IO-RAE: Information-Obfuscation Reversible Adversarial Example for Audio Privacy Protection

Jiajie Zhu[1], Xia Du[1*], Xiaoyuan Liu[1], Jizhe Zhou[2], Qizhen Xu[1], Zheng Lin[3], Chi-Man Pun[4]

[1]School of Computer and Information Engineering, Xiamen University of Technology, Xiamen, China
[2]School of Computer Science, Engineering Research Center of Machine Learning and Industry Intelligence, Sichuan University , Chengdu, China
[3]Department of Electrical and Electronic Engineering, University of Hong Kong, Pok Fu Lam, Hong Kong SAR, China
[4]Department of Computer and Information Science, University of Macau Macau, China
cosmics36@163.com, duxia@xmut.edu.cn, 2322071027@stu.xmut.edu.cn, jzzhou@scu.edu.cn, qzxu@xmut.edu.cn, linzheng@eee.hku.hk, cmpun@umac.mo

Abstract

The rapid advancements in artificial intelligence have significantly accelerated the adoption of speech recognition technology, leading to its widespread integration across various applications. However, this surge in usage also highlights a critical issue: audio data is highly vulnerable to unauthorized exposure and analysis, posing significant privacy risks for businesses and individuals. This paper introduces an Information-Obfuscation Reversible Adversarial Example (IO-RAE) framework, the pioneering method designed to safeguard audio privacy using reversible adversarial examples. IO-RAE leverages large language models to generate misleading yet contextually coherent content, effectively preventing unauthorized eavesdropping by humans and Automatic Speech Recognition (ASR) systems. Additionally, we propose the Cumulative Signal Attack technique, which mitigates high-frequency noise and enhances attack efficacy by targeting low-frequency signals. Our approach ensures the protection of audio data without degrading its quality or usability. Experimental evaluations demonstrate the superiority of our method, achieving a targeted misguidance rate of 96.5% and a remarkable 100% untargeted misguidance rate in obfuscating target keywords across multiple ASR models, including a commercial black-box system from Google. Furthermore, the quality of the recovered audio, measured by the Perceptual Evaluation of Speech Quality score, reached 4.45, comparable to high-quality original recordings. Notably, the recovered audio processed by ASR systems exhibited an error rate of 0%, indicating nearly lossless recovery. These results highlight the practical applicability and effectiveness of our IO-RAE framework in protecting sensitive audio privacy.

Introduction

Deep neural networks (DNNs) have revolutionized numerous domains due to their exceptional performance across various tasks (Brown et al. 2020; Dosovitskiy et al. 2020; Lin et al. 2024c, 2025b,a, 2024a,d,b). Among these, Automatic Speech Recognition (ASR) has emerged as a critical application and extensively integrated into many aspects of daily life (Han et al. 2020; Sainath et al. 2021; Ng et al. 2021). However, the rapid and unregulated expansion of ASR technology has raised significant concerns regarding privacy and security. Malicious commercial entities increasingly exploit private voice recordings for profit-driven objectives, highlighting the urgent need to protect the integrity of audio data.

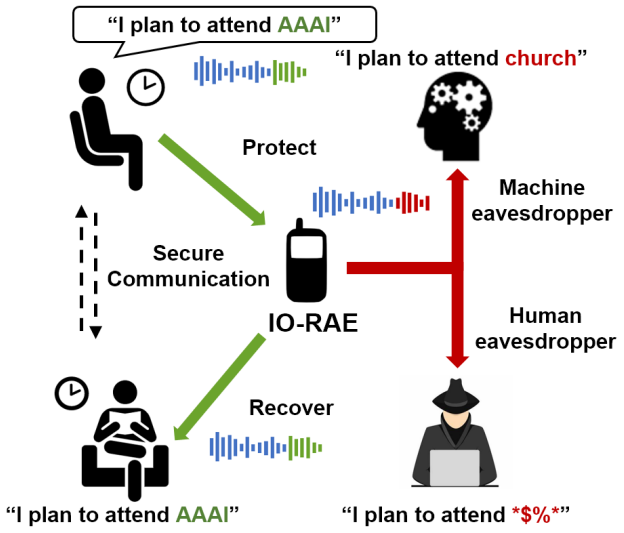

Figure 1: IO-RAE thwarts both human and machine eavesdroppers through the adversarial noise and can recover the original audio losslessly when authorized, ensuring secure communication.

Recent research utlized the Data Anonymization techniques to protect user privacy by de-identifying or distorting personal features in speech signals, preventing ASR systems from extracting sensitive information. Chou *et al.* (Chou et al. 2018) employed adversarial learning for multi-target voice conversion, which effectively removes speaker identity without the need for parallel data, facil-

*Xia Du is the corresponding author.

itating anonymization. Champion *et al.* (Champion 2023) developed a framework that isolates speaker-specific information by modifying acoustic features and embeddings. Feng *et al.* (Feng et al. 2022) introduced domain-adaptive noise injection, which preserves privacy while mitigating the risk of identity disclosure. While these methods represent significant advancements in privacy protection, they still face challenges. Malicious attackers can still extract sensitive information using advanced DNNs, and the process of anonymization often leads to a degradation in speech clarity and intelligibility, limiting their practical applicability.

Given the limitations of existing anonymization methods, more effective approaches are needed to balance privacy protection and speech quality. In this paper, we apply Reversible Adversarial Examples (RAEs) (Liu et al. 2023) to audio privacy protection: RAEs combine adversarial attacks with Reversible Data Hiding (RDH) so that crafted perturbations can mislead malicious DNNs and later be removed by authorized users to fully restore the original audio. However, current RAE designs face two major challenges: (1) enforcing imperceptible perturbations limits their obfuscation power, leaving sensitive information accessible to human listeners (Yang et al. 2021; Zhao, Dua, and Singh 2017; Elsayed et al. 2018); (2) RDH capacity constraints require tight control of embedded data size, and creating adversarial patches to cover critical content is especially hard for ASR because locating key timestamps in audio is nontrivial (Chen et al. 2024).

To address these issues, we propose the Information Obfuscation Reversible Adversarial Example (IO-RAE) framework (Fig. 1). IO-RAE uses alignment techniques to locate sensitive timestamps and applies targeted voice attenuation together with adversarial patches to those segments, improving obfuscation while preserving overall intelligibility. We further employ Large Language Models (LLMs) (Radford et al. 2018; Bai et al. 2025) to generate deceptive labels for patches, making ASR transcriptions plausibly coherent, and introduce Cumulative Signal Attack (CSA), a perturbation optimization that reduces high-frequency harshness by accumulating signal energy into lower-frequency regions.

In summary, the main contributions of this paper are as follows:

- We proposed the IO-RAE for audio privacy protection, designed to prevent both human and machine eavesdroppers from extracting key information from the protected audio. To the best of our knowledge, this approach represents the first application of RAE in the audio domain.

- We proposed an LLM-based target generation method that replaces traditional predetermined targets and utilizes the greedy algorithm to identify the optimal target. This method preserves the coherence and plausibility of the sentences while substantially enhancing attack efficacy.

- We proposed the CSA method for noise preprocessing. CSA can smooth the harsh perturbations by suppressing high-frequency signals and maintain the efficacy of the attack by boosting low-frequency signals.

Background

Adversarial Attack

Adversarial attacks pose a fundamental threat to deep learning models, with their vulnerability validated across multimodal tasks. Szegedy *et al.* (Szegedy et al. 2013) first revealed that imperceptible perturbations can significantly alter the predictions of image classification models. Goodfellow *et al.* (Goodfellow, Shlens, and Szegedy 2014) then introduced the Fast Gradient Sign Method (FGSM), a simple yet effective one-step attack defined as:

$$x_{adv} = x + \epsilon \cdot sign(\nabla_x \ell(x, y)), \tag{1}$$

where x is the input, y the true label, and ϵ the perturbation scale. Since then, numerous attacks (Zhao, Liu, and Larson 2021; Xiao et al. 2019; Dai et al. 2018) have extended beyond image tasks.

In the audio domain, Carlini and Wagner (Carlini and Wagner 2018) pioneered targeted attacks by minimizing the CTC loss (Graves et al. 2006) while constraining perturbation magnitude, enabling adversarial speech to be transcribed as arbitrary phrases. Later, Yakura *et al.* (Yakura and Sakuma 2018) modeled environmental factors—such as room acoustics and noise—for robust over-the-air attacks. Qin *et al.* (Qin et al. 2019) introduced a psychoacoustic-aware framework that leverages auditory masking to create imperceptible yet effective adversarial audio, integrating human hearing models into the attack process.

Reversible Adversarial attack

Research on reversible adversarial attacks has primarily focused on the protection of image data, while their application in the domain of audio data remains underexplored. Liu *et al.* (Liu et al. 2023) first introduced the concept of reversible adversarial attacks by combining reversible data hiding techniques with adversarial examples, pioneering this novel attack methodology. Subsequently, Xiong *et al.* (Xiong et al. 2023) extended this work by applying reversible adversarial attacks to black-box scenarios, incorporating ensemble modeling techniques to demonstrate the potential of reversible adversarial examples across different models. Meanwhile, Zhang *et al.* (Zhang et al. 2022) proposed a method based on RGAN, which replaced reversible data hiding techniques. Specifically, they efficiently generated adversarial examples through an attack encoder network and reversed these examples effectively using a recovery decoder network, achieving efficient restoration of images. Despite its outstanding performance in recovering the original state of images, this method shows limited effectiveness when applied to adversarial tasks against black-box models. More recently, the DP-RAE framework (Zhu et al. 2024) improved attack success rates in black-box scenarios through a Dual-Phase design. In this paper, we propose the method that combines LSB (Kavitha, Koshti, and Dunghav 2012) with adversarial examples to achieve effective protection of audio data.

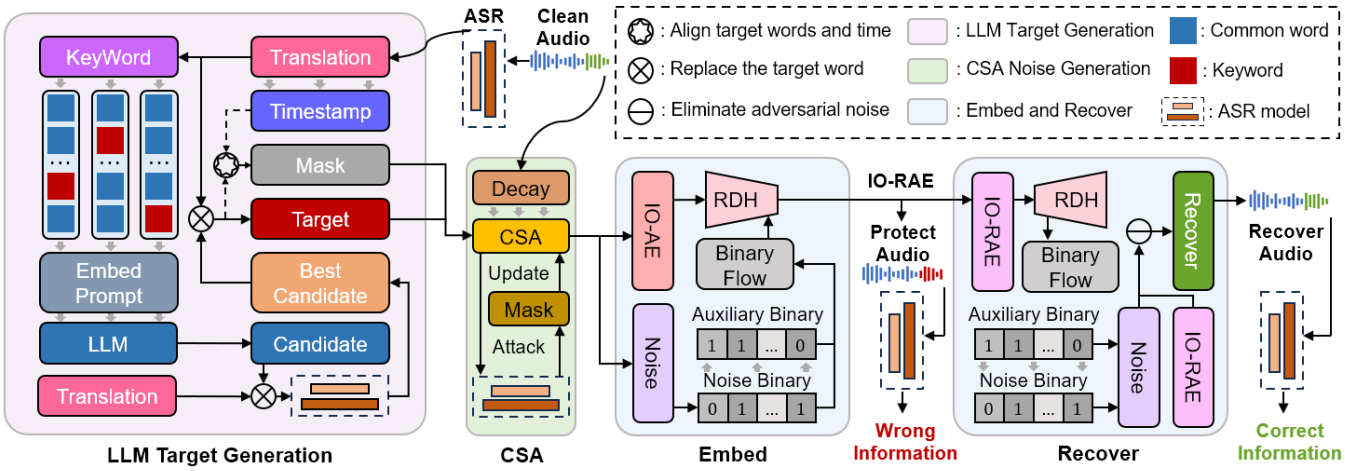

Figure 2: An overview of the IO-RAE framework.

Proposed Method

Overview

In this section, we introduce the IO-RAE framework, which generates adversarial perturbations through the CSA to obfuscate critical information. As shown in Fig . 2, IO-RAE consists of three components: LLM Target Generation, adversarial perturbation creation through the CSA, and Reversible Perturbation Embedding and Recovery. Initially, IO-RAE utilizes KeyBERT (Grootendorst 2020) to extract key information from the correct transcription. These keywords are then combined with prompts and fed into an LLM to generate logically consistent replacement words. Subsequently, the Montreal Forced Aligner MFA (McAuliffe et al. 2017) technique aligns the text and audio to obtain corresponding masks. Audio attenuation and perturbation generation are applied to these masked regions, with CSA to enhance low-frequency. Finally, auxiliary information and the perturbation matrix are embedded into the adversarial audio, forming the IO-RAE. During the recovery process, RDH extracts and decodes the perturbation matrix, effectively removing the adversarial perturbations.

LLM Target Generation

Unlike methods that generate target phrases based on fixed phrases or by maximizing word vector distance, we utilize LLMs to generate candidate replacement words. This approach avoids the confusion caused by words with the farthest vector distances and enhances diversity in the generated phrases. We denote the target transcription as $y = \{y_0, y_1, ..., y_n\}$, where y is initially set as the correct transcription and y_i represents the i_{th} word in the transcription. To achieve the goal of quoting out of context, it is necessary to identify a subset $y^* \subseteq y$ that can replace y and effectively distort the original meaning. Therefore, we employ KeyBERT to identify the most influential words in y to derive y^* and utilize LLM to generate candidate words. This generates candidate words related to y^*. Moreover, the difficulty of misleading different classes in image recognition tasks varies, a phenomenon also observed in ASR systems. This variability is determined by the value of the loss function ℓ between the input and the target y:

$$\ell\left(x, \delta, y\right) = \ell_{CTC}\left(f(x+\delta), y\right), \tag{2}$$

where δ represents the adversarial noise and x denotes original audio. Therefore, we can conclude that there exists an optimal replacement set y_* among the numerous candidate words that can mislead the ASR system at minimal cost. To efficiently determine y^*, we employ a greedy algorithm to identify the best replacement word for each position in y^*. Subsequently, we use the timestamps for each word provided by the MFA to generate the positional mask $\mathbb{M}$ for the perturbation δ. The overall algorithm is detailed in Algorithm 1.

Algorithm 1: LLM Target Generation

Input: Original audio x.
Output: Output: Target transcription y, location mask $\mathbb{M}$.
Initialize: $y = f\left(x\right), y^* = \{\}, \mathcal{L} = \{\}, \mathbb{M} \leftarrow 0\left[shape(x)\right]$

Obtain the keywords: $y^* = \text{KeyBERT}(y)$;
Utilize MFA to align the x and y;
for $i = 0 \rightarrow len(y^*) - 1$ **do**
  Embedding y_i^* into prompts;
  Obtain the candidate lexicon cl: cl = LLM(prompts);
  for $j = 0 \rightarrow len(cl) - 1$ **do**
    Replace $y_{\,idx(y_i^* \ in \ y)}$ with cl_j;
    Calculate the loss ℓ with Eq. 2;
    Append ℓ to $\mathcal{L}$;
  end for
  Find the index min with the smallest value in $\mathcal{L}$;
  Replace $y_{\,idx(y_i^* \ in \ y)}$ with cl_{min};
  Get the start and end time of the $y_{\,idx(y_i^* \ in \ y)}$;
  $\mathbb{M}\left[start : end\right] = 1$;
end for
return $y, \mathbb{M}$

The optimal target transcription y and the location mask $\mathbb{M}$ found using the LLM and greedy-based concepts will be

utilized in the subsequent CSA attack to generate targeted perturbations.

Cumulative Signal Attack

Audio data storage is unique as even short audio segments require extensive tensors for representation due to their high sampling rates, resulting in substantial data volumes in the full perturbation matrix. Given the constrained storage capacity of RDH technology, these perturbations must be compressed prior to embedding. Consequently, in CSA, we utilize composite sampling to reduce data size while preserving attack effectiveness. Moreover, CSA combats overfitting in audio attacks by preventing the generation of additional perturbations for words already successfully attacked and using cumulative signals to soften harsh noises.

The CSA method first decays the original audio content to ensure the imperceptibility of the protected material, then applies a composite sample operation on the matrix to ensure that the compressed gradient matrix retains its adversarial properties. Assuming the size of the composite sample is s and the compressed matrix is $\mathcal{L}_c$. The value of the i_{th} element in $\mathcal{L}_c$ can be expressed as:

$$\mathcal{L}_c[i] = \sum_{j=i*s}^{i*s+s} \nabla_x \ell(x, \delta, y)_j / s, \tag{3}$$

then the CSA method replaces the original gradients in the interval $[i : i + s]$ with the averaged compressed gradient $\mathcal{L}_c[i]$. This adjustment allows the perturbation to align with the compressed gradients.

Subsequently, the CSA method optimizes the generated perturbations by enhancing the low-frequency components. After undergoing the Discrete Fourier Transform (Cooley and Tukey 1965), the perturbation signal δ is represented in the frequency domain as:

$$\triangle(k) = \sum_{n=0}^{N-1} \delta[n] \, e^{-j\frac{2\pi kn}{N}}, \tag{4}$$

and the frequency domain of the cumulative signal c corresponding to δ is:

$$C(k) = \sum_{m=0}^{N-1} \delta[m] \left(\sum_{n=m}^{N-1} e^{-j\frac{2\pi kn}{N}} \right), \tag{5}$$

for each fixed m, one obtains:

$$\sum_{n=m}^{N-1} e^{-j\frac{2\pi kn}{N}} = e^{-mj\frac{2\pi kn}{N}} \sum_{n=0}^{N-1-m} e^{-j\frac{2\pi kn}{N}}. \tag{6}$$

When N is sufficiently large, Eq. 6 can be approximated as:

$$\sum_{n=0}^{N-1-m} e^{-j\frac{2\pi kn}{N}} \approx \frac{1}{1 - e^{-j\frac{2\pi k}{N}}}, \tag{7}$$

the relationship between $C(k)$ and $\triangle(k)$ can be expressed as:

$$C(k) \approx \triangle(k) \cdot \frac{1}{1 - e^{-j\frac{2\pi k}{N}}}. \tag{8}$$

Thus, the cumulative operation on the perturbed signal enhances the low-frequency components while suppressing the high-frequency components. The detailed derivation is provided in the Supplementary Material.

To prevent overfitting to specific words during the perturbation generation process, the CSA method evaluates the attack success rate for each target replacement word before each iteration, updating the success matrix $\mathbb{S}$. If a word has already been successfully attacked, no additional perturbations are applied in subsequent iterations. The complete algorithm is presented in Algorithm 2.

Algorithm 2: Cumulative Signal Attack

Input: Original audio x, target y, decay rate α, mask $\mathbb{M}$,Iteration I, unit noise ϵ, multiplicative factor t
Output: Adversarial audio x_{adv}, noise matrix δ
Initialize: $g_0 = 0, x^0_{adv} = \frac{1}{\alpha} \cdot x \odot \mathbb{M}$, CSA noise $\eta = 0, \delta \leftarrow 0[\text{shape}(x)], \eta \leftarrow 0[\text{shape}(x)], \mathbb{S} \leftarrow \mathbb{M}, \epsilon = \epsilon \cdot \frac{1+\alpha/5}{100}$

for $i = 0 \rightarrow I - 1$ **do**
  $y_{adv} = f(x^i_{adv} + \eta)$
  Verify if the attack succeeds on y_{adv}
  Update $\mathbb{S}$ based on the result of y_{adv}
  Calculate the loss for Eq. 2 and obtain g_{t+1}
  $g_{t+1} = \mu \cdot g_t + \frac{g_{t+1}}{\|g_{t+1}\|_2}$
  Compress and update g_{t+1} using Eq. 3
  $\delta = \delta + \epsilon \cdot \text{sign}(\mathbb{M} \odot \mathbb{S} \cdot g_{t+1})$
  $\delta = \text{Clip}^{t \cdot \delta}_{\delta}(\delta)$
  $\eta = \frac{1}{\alpha} \cdot \sum_{m=0}^{N-1} \delta[m]$ ▷ Cumulative Signal
  $x^{i+1}_{adv} = x^i_{adv} + \eta$
end for
return x_{adv}, δ

Subsequently, the adversarial examples and their corresponding perturbation matrices are employed in the next phase to construct reversible adversarial audio attacks. Notably, to facilitate information encoding and reduce storage overhead, the returned perturbation matrices correspond to the original matrices δ prior to the cumulative signal operation.

Embed and Recover

In the IO-RAE framework, we implement RDH techniques to embed additional data into adversarial examples. Initially, the perturbation matrix generated by the CSA method is encoded. Specifically, the perturbations are sampled at intervals of s, capturing the coefficients of the standard perturbation ϵ, the message of i_{th} element can be expressed as:

$$Mes_i = \frac{\sum_{j=i*s}^{i*s+s} \delta_j}{\epsilon \times s}. \tag{9}$$

All valid perturbation matrices are sequentially traversed, and their correspondi ng Mes_i are aggregated to form Mes, resulting in the perturbation data stream. This stream is then combined with auxiliary information—including the starting position of the perturbation, the standard perturbation magnitude, and the decay rate. The final composite information is embedded into the adversarial audio to ensure its adversarial properties are preserved.

During the audio recovery phase, the embedded information is first extracted using RDH technique. With the extracted Mes and auxiliary data, the perturbation matrix and

its corresponding embedding positions are reconstructed in reverse.Subsequently, the forward process of CSA is reapplied to the perturbed matrix to obtain the CSA noise η. Finally, the original audio is effectively restored by progressively removing the perturbations from the adversarial audio.

Experiments

Experiment setup

Datasets. We randomly selected 200 accurately transcribed audio samples from each of Mozilla Common Voice (Ardila et al. 2019), TIMIT (Garofolo et al. 1993), and LibriSpeech (Panayotov et al. 2015) to validate the method's effectiveness across diverse data sources.

Environment. We used the widely adopted DeepSpeechV3 (Battenberg et al. 2017) as the primary target model due to its relevance in speech recognition. For transferability evaluation, we tested on Conformer (Gulati et al. 2020), Whisper (Radford et al. 2023), Wenet (Yao et al. 2021), SenseVoice (Gao et al. 2023), and Wav2vec (Baevski et al. 2020). All experiments were implemented in PyTorch and conducted on an NVIDIA A40 GPU.

Evaluation Metrics. To assess audio quality and intelligibility, we used SNR, SISDR, STOI, and PESQ. For attack effectiveness, we evaluated targeted success rate (TSR), targeted word error rate (TWER), untargeted success rate (USR), and untargeted word error rate (UWER).

Parameter Setting. We set unit noise ϵ to $1/5$ and the multiplicative factor t to 15, implying each perturbation requires only 5 bits. The composite sample size s was set to 2, balancing storage and effectiveness. Audio attenuation α was set at 5 per 0.1 of maximum signal, capped at 30. The number of attack iterations I was fixed at 200. And for target word generation, we employed Qwen2.5-VL-7B.

Attack and Recover ability

In this section, we demonstrate the performance of IO-RAE in both attack and recovery on three major datasets: Common, TIMIT, and LibriSpeech, specifically targeting the DeepSpeechV3 model. The evaluation dimensions primarily include audio quality and attack performance.

Regarding attack performance, as shown in Table 1, IO-RAE exhibits high target misleading capability across multiple datasets. Notably, on the LibriSpeech dataset, the TSR reached 96.5%, while the TWER was only 0.21%, indicating minimal deviation between the transcription results and the target output even when the target attack was not fully successful. Therefore, the ASR system struggled to extract accurate information from the adversarial audio. As depicted in the Mel spectrogram (Stevens, Volkmann, and Newman 1937) in Fig . 3, the attacked audio regions were fully protected through CSA.

In terms of recovery, Fig . 3 visually illustrates that the recovered audio was free from misleading perturbations and perfectly restored to the clean sample state. Experimental results in Table 1 further confirm that the recovered audio exhibited exceptional quality in various tests, with a PESQ score of 4.45, close to the theoretical maximum of 4.5, objectively validating the superiority of the restored audio. Other critical audio quality metrics, such as SNR, SISDR, and STOI, also showed significant improvements. Re-transcription of the recovered audio using DeepSpeechV3 achieved a TSR of 0%, effectively restoring the original information and proving the effectiveness and practicality of our method.

Source: Common	Audio Quality SNR	SISDR	STOI	PESQ	Attack Performance TSR	TWER
IO-AE	1.8	-0.22	0.83	2.6	96.0	0.6
IO-RAE	1.8	-0.22	0.83	2.6	96.5	0.55
Recover	53.06	53.06	0.99	4.38	0	16.18
Source: TIMIT	Audio Quality SNR	SISDR	STOI	PESQ	Attack Performance TSR	TWER
IO-AE	-12.37	-15.35	0.78	2.41	96.0	1.08
IO-RAE	-12.37	-15.35	0.78	2.41	95.0	1.36
Recover	42.76	42.76	0.99	4.28	0	17.31
Source: Librispeech	Audio Quality SNR	SISDR	STOI	PESQ	Attack Performance TSR	TWER
IO-AE	3.37	1.99	0.89	3.02	96.5	0.21
IO-RAE	3.28	1.91	0.89	3.02	96.5	0.21
Recover	54.75	54.75	0.99	4.45	0	10.2

Table 1: SNR (dB), SISDR (dB), STOI, PESQ, TSR (%) and TWER (%) of the adversarial audio (IO-AE and IO-RAE) and recover audio. Recovered results are highlighted in red.

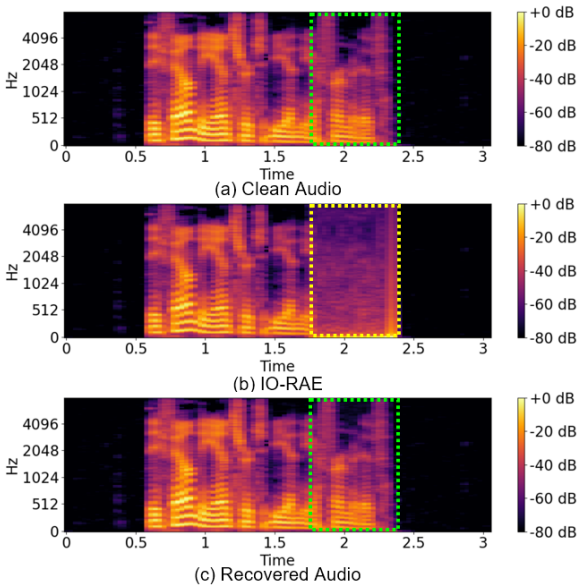

Figure 3: (a), (b), and (c) represent the Mel spectrogram of clean audio, IO-RAE, and recovered audio, respectively.

Robustness evaluation

In real-world applications, input audio typically undergoes a series of defense methods aimed at mitigating the impact of adversarial noise on the model. Therefore, ensuring the robustness of adversarial samples is crucial. To evaluate the

Source: Common	Defense method						
	AS	AT	LPF	MS	QT	MP3-V	OPUS
TSR ↑	8.0	45.0	26.0	13.5	**59.0**	25.0	19.5
TWER ↓	22.7	9.45	13.79	24.17	**6.89**	13.39	15.99
USR ↑	**100**	**100**	**100**	**100**	98.5	**100**	**100**
UWER ↑	24.13	18.33	19.29	**29.19**	17.14	18.31	19.86
Source: TIMIT	Defense method						
	AS	AT	LPF	MS	QT	MP3-V	OPUS
TSR ↑	4.5	32.0	38.0	20.5	34.5	19.5	9.0
TWER ↓	27.27	18.09	11.72	23.82	13.17	16.63	19.28
USR ↑	99.5	99.5	**100**	**100**	**100**	99.0	99.5
UWER ↑	28.98	25.75	20.45	**30.31**	23.31	21.66	22.27
Source: Librispeech	Defense method						
	AS	AT	LPF	MS	QT	MP3-V	OPUS
TSR ↑	10.5	57.5	50.5	12.5	**68.5**	32.0	30.0
TWER ↓	12.69	4.58	5.11	13.9	**3.1**	8.13	8.07
USR ↑	**99.5**	**99.5**	**99.5**	**99.5**	99.0	**99.5**	**99.5**
UWER ↑	14.01	10.69	11.03	**16.91**	10.43	11.77	11.38

Table 2: TSR (%), TWER (%), USR (%) and UWER (%) of IO-RAE under various defense methods.

robustness of IO-RAE, we applied various defense methods, including Average Smoothing (AS) (Du et al. 2020), Audio Turbulence (AT) (Yuan et al. 2018), Low Pass Filter (LPF) (Kwon, Yoon, and Park 2019), Median Smoothing (MS) (Yang et al. 2018), Quantization (QT) (Yang et al. 2018), MP3-V and OPUS (Vos et al. 2013) to the adversarial samples.

As shown in Table 2, despite the varying degrees of impact these defense methods had on targeted attacks, the adversarial examples produced by IO-RAE consistently achieve near-perfect mislead rates close to 100% across all datasets, effectively obstructing the accurate interpretation of the original information. This robustness can be attributed to IO-RAE's reversible adversarial attack nature, which allows for stronger perturbations without compromising overall audio quality. Consequently, IO-RAE demonstrates strong robustness and reliability in real-world applications.

Transferability evaluation

Since the specific architecture used by an intruder's model is often unknown, it is necessary to verify whether IO-RAE can protect critical target information across different models. In the field of image attacks, transferable perturbations can be generated by computing ensemble gradients due to the similarity in model architectures, effectively attacking unknown models. However, in the audio attack domain, significant differences in model architectures make cross-model transfer much more difficult and challenging. IO-RAE utilizes CSA to effectively conceal critical information, achieving cross-model protection. We selected several models with vastly different architectures to evaluate the transferability of IO-RAE, including Wenet, Conformer, SenseVoice, Whisper, and Wav2vec.

As shown in Table 3, IO-RAE was able to effectively protect the key content across various models, preventing multiple models from retrieving the correct original information. This demonstrates that IO-RAE possesses strong cross-model information security capabilities, maintaining the protection of critical information even when faced with different architectures.

Source: model	Common		TIMIT		ibrispeech	
	USR ↑	UWER ↑	USR ↑	UWER ↑	USR ↑	UWER ↑
Wenet	91.5	18.25	89.5	17.48	85.5	10.63
Conformer	94.0	19.72	98.0	21.61	92.0	11.38
SenseVoice	96.0	21.06	98.0	19.85	93.5	13.23
Whisper	99.5	**29.01**	**100**	**29.59**	**99.5**	**29.26**
Wav2vec	**100**	22.67	**100**	22.62	99.0	12.22

Table 3: USR (%) and UWER (%) of IO-RAE targeting various ASRs, the IO-RAE is generated by DeepspeechV3.

Commercial model attack

To further validate the effectiveness of our model in real-world scenarios, we conducted attack experiments targeting the Google API [1], the commercial ASR system. Our objective was to deceive the Google API with IO-RAE, ensuring that it failed to accurately extract the correct information from audio signals. We prepared 50 voice samples from the Common Voice dataset, all of which were initially translated accurately by the API. Subsequently, we generated the corresponding IO-RAE to deceive the API. Thanks to the disturbances introduced by CSA, the IO-RAE completely safeguarded the key content, achieving a 100% attack success rate. In addition, we also recovered the audio from IO-RAE, and the recovered audio could be correctly recognized by Google API again, which means that the recovered audio wipes out the adversarial perturbation and restores the audio quality. We uploaded the corresponding audio data to an anonymous website [2], allowing for playback and download to audition the attack and restoration effects.

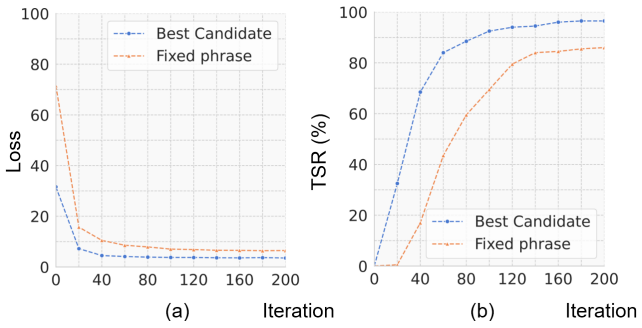

Figure 4: (a) represents the CTC loss and (b) represents the TSR (%) of two strategies in the Common dataset.

Ablation study

We initially explored the impact of the 'best candidate strategy' on attack performance. As illustrated in Fig . 4, a remarkable improvement in attack performance is observed when the 'best candidate strategy' replaces the 'fixed strategy'. This improvement occurs because a higher loss between target words and the input audio indicates a lower

[1]https://cloud.google.com/speech-to-text

[2]https://sites.google.com/view/io-rae/io-rae/

Source:	Composite Sample								Unit Noise					
Common	IO-AE				IO-RAE				IO-AE			IO-RAE		
	$s=2$	$s=3$	$s=4$	$s=5$	$s=2$	$s=3$	$s=4$	$s=5$	$\epsilon=1/4$	$\epsilon=1/5$	$\epsilon=1/6$	$\epsilon=1/4$	$\epsilon=1/5$	$\epsilon=1/6$
TSR ↑	96.0	92.0	91.0	84.5	96.5	93.0	89.5	84.5	96.0	96.0	93.0	95.5	96.5	92.5
TWER ↓	0.60	1.42	1.54	3.05	0.55	1.67	1.96	3.16	0.62	0.60	0.60	0.75	0.55	1.33
USR ↑	100	100	100	100	100	100	100	100	100	100	100	100	100	100
UWER ↑	16.21	16.2	16.27	16.8	16.16	16.33	16.27	16.8	16.08	16.21	16.21	16.08	16.16	16.22
Source:	Composite Sample								Unit Noise					
TIMIT	IO-AE				IO-RAE				IO-AE			IO-RAE		
	$s=2$	$s=3$	$s=4$	$s=5$	$s=2$	$s=3$	$s=4$	$s=5$	$\epsilon=1/4$	$\epsilon=1/5$	$\epsilon=1/6$	$\epsilon=1/4$	$\epsilon=1/5$	$\epsilon=1/6$
TSR ↑	96.0	91.0	84.5	83.5	95.0	88.5	83.5	82.0	88.5	96.0	93.5	85.5	95.0	90.5
TWER ↓	1.08	2.03	3.30	3.38	1.36	2.61	3.52	3.62	2.09	1.08	1.55	2.57	1.36	2.02
USR ↑	100	100	100	100	100	100	100	100	100	100	100	100	100	100
UWER ↑	17.74	17.6	18.07	17.92	17.74	17.90	18.1	18.08	17.59	17.74	17.40	17.89	17.74	17.52
Source:	Composite Sample								Unit Noise					
Librispeech	IO-AE				IO-RAE				IO-AE			IO-RAE		
	$s=2$	$s=3$	$s=4$	$s=5$	$s=2$	$s=3$	$s=4$	$s=5$	$\epsilon=1/4$	$\epsilon=1/5$	$\epsilon=1/6$	$\epsilon=1/4$	$\epsilon=1/5$	$\epsilon=1/6$
TSR ↑	96.5	94.5	90.5	87.5	96.5	94.0	90.5	87.0	94.5	96.5	96.0	94.0	96.5	95.5
TWER ↓	0.21	0.47	0.95	1.75	0.22	0.52	0.95	1.84	0.57	0.21	0.25	0.71	0.22	0.28
USR ↑	100	100	100	100	100	100	100	100	100	100	100	100	100	100
UWER ↑	10.22	10.34	10.44	10.70	10.22	10.34	10.44	10.70	10.34	10.22	10.23	10.39	10.22	10.26

Table 4: TSR (%), TWER (%), USR (%), UWER (%) of IO-RAE with different settings. The composite sample demonstrates the attack performance of IO-RAE with different composite sample sizes, and the unit noise represents the impact of adversarial noise of varying degrees on performance.

perceived match by the ASR, necessitating larger or more complex perturbations to coerce the ASR into outputting the target words. The 'best candidate strategy' effectively reduces this loss, thereby decreasing the required perturbation intensity and enhancing the TSR.

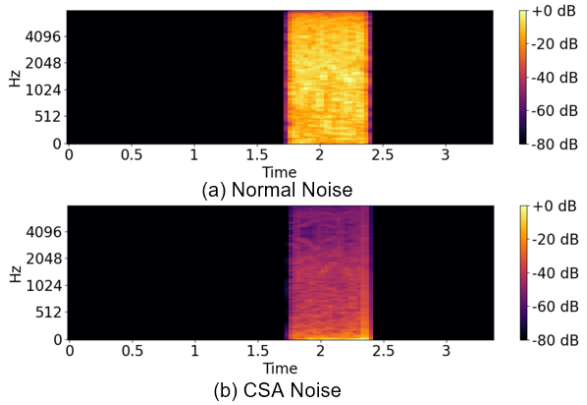

Figure 5: (a) represents the normal noise and (b) represents the CSA noise (ours).

To evaluate CSA's effectiveness on sharp noise, we compared normal and CSA-generated noise. As shown in Fig. 5, spectral analysis indicates that normal noise retains high energy in mid-to-high frequencies, whereas CSA significantly suppresses energy in these bands, reducing harshness and producing a smoother noise profile.

We then examined two key factors: composite sample size and perturbation magnitude. As shown in Table 4, larger composite sizes reduce attack performance due to gradient averaging, which weakens the precision of perturbation direction. Conversely, overly small samples increase storage demands due to high information density (Table 5). Thus, a composite size of 2 achieves a balance between efficiency and performance. For perturbation magnitude, increasing scale consistently improves attack success. We found that a scale of $1/5$ offers the best trade-off between perturbation subtlety and attack effectiveness.

	Composite Sample				
Source	$s=1$	$s=2$	$s=3$	$s=4$	$s=5$
Common	51,603	25,811	17,214	12,915	10,336
TIMIT	47,627	23,823	15,889	11,921	9,541
Librispeech	42,296	21,158	14,112	10,589	8,475

Table 5: The amount of storage (bits) for different composite sample sizes on various datasets.

Conclusion

In this study, we introduce a novel reversible adversarial attack framework for audio privacy protection. IO-RAE leverages LLMs to generate target attack phrases and employs cumulative signal techniques to seamlessly integrate noise. IO-RAE demonstrates excellent attack and recovery performance and has been tested against Google's commercial black-box model, achieving a 100% success rate in non-target attacks. In the future, we plan to explore high-transferability target attack methods in black-box scenarios to enhance its misleading capabilities, thereby achieving more practical and effective audio data protection.

Acknowledgments

This work was supported in part by the Xiamen Research Project for the Natural Science Foundation of Xiamen,China (3502Z202472028), the Xiamen Science and Technology Plan Project (3502Z20231042), the Fundamental Research Funds for the Central Universities(1082204112364) and the Science and Technology Development Fund, Macau SAR, under Grant 0193/2023/RIA3 and 0079/2025/AFJ.

References

Ardila, R.; Branson, M.; Davis, K.; Henretty, M.; Kohler, M.; Meyer, J.; Morais, R.; Saunders, L.; Tyers, F. M.; and Weber, G. 2019. Common voice: A massively-multilingual speech corpus. *arXiv preprint arXiv:1912.06670*.

Baevski, A.; Zhou, Y.; Mohamed, A.; and Auli, M. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33: 12449–12460.

Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Song, S.; Dang, K.; Wang, P.; Wang, S.; Tang, J.; et al. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Battenberg, E.; Chen, J.; Child, R.; Coates, A.; Li, Y. G. Y.; Liu, H.; Satheesh, S.; Sriram, A.; and Zhu, Z. 2017. Exploring neural transducers for end-to-end speech recognition. In *2017 IEEE automatic speech recognition and understanding workshop (ASRU)*, 206–213. IEEE.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901.

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

Champion, P. 2023. Anonymizing speech: Evaluating and designing speaker anonymization techniques. *arXiv preprint arXiv:2308.04455*.

Chen, Z.; Chai, X.; Gan, Z.; Wang, B.; and Zhang, Y. 2024. RAE-VWP: A Reversible Adversarial Example-Based Privacy and Copyright Protection Method of Medical Images for Internet of Medical Things. *IEEE Internet of Things Journal*.

Chou, J.-c.; Yeh, C.-c.; Lee, H.-y.; and Lee, L.-s. 2018. Multi-target voice conversion without parallel data by adversarially learning disentangled audio representations. *arXiv preprint arXiv:1804.02812*.

Cooley, J. W.; and Tukey, J. W. 1965. An algorithm for the machine calculation of complex Fourier series. *Mathematics of computation*, 19(90): 297–301.

Dai, H.; Li, H.; Tian, T.; Huang, X.; Wang, L.; Zhu, J.; and Song, L. 2018. Adversarial attack on graph structured data. In *International conference on machine learning*, 1115–1124. PMLR.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Du, T.; Ji, S.; Li, J.; Gu, Q.; Wang, T.; and Beyah, R. 2020. Sirenattack: Generating adversarial audio for end-to-end acoustic systems. In *Proceedings of the 15th ACM Asia conference on computer and communications security*, 357–369.

Elsayed, G.; Shankar, S.; Cheung, B.; Papernot, N.; Kurakin, A.; Goodfellow, I.; and Sohl-Dickstein, J. 2018. Adversarial examples that fool both computer vision and time-limited humans. *Advances in neural information processing systems*, 31.

Feng, T.; Hashemi, H.; Annavaram, M.; and Narayanan, S. S. 2022. Enhancing privacy through domain adaptive noise injection for speech emotion recognition. In *ICASSP 2022-2022 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, 7702–7706. IEEE.

Gao, Z.; Li, Z.; Wang, J.; Luo, H.; Shi, X.; Chen, M.; Li, Y.; Zuo, L.; Du, Z.; Xiao, Z.; and Zhang, S. 2023. FunASR: A Fundamental End-to-End Speech Recognition Toolkit. In *INTERSPEECH*.

Garofolo, J. S.; Lamel, L. F.; Fisher, W. M.; Fiscus, J. G.; and Pallett, D. S. 1993. DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM. NIST speech disc 1-1.1. *NASA STI/Recon technical report n*, 93: 27403.

Goodfellow, I. J.; Shlens, J.; and Szegedy, C. 2014. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*.

Graves, A.; Fernández, S.; Gomez, F.; and Schmidhuber, J. 2006. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd international conference on Machine learning*, 369–376.

Grootendorst, M. 2020. KeyBERT: Minimal keyword extraction with BERT.

Gulati, A.; Qin, J.; Chiu, C.-C.; Parmar, N.; Zhang, Y.; Yu, J.; Han, W.; Wang, S.; Zhang, Z.; Wu, Y.; et al. 2020. Conformer: Convolution-augmented transformer for speech recognition. *arXiv preprint arXiv:2005.08100*.

Han, W.; Zhang, Z.; Zhang, Y.; Yu, J.; Chiu, C.-C.; Qin, J.; Gulati, A.; Pang, R.; and Wu, Y. 2020. Contextnet: Improving convolutional neural networks for automatic speech recognition with global context. *arXiv preprint arXiv:2005.03191*.

Kavitha, K. K.; Koshti, A.; and Dunghav, P. 2012. Steganography using least significant bit algorithm. *International Journal of Engineering Research and Applications*, 2(3): 338–341.

Kwon, H.; Yoon, H.; and Park, K.-W. 2019. POSTER: Detecting audio adversarial example through audio modification. In *Proceedings of the 2019 ACM SIGSAC conference on computer and communications security*, 2521–2523.

Lin, Z.; Chen, Z.; Fang, Z.; Chen, X.; Wang, X.; and Gao, Y. 2024a. Fedsn: A federated learning framework over heterogeneous leo satellite networks. *IEEE Transactions on Mobile Computing*.

Lin, Z.; Qu, G.; Chen, X.; and Huang, K. 2024b. Split learning in 6G edge networks. *IEEE Wireless Communications*, 31(4): 170–176.

Lin, Z.; Qu, G.; Wei, W.; Chen, X.; and Leung, K. K. 2024c. Adaptsfl: Adaptive split federated learning in resource-constrained edge networks. *arXiv preprint arXiv:2403.13101*.

Lin, Z.; Wei, W.; Chen, Z.; Lam, C.-T.; Chen, X.; Gao, Y.; and Luo, J. 2025a. Hierarchical split federated learning: Convergence analysis and system optimization. *IEEE Transactions on Mobile Computing*.

Lin, Z.; Zhang, Y.; Chen, Z.; Fang, Z.; Wu, C.; Chen, X.; Gao, Y.; and Luo, J. 2025b. Leo-split: A semi-supervised split learning framework over leo satellite networks. *arXiv preprint arXiv:2501.01293*.

Lin, Z.; Zhu, G.; Deng, Y.; Chen, X.; Gao, Y.; Huang, K.; and Fang, Y. 2024d. Efficient parallel split learning over resource-constrained wireless edge networks. *IEEE Transactions on Mobile Computing*, 23(10): 9224–9239.

Liu, J.; Zhang, W.; Fukuchi, K.; Akimoto, Y.; and Sakuma, J. 2023. Unauthorized AI cannot recognize me: Reversible adversarial example. *Pattern Recognition*, 134: 109048.

McAuliffe, M.; Socolof, M.; Mihuc, S.; Wagner, M.; and Sonderegger, M. 2017. Montreal forced aligner: Trainable text-speech alignment using kaldi. In *Interspeech*, volume 2017, 498–502.

Ng, E. G.; Chiu, C.-C.; Zhang, Y.; and Chan, W. 2021. Pushing the limits of non-autoregressive speech recognition. *arXiv preprint arXiv:2104.03416*.

Panayotov, V.; Chen, G.; Povey, D.; and Khudanpur, S. 2015. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, 5206–5210. IEEE.

Qin, Y.; Carlini, N.; Cottrell, G.; Goodfellow, I.; and Raffel, C. 2019. Imperceptible, robust, and targeted adversarial examples for automatic speech recognition. In *International conference on machine learning*, 5231–5240. PMLR.

Radford, A.; Kim, J. W.; Xu, T.; Brockman, G.; McLeavey, C.; and Sutskever, I. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, 28492–28518. PMLR.

Radford, A.; Narasimhan, K.; Salimans, T.; Sutskever, I.; et al. 2018. Improving language understanding by generative pre-training.

Sainath, T. N.; He, Y.; Narayanan, A.; Botros, R.; Pang, R.; Rybach, D.; Allauzen, C.; Variani, E.; Qin, J.; Le-The, Q.-N.; et al. 2021. An Efficient Streaming Non-Recurrent On-Device End-to-End Model with Improvements to Rare-Word Modeling. In *Interspeech*, volume 8, 1777–1781.

Stevens, S. S.; Volkmann, J.; and Newman, E. B. 1937. A scale for the measurement of the psychological magnitude pitch. *The journal of the acoustical society of america*, 8(3): 185–190.

Szegedy, C.; Zaremba, W.; Sutskever, I.; Bruna, J.; Erhan, D.; Goodfellow, I.; and Fergus, R. 2013. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*.

Vos, K.; Sørensen, K. V.; Jensen, S. S.; and Valin, J.-M. 2013. Voice coding with Opus. In *Audio Engineering Society Convention 135*. Audio Engineering Society.

Xiao, C.; Deng, R.; Li, B.; Lee, T.; Edwards, B.; Yi, J.; Song, D.; Liu, M.; and Molloy, I. 2019. Advit: Adversarial frames identifier based on temporal consistency in videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3968–3977.

Xiong, L.; Wu, Y.; Yu, P.; and Zheng, Y. 2023. A black-box reversible adversarial example for authorizable recognition to shared images. *Pattern Recognition*, 140: 109549.

Yakura, H.; and Sakuma, J. 2018. Robust audio adversarial example for a physical attack. *arXiv preprint arXiv:1810.11793*.

Yang, X.; Dong, Y.; Pang, T.; Su, H.; Zhu, J.; Chen, Y.; and Xue, H. 2021. Towards face encryption by generating adversarial identity masks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3897–3907.

Yang, Z.; Li, B.; Chen, P.-Y.; and Song, D. 2018. Characterizing audio adversarial examples using temporal dependency. *arXiv preprint arXiv:1809.10875*.

Yao, Z.; Wu, D.; Wang, X.; Zhang, B.; Yu, F.; Yang, C.; Peng, Z.; Chen, X.; Xie, L.; and Lei, X. 2021. Wenet: Production oriented streaming and non-streaming end-to-end speech recognition toolkit. *arXiv preprint arXiv:2102.01547*.

Yuan, X.; Chen, Y.; Zhao, Y.; Long, Y.; Liu, X.; Chen, K.; Zhang, S.; Huang, H.; Wang, X.; and Gunter, C. A. 2018. {CommanderSong}: a systematic approach for practical adversarial voice recognition. In *27th USENIX security symposium (USENIX security 18)*, 49–64.

Zhang, J.; Wang, J.; Wang, H.; and Luo, X. 2022. Self-recoverable adversarial examples: A new effective protection mechanism in social networks. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(2): 562–574.

Zhao, Z.; Dua, D.; and Singh, S. 2017. Generating natural adversarial examples. *arXiv preprint arXiv:1710.11342*.

Zhao, Z.; Liu, Z.; and Larson, M. 2021. On success and simplicity: A second look at transferable targeted attacks. *Advances in Neural Information Processing Systems*, 34: 6115–6128.

Zhu, J.; Du, X.; Zhou, J.; Pun, C.-M.; Xu, Q.; and Liu, X. 2024. Dp-rae: A dual-phase merging reversible adversarial example for image privacy protection. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 671–680.