, Changsha 410073, China

Editors: Vu Nguyen and Hsuan-Tien Lin

Abstract

Recent studies have exposed the substantial vulnerability of voice-activated smart devices to adversarial examples, predominantly targeting the robustness of automatic speech recognition (ASR) systems. Most of adversarial examples are generated by introducing adversarial perturbations within the l_p norm bounds to benign audio inputs. However, these attacks are constrained by the parametric bounds of perturbations or the features of disturbance, which limits their effectiveness. To improve the acoustic realism of adversarial examples and enhance attack performance, we propose a novel attack framework called Diffusion-based Adversarial Attack. By leveraging DiffVC, a diffusion-based voice conversion model, to map audio into a latent space and employing Adversarial Latent Perturbation (ALP), we manage to achieve robust and imperceptible adversarial perturbations embedding. Extensive evaluations demonstrate that our method enhances targeted attack performance. Specifically, our method has remarkably achieved a Word Error Rate (WER) of 103.7%, alongside a Success Rate (SR) of 99%, demonstrating a notable improvements of 25% and 11% respectively over the state-of-the-art attack. Additionally, our approach also stands out for its high audio quality and efficiency.

Keywords: Trustworthy Machine Learning; Adversarial Examples; Automatic Speech Recognition; Voice Conversion; Deep Learning; Machine Learning.

1. Introduction

In recent years, deep neural networks (DNNs) have achieved remarkable progress in various fields, notably as natural language processing (NLP) Zhang et al. (2020), autonomous driving Liu et al. (2024), and speech recognition Prabhavalkar et al. (2023). Despite their tremendous success, DNNs are vulnerable to adversarial attacks Goodfellow et al. (2015). These attacks inject imperceptible perturbations to the input data, which can cause the model to produce incorrect outputs or misclassify the data. **Our work primarily focuses on the automatic speech recognition (ASR) domain, as adversarial attacks present challenges to the security and reliability of speech-related applications.** Attackers can craft and inject malicious speech commands that can lead to severe security and safety consequences, including economic damages Samuel et al. (2020), health risks Venkatraman et al. (2021), and even physical harm Sugawara et al. (2020).

To investigate the threats posed by audio adversarial examples, extensive works Carlini and Wagner (2017, 2018); Qin et al. (2019); Chen et al. (2020); Zheng et al. (2021) have

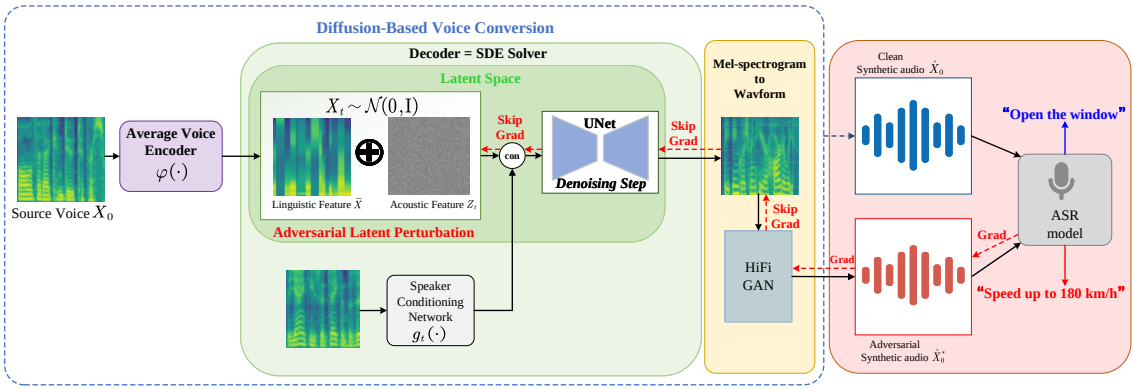


Figure 1: Pipeline of Diffusion-based Adversarial Attack. Firstly, we use Diffusion-Based Voice Conversion Model to construct latent space. Next, Adversarial Latent Perturbation is used to generate adversarial examples. Eventually, the generated adversarial examples can fool the ASR model.

proposed various methods. In order to maintain human acoustic imperceptibility and realism, most of these work introduce adversarial perturbations within the constraint of l_p norm , thereby restricting the distance between the adversarial audio and benign audio. However, this approach may overly constrain the perturbations, limiting the effectiveness of adversarial attacks. To address this gap, Qu et al. (2022) proposed a method named Speech Synthesizing Attack (SSA) to optimize the audio style vector z in the CVAE model Kim et al. (2021) for the first time, which controls the pitches and rhythms of synthesized waveforms, fooling the ASR model to transcribe incorrectly or even transcribe to a given target text. However, they found that some synthesized adversarial audios do not sound as natural as those original synthesized audios, which needs efforts to enhance the quality of adversarial audio synthesis.

Considerable efforts have made in the aforementioned works, several major challenges still remain in generating realistic yet deceptive audio adversarial examples. 1) **Challenges in balancing the Audio Quality and Attack Efficiency**. Despite efforts to maintain audio quality, adversarial perturbations often introduce audible artifacts or distortions, degrading the attack's effectiveness against human perception. To elevate the fidelity of adversarial audio, it is imperative to allocate additional computational resources and time in optimizing adversarial perturbations. 2) **Dependence on existing audio**. Most audio adversarial examples are typically *audio dependent attack* (ADA), requiring construction based on existing benign audio samples. In cases where the human speaker or the original benign audio is unavailable, these attacks become not accessible. Mitigating above limitations is critical for developing more robust audio adversarial attacks and effective countermeasures.

To tackle that aforementioned challenges, we propose a novel adversarial attack called "Diffusion-based Adversarial Attack", which enables imperceptible and robust perturbations injection in the latent space that covering all aspects of audio information. As shown in Figure 1, we first map audios onto a low-dimensional manifold using a advanced diffusion-

based voice conversion model, ensuring high-quality conversion that maintains speaker similarity and speech naturalness, even for unseen speakers. Within this latent space, multiple features enable the generation of diverse and natural-sounding adversarial audios. By optimizing the adversarial objective through Adversarial Latent Perturbation (ALP), we enrich the audios with a diverse range of potent adversarial features, enhancing the attack performance without compromising the audio's naturalness. Moreover, a novel sampling scheme accelerates synthesis, guided by gradient descent for rapid adversarial attack. Our method integrates advanced modeling with adversarial optimization, enhancing the generation of high-fidelity, effective adversarial audios. In conclusion, our contributions can be summarized as follows:

- We propose a novel audio independent adversarial attack framework called Diffusion-based Adversarial Attack, which utilizes high-capacity and low-dimensional manifolds to generate more diverse and natural adversarial audios. To the best of our knowledge, our framework is the first to employ a diffusion-based model for generating adversarial audio examples.
- We present the Adversarial Latent Perturbation, that add perturbations in the latent space which contains diverse features. By utilizing **Adversarial Latent Optimization** and **Skip Gradient** method, we optimize perturbations in the latent space, rapidly generating realistic yet deceptive adversarial audios.
- Our attack's efficacy has been validated through extensive experiments on the state-of-the-art (SOTA) ASR model, Whisper. Notably, we have achieved an average improvement in WER ranging from 3 to 25 absolute points and in SR ranging from 11 to 16 absolute points, surpassing the performance of previous attacks. By enhancing audio quality and optimizing attack efficiency, our approach successfully generates high-quality adversarial audios within a short time.

2. Related Works and Background

2.1. Adversarial Attack on ASR Model

Adversarial examples are purposefully crafted alterations to input data that can lead classifiers to misclassify them. In the realm of speech recognition, these examples serve as a critical tool for evaluating the robustness of speech processing systems. A variety of methods have been proposed for generating such examples, which can be generally categorized into the following types:

2.1.1. **Audio Dependent Attack.**

Audio adversarial examples which are constructed depending on existing benign audios can be deemed as *Audio Dependent Attack* (ADA). These attacks are designed on the principle that adversarial perturbations must be imperceptible to human perception to preserve the semantic integrity of the audio while significantly deceiving automatic speech recognition (ASR) models. A great deal of methods Carlini and Wagner (2018); Qin et al. (2019); Silver et al. (2017); Song et al. (2018) have been developed under this framework, assuming that

the perturbations can be constrained to be as unnoticeable as possible. For example, a psychoacoustic rule of auditory masking Qin et al. (2019) is employed to add perturbations in regions that are least perceptible to human hearing. However, ADAs are contingent upon the availability of benign audios, which may not always be feasible in practical scenarios where the original human speaker or audio is inaccessible. Furthermore, the perturbation principle underlying the ADA approach requires the introduced perturbations be rigorously constrained to evade detection by human perception. This constraint is essential to ensure the efficacy of the adversarial attack while preserving the naturalness of the audio. However, the reliance on imperceptibility and the dependency on benign audios highlight the limitations and challenges of ADA in real-world applications.

2.1.2. **Audio Independent Attack.**

Distinct from Audio Dependent Attacks (ADAs) that necessitate access to specific audio input, *Audio Inependent Attacks* (AIAs) operate on the principle that any audio capable of deceiving ASR models while remaining imperceptible to humans poses a security threat within the domain of voice recognition. Unlike ADA, AIA enables a novel threat model that generates adversarial audio from the scratch, bypassing the need to add perturbations to benign audios. The majority of these unrestricted adversarial attacks have been concentrated within the image domain, as evidenced by studies Shi et al. (2022). Carlini and Wagner's work Carlini and Wagner (2018) explored an audio attack that initiates from non-speech sources, such as classical music, while still necessitates the use of existing audio to apply perturbations. Similarly, the research conducted by Roy et al. (2018) introduced methods to inaudibly modulate voice commands onto ultrasonic frequencies, enabling silent adversarial control over voice command systems. In contrast to prior works, SSA Qu et al. (2022), leveraging advances in neural speech synthesis, directly synthesizes adversarial audio that retains the intended semantic content and effectively deceives ASR models into producing incorrect or targeted transcriptions. This novel approach not only mitigates the limitations of ADA by eliminating the reliance on benign audios but also broadens the scope of adversarial attacks within speech recognition systems. However, this method only choose the audio style vector, which controls the pitches and rhythms of synthesised waveform, for minor modifications, rather than altering all features. Yet, it will sacrifices the flexibility and attack performance of AIA.

2.2. Whisper ASR Model

Whisper ASR model is a transforner sequence-to-sequence model trained on very large amounts of supervised data, which achieve very impressive robustness against noise and out-of-distribution data. It has enabled a range of applications, including captioning, translation, and audio analysis, across various industries and domains. However, Olivier and Raj (2022) show that Whisper model is vulnerable to white-box adversarial attacks. Based on this, we also conduct adversarial attacks on the Whisper model. In contrast to previous automatic speech recognition (ASR) models that utilize the Connectionist Temporal Classification (CTC) loss, the Whisper model employs the Cross-Entropy (CE) loss between its predictions and ground truth text transcriptions. Through this CE loss, the model implic-

itly learns to align its token predictions with the corresponding labels, circumventing the need for explicit alignment mechanisms.

3. Method

3.1. Problem Definition

Given a clean source voice X and a target voice Y, attackers aim to construct an adversarial voice $\hat{X}$ that is naturally sounded but able to deceive well trained ASR model $\mathcal{F}(\cdot)$ in predicting targeted transcription y_t:

$$\mathcal{F}(\text{Attack}(X)) = \mathcal{F}(\hat{X}) = y_t, \tag{1}$$

where

$$\hat{X} = \mathit{DiffVC}(C(X), S(Y)) \tag{2}$$

Here, Attack$(\cdot)$ is the attack approach, while $C(\cdot)$ and $S(\cdot)$ denote the extraction of linguistic information and acoustic information from the voice, respectively. Notice that equaption (2) implies that $\hat{X}$ generated by DiffVC model contains the linguistic information from the source speech and the acoustic information from the target speech.

Different from previous attacks that apply l_p norm to constraints audio perturbation values on raw waveform ($\|\delta\|_p < \epsilon$, where δ is the perturbation and ϵ is a small positive constant), we add perturbations in latent space of the diffusion model and exploit the properties of the diffusion model to generate acoustically natural and successful adversarial attacks. We elaborate on our approach in the following paragraphs.

3.2. Diffusion-Based Adversarial Attack

We observe that perturbations in the latent space are less perceptible and more robust, which makes it an ideal candidate for launching adversarial attacks. In order to achieve the latent-based adversarial attack above-mentioned, known as **Diffusion-based Adversarial Attack**, we propose to first leverage DiffVC, which implements **Diffusion-Based Voice Conversion** to map audios onto a low-dimensional manifold that represents the feature latent space covering all aspects of audio information. Then, we propose an **Adversarial Latent Perturbation (ALP)**, as shown in Figure 1, through the interaction of generative model and ASR models to embed perturbations in the latent space. By the joint optimization of perturbations and signed gradients, we generate realistic yet deceptive adversarial audios.

3.2.1. Diffusion-Based Voice Conversion.

We use DiffVC proposed by Popov et al. (2021) for efficient synthesis of high-quality speech. In practical, DiffVC employs an encoder to parameterize the terminal distribution of the forward diffusion, extracting linguistic content from the source speech X to predict average voice features $\bar{X}$. The reverse diffusion is parameterized with the decoder, which generates speech with the target voice Y conditioning using the encoded content, enabling the model

to adapt to new voices from a single reference. The forward and reverse processes can be represented by the following Stochastic Differential Equations (SDEs):

$$dX_t = \frac{1}{2}\beta_t(\bar{X} - X_t)d_t + \sqrt{\beta_t}d\overrightarrow{W_t}, \tag{3}$$

$$d\hat{X}_t = (\frac{1}{2}(\bar{X} - \hat{X}_t) - s_\theta(\hat{X}_t, \bar{X}, g_t(Y), t))\beta_t dt + \sqrt{\beta_t}d\overleftarrow{W_t}, \tag{4}$$

where t ∈ [0,1], $\overrightarrow{W_t}$ and $\overleftarrow{W_t}$ are independent Wiener processes, β_t is noise schedule, s_θ is the score function with parameters θ.

With a well-trained reverse diffusion process, its trajectories closely approximate the forward diffusion process. Consequently, data generation can be achieved by sampling $\hat{x}_1$ from the prior distribution $\mathcal{N}(\bar{x}, \mathrm{I})$ and solving the stochastic differential equation (SDE) (4) in reverse time. In this way, we can sample $\hat{x}_{t-h}$ from the reverse SDE solvers:

$$\hat{X}_{t-h} = \hat{X}_t + \beta_t h\left(\left(\frac{1}{2} + \hat{\omega}_{t,h}\right)(\hat{X}_t - \bar{X}) + (1 + \hat{\kappa}_{t,h})s_\theta(\hat{X}_t, \bar{X}, g_t(Y), t)\right) + \hat{\sigma}_{t,h}\xi_t, \tag{5}$$

where $t = \{1, 1-h, .., h\}$, $h = 1/N$, ($N \in \mathbb{N}$ is the number of SDE solver steps) and ξ_t are i.i.d. samples from $\mathcal{N}(0, \mathrm{I})$. The score function s_θ conditioned on the speaker conditioning network $g_t(Y)$, is trained to approximate gradient of the log-density of noisy data X_t. For $\hat{\kappa}_{t,h}, \hat{\omega}_{t,h}, \hat{\sigma}_{t,h}$, the described derivation process was introduced in Popov et al. (2022).

3.2.2. Adversarial Latent Perturbation.

An ideal AIA attack should ensure the acoustic realism and semantic integrity of adversarial examples, exhibits diverse adversarial audio features, and demonstrate potential attack performance. In light of the current absence of comprehensive strategies, we propose an audio independent Latent-based attack called Adversarial Latent Perturbation (ALP).

Design Intuition. Extensive research Pope et al. (2021); Shamir et al. (2021) have revealed and verified that real-world data distribution can be described by low dimensional manifold. Such manifolds facilitate the learning process for neural networks, enabling them to construct complex decision boundaries from a relatively small set of training examples. When a well-trained model captures the essence of natural audios on these low-dimensional manifolds, it inherently ensures the realism of the synthesized audios and retains the richness of content, including acoustic and linguistic features. By projecting an audio onto this low-dimensional manifold, moving it along the adversarial direction on the manifold with accurate gradient descent guidance, the direct search method efficiently yields an adversarial audio. Moreover, since ASR systems also conform to the distribution of these manifolds, adversarial examples crafted along the manifold have more adversarial features and and exhibit enhanced attack performance, as shown in Figure 2.

Design Overview. To mitigate perceptible disturbances in the audio latent space while enhancing the efficiency of adversarial attack, we add perturbations and propose an optimization method named ***Adversarial Latent Optimization***. After the input source voice

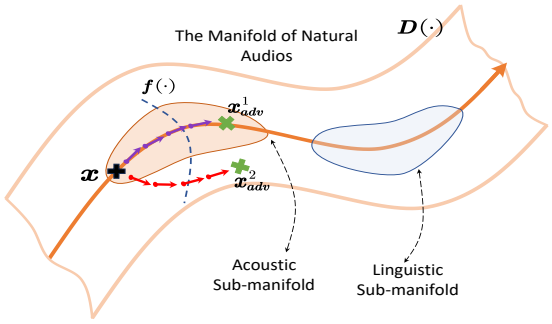


Figure 2: Adversarial examples generated by AIA and ADA, donated as x^1_{adv} and x^2_{adv}. AIA generate adversarial example x^1_{adv} along the adversarial direction of the acoustic manifold and remain within the distribution of the original audio data. In contrast, adversarial examples x^2_{adv} generated by ADA falls out-of-distribution. Our attack (orange arrow) can manipulate attack in the low-dimensional manifold of natural audios, which combine acoustic and linguistic sub-manifold.

is converted into an 'average voice' through the forward diffusion process, the resulting 'average voice' mel-spectrograms preserve the semantic information of the synthesized audio, while the speaker conditioning network $g_t(\cdot)$ contains the acoustic information, ensuring the speaker similarity and speech naturalness of the audio. Considering that $\hat{X}_1 \sim \mathcal{N}(0, \mathrm{I})$ substantially contain both the acoustic features and linguistic features in the latent space, we determine to perturb it to enrich adversarial examples with a spectrum of deceptive features, such as pitch, rhythm, and semantics, thereby enhancing attack performance. With ***Skip Gradient***, we calculate signed gradients which provides the adversarial direction, and manipulate Adversarial Latent Perturbation. The algorithm for ALP is presented in Algorithm 1, and we integrate the diffusion-based voice conversion model to design the optimization methods.

Adversarial Latent Optimization. Based on the average mel-spectrogram $\bar{x}$ by the encoder $\varphi(\cdot)$, the reverse denoising process of DiffVC can be defined as $D(\cdot)$ through equation (5), and it involves t iterations:

$$D(\hat{X}_1, \{g_t(Y)\}_{t=0}^1, \bar{X}, 1) = \hat{X}_0(\hat{X}_h(..., (\hat{X}_{1-h}, g_{1-h}(Y), 1-h), \\ ..., g_h(Y), h), g_0(Y), 0), \tag{6}$$

Therefore, the synthesis audio is indicated by $\hat{X}_0 = D(\hat{X}_1, \{g_t(Y)\}_{t=0}^1, \bar{X}, 1)$. Combined with equaption (1), the optimization of adversarial latent can be defined as fellows:

$$\min_{\delta} \mathcal{L}\left(\mathcal{F}(\hat{X}_0^*), y_t\right), \quad s.t. ||\delta||_\infty \leq \kappa, \tag{7}$$

where

$$\hat{X}_0^* = D(\hat{X}_1 + \delta, \{g_t(Y)\}_{t=0}^1, \bar{X}, 1) = \hat{X}_0^* \text{ and } \hat{X}_0^* \text{ sounds natural}, \tag{8}$$

Here, δ represents the adversarial perturbation on the latent space. Once perturbation has been integrated, we subsequently refine the initial estimate $\hat{X}_0$ to $\hat{X}_0^*$, denoting the

audio adversarial examples. To achieve our objective of generating natural-sounding audio and misleading ASR model into incorrect transcriptions, we formulate our loss function of two parts: **i**) cross-entropy loss $\mathcal{L}_{ce}$, which is used in the training process of the Whisper model. Thus, adversarial examples optimized with $\mathcal{L}_{ce}$ will guide the Whisper model toward misclassification. **ii**) mel-spectrogram loss $\mathcal{L}_{mel}$ mainly guides to minimize the l_1 norm distance between the mel-spectrograms of the generated adversarial examples $\hat{X}_0^*$ and those of the clean synthesis audio $\hat{X}_0$. Consequently, the total loss function $\mathcal{L}$ is formulated as follows:

$$\mathcal{L}(\mathcal{F}(\hat{X}_0^*), y_t, \hat{X}_0) = \mathcal{L}_{ce}(\mathcal{F}(\hat{X}_0^*), y_t)) - \beta \cdot \mathcal{L}_{mel}(\hat{X}_0^*, \hat{X}_0), \tag{9}$$

where

$$\mathcal{L}_{mel}(\hat{X}_0^*, \hat{X}_0) = ||\phi(\hat{X}_0) - \phi(\hat{X}_0^*)||_1 \tag{10}$$

In this context, $\mathcal{L}_{mel}$ is controlled by β, and ϕ denotes the function that converts audio into the corresponding mel-spectrogram.

Similar to most of adversarial attacks, we use the sign gradient descent approach to estimate δ through: $\delta \simeq \alpha \nabla_{\hat{X}_1^*} \mathcal{L}\left(\mathcal{F}(\hat{X}_0^*), y_t\right)$, in which α represents the magnitude of perturbations directed by the gradient. Additionally, to maintain the congruence between $\hat{X}_0$ and $\hat{X}_0^*$, we constrain α in the l_∞-ball of radius ϵ. Specifically, if the perturbation δ exceeds the threshold ϵ in optimization steps, it will be projected back onto the boundary of the ball to ensure that $||\delta||_\infty \leq \epsilon$. The gradient $\delta \simeq \alpha \nabla_{\hat{X}_1^*} \mathcal{L}\left(\mathcal{F}(\hat{X}_0^*), y_t\right)$ can be expanded by the chain rule as follows:

$$\nabla_{\hat{X}_1^*} \mathcal{L}\left(\mathcal{F}(\hat{X}_0^*), y\right) = \frac{\partial \mathcal{L}}{\partial \hat{X}_0^*} \cdot \frac{\partial \hat{X}_0^*}{\partial \hat{X}_h^*} \cdot \frac{\partial \hat{X}_h^*}{\partial \hat{X}_{2h}^*} \cdots \frac{\partial \hat{X}_{1-h}^*}{\partial \hat{X}_1^*}. \tag{11}$$

Skip Gradient. For a complete denoising process that encompasses T sequential calculation graph, the direct computation of gradients is unfeasible due to GPU memory overflow. However, this challenge can be addressed through the adoption of a skip-gradient method, which was originally proposed in Chen et al. (2024).

The entire calculation graph consists of two primary parts. The initial segment, denoted by $\frac{\partial \mathcal{L}}{\partial \hat{X}_0^*}$ represents the gradient of the classifier's loss function with respect to the reconstructed mel-spectrogram $\hat{X}_0^*$, and it specifies the direction of the adversarial gradient. Following this, the subsequent component involves the computation of the iterative product $\frac{\partial \hat{X}_t^*}{\partial \hat{X}_{t+1}^*}$, which corresponds to the backpropagation process through the generative model. This process approximate $\frac{\partial \hat{X}_0^*}{\partial \hat{X}_1^*} = \frac{\partial \hat{X}_0^*}{\partial \hat{X}_h^*} \cdot \frac{\partial \hat{X}_h^*}{\partial \hat{X}_{2h}^*} \cdots \frac{\partial \hat{X}_{1-h}^*}{\partial \hat{X}_1^*}$ employing a skip gradient strategy to mitigate the gradient issue. In diffusion process, the denoising process is designed to eliminate the Gaussian noise introduced during sampling Song et al. (2020). Denoting $\bar{\boldsymbol{\alpha}}_{\boldsymbol{t}} = \gamma_{0,t}$ and $\bar{\boldsymbol{\beta}}_{\boldsymbol{t}} = \sqrt{1 - \gamma_{0,t}^2}$, the Stochastic Differential Equation (SDE) framework resembles Denoising Diffusion Implicit Models (DDIM), enabling the closed-form sampling of X_t at any arbitrary time step t through the reparameterization trick:

$$X_t = \bar{\boldsymbol{\alpha}}_{\boldsymbol{t}} X_0 + \bar{\boldsymbol{\beta}}_{\boldsymbol{t}} \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, I). \tag{12}$$

Subsequently, we execute a transformation by rearranging equation (12), yielding the expression $X_0 = \frac{1}{\bar{\boldsymbol{\alpha}}_{\boldsymbol{t}}} X_t - \frac{\bar{\boldsymbol{\beta}}_{\boldsymbol{t}}}{\bar{\boldsymbol{\alpha}}_{\boldsymbol{t}}} \varepsilon$. As the timestep t approaches to 1, we deduce that $\lim_{t \to 1} \frac{\partial \hat{X}_0^*}{\partial \hat{X}_1^*} = \lim_{t \to 1} \frac{1}{\bar{\boldsymbol{\alpha}}_{\boldsymbol{t}}} \approx \rho$, where ρ is a constant. This approximation results in a simplified gradient for the denoising process, given by $\nabla_{\hat{X}_1^*} \mathcal{L}\left(\mathcal{F}(\hat{X}_0^*), y\right) = \rho \frac{\partial \mathcal{L}}{\partial \hat{X}_0^*}$, which significantly reduces computational complexity and memory usage. Consequently, the gradient can be efficiently computed using the loss function evaluated with respect to the adversarial audio $\hat{X}_0^*$.

Algorithm 1 Diffusion-Based Adversarial Attack

Input: source voice X, target voice Y, target transcription y_t, average voice encoder $\varphi(\cdot)$, speaker conditioning network $g_t(\cdot)$, ASR model $\mathcal{F}(\cdot)$, SDE solver steps N, attack iterations N_a, and momentum factor μ

Calculate $\bar{X}$ by average voice encoder $\varphi(X)$
Initialize $h = 1/N, \delta_0 \leftarrow 0, g_0 \leftarrow 0$
// ***Diffusion-Based Voice Conversion***
for $t = 1, 1-h, \ldots, h$ **do**
  $d\hat{X}_t = (\frac{1}{2}(\bar{X} - \hat{X}_t) - s_\theta(\hat{X}_t, \bar{X}, g_t(Y), t))\beta_t dt + \sqrt{\beta_t} d\overleftarrow{W}_t$
  $\hat{X}_{t-h} \leftarrow \hat{X}_t - d\hat{X}_t$
end for
// ***Adversarial Latent Perturbation***
for $k = 1, \ldots, N_a$ **do**
  $\hat{X}_0^* \leftarrow D(\hat{X}_1 + \delta, \{g_t(Y)\}_{t=0}^1, \bar{X}, 1)$
  $\nabla_{\hat{X}_1^*} \mathcal{L}\left(\mathcal{F}(\hat{X}_0^*), y\right) \leftarrow \rho \frac{\partial \mathcal{L}}{\partial \hat{X}_0^*}$
  $g_k \leftarrow \mu \cdot g_{k-1} + \frac{\nabla_{\hat{X}_1^*} \mathcal{L}(\mathcal{F}(\hat{X}_0^*, y))}{||\nabla_{\hat{X}_1^*} \mathcal{L}(\mathcal{F}(\hat{X}_0^*, y))||_1}$
  $\delta_k \leftarrow \prod_\kappa (\delta_{k-1} + \eta \cdot \text{sign}(g_k))$
end for

Output: The high-quality audio adversarial examples $\hat{X}_0^*$.

4. Experiments

4.1. Experiments Setup

4.1.1. Datasets.

In our experiments, we employ the LibriSpeech dataset, a comprehensive corpus of 1,000 hours of audiobooks designed for speech recognition and voice conversion tasks. This dataset is split into three subsets: a 100-hour set, a 360-hour set, and a 500-hour set, which collectively facilitate a broad spectrum of research applications. We utilize the test-clean-100 subset for our voice conversion attacks.

4.1.2. Models.

In our attack framework, DiffVC and the Whisper models are utilized as voice conversion synthesizers and speech recogniser, respectively. The DiffVC model trained on Librispeech datasets, is capable of high-quality voice conversion with 30 reverse diffusion steps. The speaker conditioning network is set to the *wodyn* type to condition the decoder at time t with the noisy target mel-spectrogram Y_t. And the Whisper model includes 5 model sizes:

tiny (39M parameters) to large (1550M). Constrained by the device, we only implement our attack on tiny to medium (769M). In practical, we use the Whisper inference package, and follows the loss computation functions in Olivier and Raj (2022). More details about the package provided by OpenAI are available at the provided link[1]. Radford et al. (2023) have demonstrated that the Whisper model achieves accuracy on speech recognition and translation that is comparable to other state-of-the-art systems, such as DeepSpeech, thus the success of our method on the Whisper model suggests that it would likely perform well on other ASR models as well.

4.1.3. Experiment Settings.

Our experiments using an NVIDIA GeForce RTX 4090 GPU with Pytorch. SDE solver steps $N = 30$, attack iterations $N_a = 100, \beta = 0.2, \eta = 0.02, \kappa = 0.1$, and $\mu = 1$.

4.1.4. Evaluation Metrics.

We will evaluate our approach from three perspectives: attack effectiveness, efficiency, and the quality of adversarial examples generated. In terms of attack effectiveness, since our method belongs to the category of targeted attacks, we evaluate it with the Word Error Rate (WER) and Success Rate (SR) metrics. Regarding attack efficiency, we intend to measure this by the time consumed per attack or the number of iterations required. Finally, for the assessment of adversarial sample quality, we will employ the Mean Opinion Score (MOS) to evaluate the quality of the generated adversarial audio.

4.2. Attack on Fixed Target Text

To illustrate the superiority of our proposed attack method, we conduct two parts of experiments. For the first part, we choose 100 audio samples as source speakers and perform voice conversion with the same single target speaker to execute the adversarial attack. The objective of our attacks is to deceive recognition systems, with the C&W attack and the ALP targeting the Whisper model, and the SSA focusing on the DeepSpeech model. The targeted text for these attacks is explicitly defined as: 'OK, Google. Browse to evil.com.'. From the experiments presented in this section, it is evident that our attack possesses significant advantages over the baseline work in terms of attack effectiveness, adversarial examples synthesis quality, and attack speed.

4.2.1. Targeted Attack Performance.

We found that ALP attack shows remarkable performance in terms of WER and SR on the fixed target text. Specifically, our method is largely successful in degrading the performance of all Whisper models, by **103.7**% and **99**% regarding WER and SR respectively, significantly outperforming the baselines. The detailed results of our method and comparisons with baselines are presented in Table 1. Our approach demonstrates a relative increase in WER ranging from **5**% to **94**% under the C&W attack, and **3**% under the SSA. Furthermore, the SR has achieved a relative increase of **5**% to **30**% over the C&W attack and **19**% over the SSA. Notably, the Whisper model is considered more advanced than the

1. https://github.com/openai/whisper

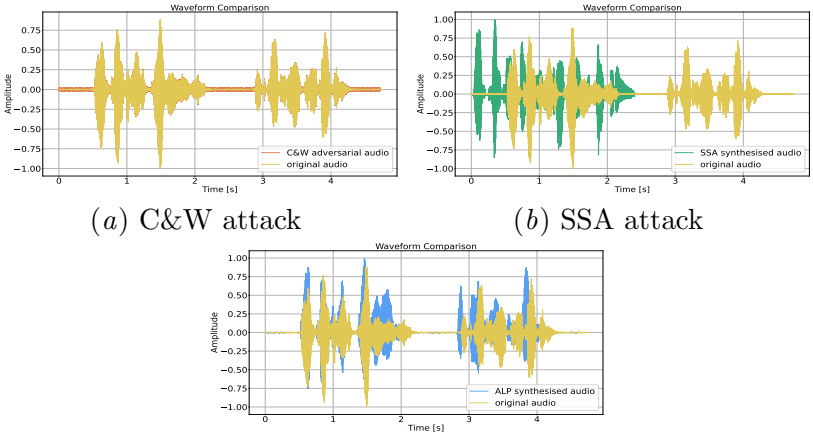
Table 1: Targeted attack results of ALP on the librispeech dataset and comparisons with baselines.

Model	Params.	Clean	C&W		ALP(Ours)		SSA		Model
		WER	WER↑	SR	WER↑	SR	WER↑	SR	
tiny.en	39M	3.4%	94.1%	100%	103.7%	99%	100.7%	83%	Deep Speech
base.en	74M	3.0%	98.3%	94%					
small.en	244M	1.9%	66.0%	82%					
medium.en	796M	1.7%	53.2%	76%					
Average		2.5%	77.9%	88%					

DeepSpeech model. Given our attack's strong performance against the Whisper model, we anticipate that it would also exhibit superior efficacy against the DeepSpeech model.

4.2.2. Quality Evaluation Comparison.

Despite constrained perturbations in the latent space, the synthesized adversarial examples maintain acoustic realism. In our experiments, we subjected 100 source audios to adversarial voice conversion, discovering that our attack is universally applicable across different speakers. As depicted in Figure 3, our attack leverages the strengths of both the SSA attack and C&W attack. The waveforms produced by our ALP attack retain the fundamental shape of high-quality original synthetic audio, enabling a natural auditory perception. Moreover, our adversarial audio significantly differs from the C&W approach, which is limited to minor perturbations. The ALP attack surpasses this constraint, offering enhanced optimization capabilities.

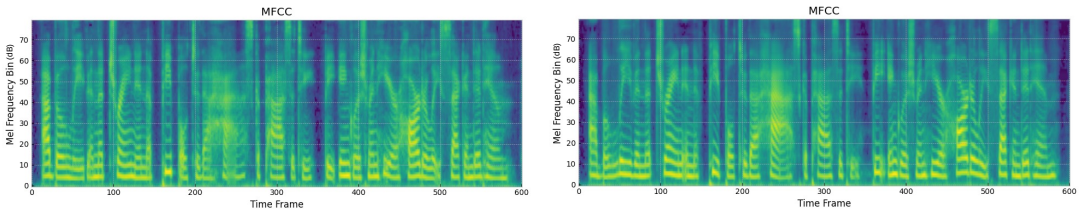


(a) C&W attack (b) SSA attack

(c) Our attack

Figure 3: The comparison on waveforms between the baselines and our ALP.

Furthermore, Figure 4 presents mel-spectrograms synthesized before and after our adversarial attack, illustrating the higher quality of the adversarial examples. The figure demonstrates that the introduction of noise during mel-spectrogram generation does not markedly distort the synthesized mel-spectrogram, ensuring minimal alteration of audio features and preserving the naturalness of the adversarial samples.



(*a*) original synthetic audio (*b*) adversarial synthetic audio

Figure 4: Mel-spectrogram of adversarial audio and original synthetic audio

Additionally, we conducted Mean Opinion Score (MOS) tests to further assess the quality of the synthesized adversarial audios. Fifty participants rated 50 pairs of audio samples, each consisting of the original audio synthesized by DiffVC and its corresponding adversarial audio generated by our ALP attack. The participants listened to the audio samples in a randomized order and rated them on a scale from 1 to 5. The MOS results, detailed in Table 2, indicate that our ALP attack's MOS score (3.75) closely matches the original synthesis (3.96), showcasing the robust generative capacity of our diffusion model. Furthermore, our attack's MOS surpasses that of the baseline SSA attack.

Table 2: The MOS comparison between original and adversarial synthesised audios.

Type of audios		MOS
ALP	Before Attack (original synthesised audios)	3.96
	After Attack (ALP synthesised audios)	3.75
SSA	Before Attack (original synthesised audios)	4.09
	After Attack (SSA synthesised audios)	3.39

4.2.3. Time Consuming.

In this section, we discuss the attack speed of our work. Leveraging a diffusion-based voice conversion model enhanced by a novel sampling scheme, we significantly accelerate the synthesis process. As depicted in Table 3, ALP attack demonstrates a substantially reduced computational time compared to the SSA attack, which requires 8000 iterations, and the C&W attack, which involves minutes of computation. On average, the ALP attack requires only 35.21 seconds, underscoring its efficiency.

Table 3: Attack speed of ALP for different Whisper types and comparisons with baselines.

Attack	Time(sec)					Attack
	tiny.en	base.en	small.en	medium.en	DeepSpeech	
C&W	51.32	98.60	145.74	498.53	1554	SSA
ALP(Ours)	35.21	75.33	95.11	137.21		

Table 4: Time consuming and MOS with different perturbation constraint values.

Perturbation	0.1	0.25	0.5
Time Consuming (sec)	75.33	74.51	75.18
MOS	3.75	3.52	3.21

4.2.4. Perturbation Analysis.

We have experiments with different perturbation constraint values on the Whisper small.en model, specifically 0.1, 0.25, and 0.5. Table 4 indicate that different perturbation ranges minimally affects the time consumption for the attack but significantly impacts the Mean Opinion Score (MOS) values. This could be attributed to the fact that minor changes in the latent space can lead to substantial alterations in the synthesized audio. To ensure the authenticity of the synthesized audio, we have limited the perturbations to a range below 0.1.

4.3. Attack on Multiple Target Text

To explore the impact of different source and target speakers on the performance of our attack, we design following experiments.

For this part, we select 10 distinct source voices and 10 target voices. The source and target voices are paired to synthesize 100 adversarial audios, targeting the Whisper base.en model. Adversarial audios are generated to replicate the source speaker's semantic content and are aligned with the target voice's. This experiments show the robustness of our approach across a diverse spectrum of speakers. Furthermore, by aligning with the target audio's semantic content, we demonstrate our attack method's flexibility in various target text length.

By analyzing the Word Error Rate (WER) across varying text lengths for both source and target speaker audios during targeted attacks, we derived Figure 5. The figure shows that the WER reach up to 247% when the conditional text is 49 words and the target text 19 words, and it declines as the length of the conditional text shortens. This suggests that longer conditional texts will enrich more information in the style latent and content spaces of the synthesized speech, allow for a broader spectrum of acoustic and linguistic features to be perturbed, making it easier to achieve the attack objective. Conversely, when the target text becomes longer, it will increase in the distance between the conditional text and the target text, thereby degrades attack performance. This observation indicates that while the method remains effective across various lengths, the optimal balance between the lengths of the conditional and target texts is pivotal for maximizing the attack performance. Therefore, an optimal text length ratio between the source and target texts is key to enhancing attack performance.Importantly, our analysis also reveals that even when the WER is minimized, it remains significantly high at around 100%. This finding confirms the effectiveness of our attack, proving that with just a few seconds of the victim's audio, we can still effectively execute attacks on extended speech content.

Moreover, we investigated the execution time of our attacks and found a positive correlation between audio length and the time required, as shown in Figure 6. Although longer audio segments increase the duration due to a wider search space for perturbation, the non-

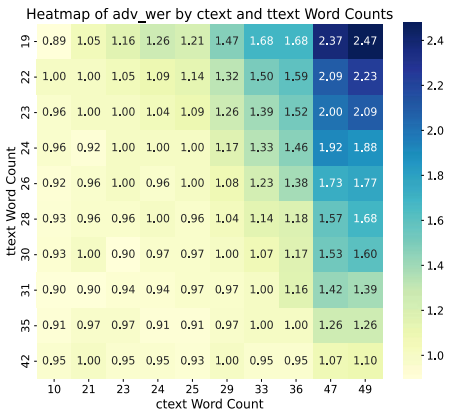


Figure 5: WER with different conditional text length and target text length

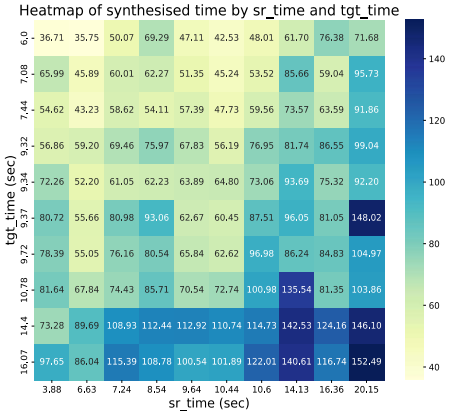


Figure 6: Adversarial synthesis time to different source and target audio durations

linear correlation implies that attack complexity does not increase uniformly with audio length. This insight imply that optimizing audio length can provide more opportunities for perturbation, potentially makes our attacks easier to execute.

In summary, our comprehensive analysis underscores the nuanced interplay between text length, audio duration, and attack performance. It highlights the methodical approach required to fine-tune attacks for maximum efficacy, confirming the versatility and potency of our attack methodology across a spectrum of conditions.

5. Conclusion

In this paper, we propose an audio adversarial attack framework called Diffusion-based Adversarial Attack. We introduce adversarial perturbations within the feature latent space of the diffusion-based voice conversion process to synthesize natural-sounding adversarial samples. By moving the latents along the adversarial gradients during audio synthesis, we are able to generate adversarial voices. Building upon this concept and leveraging diffusion models, we have implemented the Adversarial Latent Perturbation (ALP). Our experiments further demonstrate the superiority of our approach in terms of attack effectiveness, audio synthesis quality, and time efficiency. With the development of diffusion-based generative models, we anticipate a continuous improvement in the quality of synthesized adversarial voice samples and a reduction in computational costs. The availability of high-quality adversarial samples at a low cost presents convenience for attackers and raises concerns regarding the security of Automatic Speech Recognition (ASR) models. In the following research, we will focus on the security and robustness of ASR models in real-world applications.

References

Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 IEEE symposium on security and privacy (SP)*, pages 39–57. IEEE, 2017.

Nicholas Carlini and David Wagner. Audio adversarial examples: Targeted attacks on speech-to-text. In *2018 IEEE security and privacy workshops (SPW)*, pages 1–7. IEEE, 2018.

Yuxuan Chen, Xuejing Yuan, Jiangshan Zhang, Yue Zhao, Shengzhi Zhang, Kai Chen, and XiaoFeng Wang. Devil's whisper: A general approach for physical adversarial attacks against commercial black-box speech recognition devices. In *29th USENIX Security Symposium (USENIX Security 20)*, pages 2667–2684, 2020.

Zhaoyu Chen, Bo Li, Shuang Wu, Kaixun Jiang, Shouhong Ding, and Wenqiang Zhang. Content-based unrestricted adversarial attack. *Advances in Neural Information Processing Systems*, 36:125–139, 2024.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *stat*, 1050:20, 2015.

Jaehyeon Kim, Jungil Kong, and Juhee Son. Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. In *International Conference on Machine Learning*, pages 5530–5540. PMLR, 2021.

Jia Liu, Yunduan Cui, Jianghua Duan, Zhengmin Jiang, Zhongming Pan, Kun Xu, and Huiyun Li. Reinforcement learning-based high-speed path following control for autonomous vehicles. *IEEE Transactions on Vehicular Technology*, pages 1–14, 2024.

Raphael Olivier and Bhiksha Raj. There is more than one kind of robustness: Fooling whisper with adversarial examples. *arXiv preprint arXiv:2210.17316*, 2022.

Phillip Pope, Chen Zhu, Ahmed Abdelkader, Micah Goldblum, and Tom Goldstein. The intrinsic dimension of images and its impact on learning. *arXiv preprint arXiv:2104.08894*, 2021.

Vadim Popov, Ivan Vovk, Vladimir Gogoryan, Tasnima Sadekova, and Mikhail Kudinov. Grad-tts: A diffusion probabilistic model for text-to-speech. In *International Conference on Machine Learning*, pages 8599–8608. PMLR, 2021.

Vadim Popov, Ivan Vovk, Vladimir Gogoryan, Tasnima Sadekova, Mikhail Kudinov, and Jiansheng Wei. Diffusion-based voice conversion with fast maximum likelihood sampling scheme, 2022.

Rohit Prabhavalkar, Takaaki Hori, Tara N Sainath, Ralf Schlüter, and Shinji Watanabe. End-to-end speech recognition: A survey. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, pages 325–351, 2023.

Yao Qin, Nicholas Carlini, Garrison Cottrell, Ian Goodfellow, and Colin Raffel. Imperceptible, robust, and targeted adversarial examples for automatic speech recognition. In *International conference on machine learning*, pages 5231–5240. PMLR, 2019.

Xinghua Qu, Pengfei Wei, Mingyong Gao, Zhu Sun, Yew Soon Ong, and Zejun Ma. Synthesising audio adversarial examples for automatic speech recognition. In *Proceedings of*

the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, pages 1430–1440, 2022.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR, 2023.

Nirupam Roy, Sheng Shen, Haitham Hassanieh, and Romit Roy Choudhury. Inaudible voice commands: The long-range attack and defense. In *15th USENIX Symposium on Networked Systems Design and Implementation (NSDI 18)*, pages 547–560, 2018.

Isaac Samuel, Fiyinfoba A Ogunkeye, Ayobami Olajube, and Ayokunle Awelewa. Development of a voice chatbot for payment using amazon lex service with eyowo as the payment platform. In *2020 International Conference on Decision Aid Sciences and Application (DASA)*, pages 104–108. IEEE, 2020.

Adi Shamir, Odelia Melamed, and Oriel BenShmuel. The dimpled manifold model of adversarial examples in machine learning. *arXiv preprint arXiv:2106.10151*, 2021.

Cong Shi, Tianfang Zhang, Zhuohang Li, Huy Phan, Tianming Zhao, Yan Wang, Jian Liu, Bo Yuan, and Yingying Chen. Audio-domain position-independent backdoor attack via unnoticeable triggers. In *Proceedings of the 28th Annual International Conference on Mobile Computing And Networking*, pages 583–595, 2022.

David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, et al. Mastering the game of go without human knowledge. *nature*, 550(7676):354–359, 2017.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

Yang Song, Rui Shu, Nate Kushman, and Stefano Ermon. Constructing unrestricted adversarial examples with generative models. *Advances in neural information processing systems*, 31, 2018.

Takeshi Sugawara, Benjamin Cyr, Sara Rampazzi, Daniel Genkin, and Kevin Fu. Light commands: Laser-based audio injection attacks on voice-controllable systems. In *29th USENIX Security Symposium (USENIX Security 20)*, pages 2631–2648, 2020.

Sitalakshmi Venkatraman, Anthony Overmars, and Minh Thong. Smart home automation—use cases of a secure and integrated voice-control system. *Systems*, 9(4):77, 2021.

Wei Emma Zhang, Quan Z Sheng, Ahoud Alhazmi, and Chenliang Li. Adversarial attacks on deep-learning models in natural language processing: A survey. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 11(3):1–41, 2020.

Baolin Zheng, Peipei Jiang, Qian Wang, Qi Li, Chao Shen, Cong Wang, Yunjie Ge, Qingyang Teng, and Shenyi Zhang. Black-box adversarial attacks on commercial speech platforms with minimal information. In *Proceedings of the 2021 ACM SIGSAC conference on computer and communications security*, pages 86–107, 2021.