Voice Guard: Protecting Voice Privacy with Strong and Imperceptible Adversarial Perturbation in the Time Domain

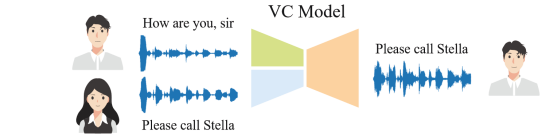
Jingyang Li[1] , **Dengpan Ye**[1*] , **Long Tang**[1] , **Chuanxi Chen**[1] and **Shengshan Hu**[2]
[1]Key Laboratory of Aerospace Information Security and Trusted Computing, Ministry of Education, School of Cyber Science and Engineering, Wuhan University
[2]Huazhong University of Science and Technology
{ljy_miao, yedp, l_tang, chencx}@whu.edu.com, hushengshan@hust.edu.cn

Abstract

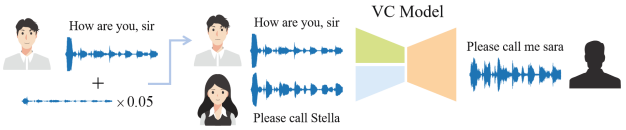
Adversarial example is a rising tool for voice privacy protection. By adding imperceptible noise to public audio, it prevents tampers from using zero-shot Voice Conversion (VC) to synthesize high quality speech with target speaker identity. However, many existing studies ignore the human perception characteristics of audio data, and it is challenging to generate strong and imperceptible adversarial audio. In this paper, we propose the Voice Guard defense method, which uses a novel method to advance the adversarial perturbation to the time domain to avoid the loss caused by cross-domain conversion. And the psychoacoustic model is introduced into the defense of VC for the first time, which greatly improves the disruption ability and concealment of adversarial audio. We also standardize the evaluation metrics of adversarial audio for the first time, combining multi-dimensional metrics to define the criteria for defense. We evaluate Voice Guard on several state-of-the-art zero-shot VC models. The experimental results show that our method can ensure the perceptual quality of adversarial audio while having a strong defense capability, and is far superior to previous works in terms of disruption ability and concealment.

1 Introduction

With the development of deep learning, it is easier for deep neural networks to fake realistic data, which has many applications in film and television creation [Sinha *et al.*, 2022; Perov *et al.*, 2020] , audio reading [Huang *et al.*, 2022; Ye *et al.*, 2022] , text generation [Mu and Li, 2022; Shu *et al.*, 2021] and other fields. However, as shown in Fig.1(a), some criminals employ VC, a technique that alters the speaker's identity while maintaining the speech content unchanged, in order to fabricate high-quality speech and perpetrate large-scale network voice fraud, which poses a serious threat to citizen privacy protection and security authentication [Wenger *et al.*, 2021]. In particular, zero-shot VC models such as AdaIN-VC [Chou *et al.*, 2019], VQMIVC [Wang *et al.*, 2021] are the most dangerous, since they can convert voices between any two speakers. Although there has been a lot of research on the detection of forged audio [Yamagishi *et al.*, 2021; Wang and Yamagishi, 2021; Jung *et al.*, 2022; Yi *et al.*, 2022], this passive defense method can only reduce the damage of the attack after it occurs. Therefore, this paper aims to resist such attacks in an active fashion [Ruiz *et al.*, 2020] as shown in Fig.1(b), which makes a precise and comprehensive defense available to the public before the attack occurs.



(a) Tampering of speaker identity based on VC.



(b) Active defense against VC. Disrupting with speaker identity and speech content of converted audio.

Figure 1: Description of the forgery and defense scenario.

Adversarial examples can change the output of a neural network using imperceptible noise [Yuan *et al.*, 2019], and have shown excellent disruption performance on image [Xie *et al.*, 2020] and speech [Qin *et al.*, 2019] classification tasks. In contrast, there are few studies on adversarial perturbation for generative models, and it is difficult to achieve the same level of disruption performance as the former. We try to add adversarial noise to the target audio that does not affect the original content, so that the defended audio loses its original feature information after performing voice conversion to achieve active defense.

Typically, there are two ways to produce adversarial audio: by adding adversarial perturbations to a waveform in the time domain, as well as by adding adversarial perturbations to frequency domain feature spectrograms such as Mel and Mel-scale Frequency Cepstral Coefficients (MFCC) [Kassis and Hengartner, 2021]. However, the adversarial audio generated in the frequency domain needs to be restored to the

*Corresponding author

speech waveform by the vocoder, and this upsampling process will weaken the disruption performance of the adversarial audio. Therefore, we move up the perturbation process from frequency domain to time domain to maintain the disruption performance of adversarial audio.

In the production of adversarial audio, researchers often directly regard the time domain or frequency domain audio data as images, and use the same generation steps as the adversarial image. However, human eyes utilize a different signal processing mechanism than human ears. Therefore, adversarial noise invisible to the human eye is not necessarily inaudible to the human ear [Lin and Abdulla, 2015]. In order to ensure the imperceptibility of adversarial noise while implementing powerful disruption, we use the masking effect in acoustics, which is a phenomenon that the larger energy signal masks the smaller energy signal with similar frequency and time to make it inaudible. We introduce a psychoacoustic model to constrain adversarial noise and improve its concealment.

Previous works often evaluate the disruption ability and concealment of adversarial audio separately, which leads to adversarial audio being either weak or perceptible. Only strong and imperceptible adversarial examples are effective. In addition, the evaluation metrics of voice conversion performance should not be limited to the change in speaker identity. Speech quality and speech content are also significant indicators. Therefore, we combine multiple dimensions and simultaneously consider the concealment and disruption ability of adversarial audio, and propose the adversarial audio evaluation metrics for voice conversion for the first time.

Overall, this work makes the following contributions:

(1) We propose Voice Guard, a strong and imperceptible voice privacy protection method that enhances disruption ability through a novel time-domain perturbation. It also improves the concealment of the perturbation by introducing a psychoacoustic model into the voice conversion defense for the first time.

(2) We first standardize the performance evaluation metrics of adversarial examples for voice conversion models, and evaluate the concealment and disruption ability of adversarial examples in multiple dimensions.

(3) We verify the effectiveness of Voice Guard through systematic and comprehensive experiments, obtaining a state-of-the-art defense success rate of over 80%. Compared with previous works, Voice Guard has been significantly improved in disruption ability and concealment.

2 Related Work

2.1 Voice Conversion

The state-of-the-art zero-shot voice conversion model not only breaks through the need for parallel datasets, but also performs voice conversion between arbitrary speakers, synthesizing sufficiently realistic audio. AdaIN-VC [Chou *et al.*, 2019] is based on a simple autoencoder structure, and introduces adaptive instance normalization in the content encoder and decoder to assist the separation of speaker information and content information. To further improve the model's disentanglement ability, VQMIVC [Wang *et al.*, 2021] introduces mutual information as a constraint based on the traditional feature disentanglement method.

2.2 Adversarial Examples for Generative Model

Kos et al. [2018] propose three defense flows to generate adversarial examples against VAE, VAE-GAN and other generative models in the image reconstruction task. In experiments, the researchers find that misclassification does not necessarily lead to the reconstruction of the target class. This indicates that the classifier is more vulnerable to adversarial examples than the generative model. This idea is also tested by the work of Joshi et al [2021]. Researchers find that in speaker classification tasks, compared with traditional defense methods such as random smoothing and adversarial training, using generative models such as GAN, VAE and vocoder to preprocess audio can more effectively defend against adversarial examples. Huang et al. [2021] transfer Kos et al.'s work from image generation task to voice conversion task and demonstrate the feasibility of disrupting voice conversion. However, their method has major defects: it cannot balance the disruption ability and concealment of adversarial audio; speech quality and speech content are not included in the evaluation index.

2.3 Psychoacoustic Model

The emergence of psychoacoustic models [Lin and Abdulla, 2015], which is a quantitative model closely matching the auditory mechanism, promotes research on the concealment of adversarial audio. Researchers can analyze auditory thresholds using empirically determined masking models. Qin et al. [2019] are the first to introduce a psychoacoustic model to disrupt neural network-based speech recognition systems. In order to avoid the perception of adversarial noise, they calculate the masking threshold using a psychoacoustic model and implement perturbations below the masking threshold. Wang et al. [2020] disrupt x-vector based speaker recognition systems using psychoacoustic models. Their method uses a masking threshold instead of the common l_p norm to limit the strength of adversarial perturbations. This allows them to achieve an excellent disruption effect while ensuring the imperceptibility of adversarial examples.

3 Threat Model

The tamper collects high-quality audio of citizens on the public network, and then uses zero-shot VC to generate forged audio of the target speakers for telecom fraud. Formally, let $F(\cdot,\cdot)$ be the VC model used by the tamper and $ASV(\cdot)$ be the speaker classification model. Speakers X, Y, and T are the providers of audio $\hat{x}$, $\hat{y}$, and $\hat{t}$ respectively. These speakers are defined consistently throughout, but play different roles for different systems.

When the tamper attacks speaker X using VC system, audio $\hat{x}$ acts as the target speaker and provides the speaker information. Audio $\hat{t}$ acts as the source speaker and provides the speech content. The choice of speaker T is arbitrary and unimportant, because audio $\hat{t}$ only needs to provide specific content and can come from anyone. When the attack succeeds, the tamper gets forged audio with the identity of the target speaker X. That is, $ASV(F(\hat{x},\hat{t})) = X$.

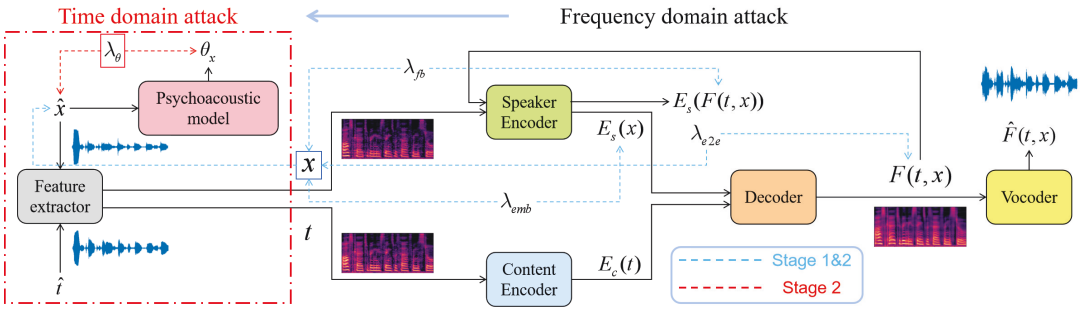


Figure 2: Overview of the Voice Guard defense pipeline. The speech $\hat{t}$ and $\hat{x}$ are extracted by feature extraction to obtain the spectrogram t and x. $E_c(t)$ and $E_s(x)$ are the content features and speaker features extracted by the content encoder and the speaker encoder, respectively. The decoder combines $E_c(t)$ and $E_s(x)$ to generate a spectrogram $F(t,x)$, and then restores it to a speech waveform $\hat{F}(t,x)$ through the vocoder. θ_x is the masking threshold for speech $\hat{x}$ calculated by the psychoacoustic model.

To prevent this forgery, we propose Voice Guard, which prevents zero-shot VC from generating audio with speaker X identity information by adding time domain adversarial noise $\hat{\delta}$ to audio $\hat{x}$ beforehand, such that, $ASV(F(\hat{x}+\hat{\delta},\hat{t})) \neq X$. We adopt the concept of least likely class in adversarial examples and use it to guide the generation of noise $\hat{\delta}$ by selecting the speaker Y who is the most dissimilar to speaker X.

4 Methodology

Zero-shot VC models are usually composed of a speaker encoder, a content encoder, and a decoder. With the help of a known zero-shot VC model, our proposed Voice Guard generates strong and imperceptible adversarial audio in two stages. In stage 1, we generate adversarial noise strong enough to perturb the zero-shot VC, and in stage 2, we hide it. Fig.2 shows an overview of the Voice Guard defense pipeline.

4.1 Strong Adversarial Audio

Based on previous research, there are three methods to defend voice conversion models based on adversarial audio: end-to-end defense (*e2e*), feedback defense (*fb*), and embedding defense (*emb*). *emb* can change the output of the speaker encoding to protect the speaker identity of the input audio. *e2e* directly alters the output of the voice conversion model. *fb* transmits the output result back to the speaker encoder to ensure that the speaker information can be changed in the converted spectrogram. All three defenses add adversarial noise in the frequency domain and rely on speaker features to generate targeted disruption. Among them, *emb* has the highest efficiency and *fb* has the highest success rate. The above defense flows can be expressed as follows,

$$\begin{aligned}
\min_{w} \mathcal{L}_{emb}(x,y,\delta) &= \mathcal{L}(E_s(x+\delta),E_s(y)) \\
&\quad -\lambda\mathcal{L}(E_s(x+\delta),E_s(x)) \\
\min_{w} \mathcal{L}_{e2e}(x,y,\delta,t) &= \mathcal{L}(F(t,x+\delta),F(t,y)) \\
&\quad -\lambda\mathcal{L}(F(t,x+\delta),F(t,x)) \\
\min_{w} \mathcal{L}_{fb}(x,y,\delta,t) &= \mathcal{L}(E_s(F(t,x+\delta)),E_s(y)) \\
&\quad -\lambda\mathcal{L}(E_s(F(t,x+\delta)),E_s(x)) \\
subject\ to\ \ \delta &= \epsilon\cdot tanh(w)
\end{aligned} \tag{1}$$

where $w \in \mathbb{R}^{M\times L_f}$, M and L_f are the total number of frequency components and time frames respectively. w uses the variable transformation method [Carlini and Wagner, 2017] of the $tanh(\cdot)$ function to effectively constrain the frequency domain noise δ to stay in the range $[-\epsilon,\epsilon]$. $\mathcal{L}(\cdot,\cdot)$ is the distance between two vectors or two feature spectrograms, and ϵ is a constraint on the strength of the adversarial noise. x, y and t are frequency domain features of audio $\hat{x}$, $\hat{y}$ and $\hat{t}$ converted from time domain, respectively. The expression of each defense flow has two terms. The first item makes the output speech sound as if it was uttered by speaker Y. The second item aims to remove the identity of the speaker X. λ is a positive hyperparameter that balances these two terms.

However, previous defenses have certain limitations. The adversarial examples generated in the frequency domain need to be restored to waveforms through lossy upsampling, and this process will reduce the disruption ability of the adversarial examples. To avoid the loss caused by this cross-domain conversion, we propose to replace the frequency domain adversarial noise δ with the time domain adversarial noise $\hat{\delta}$. But since the data input to the model are often spectrograms of acoustic features such as Mel and MFCC, methods that provide extraction of such features, such as librosa, do not support backward gradient propagation. To implement our proposed defense, we need a way to backpropagate gradients computed in the frequency domain to the time domain. We re-implement the feature extractor $H(\cdot)$ that supports gradient backpropagation in Pytorch [Paszke *et al.*, 2019]. Formally, $x = H(\hat{x})$. To propagate the gradient back to the original signal layer, we apply the chain rule:

$$\frac{\partial F(H(\hat{t}),H(\hat{x}+\hat{\delta}))}{\partial\hat{\delta}} = \frac{\partial F(H(\hat{t}),H(\hat{x}+\hat{\delta})))}{\partial H(\hat{\delta})}\cdot\frac{\partial H(\hat{\delta})}{\partial\hat{\delta}} \tag{2}$$

Based on Eq.2, we are able to construct adversarial noise from the frequency domain to the time domain, and it is effective for all three defense methods mentioned in Eq.1. Considering the defense effect and production efficiency of adversarial examples, we propose the Voice Guard defense method based on embedding defense. The generation method of adversarial examples can be expressed as follows,

$$\min_{\hat{w}} \mathcal{L}_{vc}(\hat{x}, \hat{y}, \hat{\delta}) = \mathcal{L}(E_s(H(\hat{x}+\hat{\delta})), E_s(H(\hat{y}))) - \lambda \mathcal{L}(E_s(H(\hat{x}+\hat{\delta})), E_s(H(\hat{x}))) \quad (3)$$
$$subject\ to\ \ \hat{\delta} = \epsilon \cdot tanh(\hat{w})$$

where $\hat{w} \in \mathbb{R}^{1 \times L_t}$, and L_t is the length of the waveform $\hat{x}$.

4.2 Imperceptible Adversarial Audio

The imperceptibility of adversarial audio is very crucial in real scenarios. The tamper will select clean audio examples for speaker identity tampering, and if the generated adversarial audio sounds noisy, it will be directly screened by the tamper. Consequently, the noise needs to be constrained based on the human perception mechanism for audio signals.

To exploit masking effects, we constructed psychoacoustic models [Lin and Abdulla, 2015] using Numpy [Van Der Walt *et al.*, 2011]. The calculation of the masking threshold requires the use of the normalized power spectral density (PSD), which is calculated for the input audio $\hat{x}$ and can be expressed as follows,

$$P_x(k) = 10log_{10}|\frac{1}{N}s_x(k)|^2 \quad (4)$$

$$\bar{P}_x(k) = 96 - max\{P_x(k)\} + P_x(k) \quad (5)$$

where $s_x(k)$ is the output of $\hat{x}$ with frequency k after passing the short-time Fourier Transform (STFT). The calculation of the masking threshold can be expressed as follows,

$$\theta_x = P(\bar{P}_x(k)) \quad (6)$$

where $P(\cdot)$ is the psychoacoustic model, detailed calculation steps are given in **Appendix A**, and θ_x is the masking threshold of the input audio $\hat{x}$. In order to construct imperceptible adversarial audio, we constrain the generated adversarial perturbations based on a masking threshold, which can be calculated as follows:

$$\mathcal{L}_\theta(\hat{x}, \hat{\delta}) = \mathbb{E}_k max\{\bar{P}_\delta(k) - \theta_x(k)\} \quad (7)$$

4.3 Strong and Imperceptible Adversarial Audio

Combining the methods proposed in 4.1 and 4.2, we implement Voice Guard, a strong and imperceptible voice privacy protection method. We formulate the problem of computing time-domain adversarial noise $\hat{\delta}$ as minimizing a loss function $\mathcal{L}(\hat{x}, \hat{y}, \hat{\delta})$, defined as follows:

$$\min_{\hat{w}} \mathcal{L}(\hat{x}, \hat{y}, \hat{\delta}) = \mathcal{L}_{vc}(\hat{x}, \hat{y}, \hat{\delta}) + \alpha \cdot \mathcal{L}_\theta(\hat{x}, \hat{\delta})$$
$$subject\ to\ \ \hat{\delta} = \epsilon \cdot tanh(\hat{w}) \quad (8)$$

The adaptive parameter α is used to balance the weight of disruption ability and concealment. In the Eq.8, the first term $\mathcal{L}_{vc}$ prompts the audio synthesized by the voice conversion model to have speaker features of audio $\hat{y}$. The second term $\mathcal{L}_\theta$ forces the normalized PSD estimate of the adversarial perturbation $\bar{P}_\delta(k)$ to be below the frequency masking threshold $\theta_x(k)$ of the original audio.

Algorithm 1 Adversarial noise generation for Voice Guard

Input: $\hat{x}$ (defend wav), $\hat{y}$ (defend target wav)
Parameter: T_1/T_2 (defense step 1/2 iterations), F (Number of binary searches)
Output: $\hat{\delta}$ (time domain noise)

```
Random Init δ̂ ~ N(0,1)
for t1 in T1 do
    δ̂ ← δ̂ − lr1 · ∇_δ̂ L_vc(x̂, ŷ, δ̂)
end for
for f in F do
    for t2 in T2 do
        δ̂ ← δ̂ − lr2 · ∇_δ̂ L(x̂, ŷ, δ̂)
    end for
    Update α by Binary Search
end for
return δ̂
```

The whole defense process of Voice Guard is divided into two defense stages. In stage 1, we consider the disruption ability of the perturbation. The performance of voice conversion is greatly reduced by using the Eq.3 to construct strong adversarial audio. In stage 2, we hide the strong adversarial audio generated in stage 1 by introducing the loss of the masking threshold (Eq.7). We perform this stage of defense several times and judge whether the defense is successful or not, and then update the adaptive parameter α using binary search according to the results, until we find the largest α that can successfully defend. The adversarial noise generation of Voice Guard can be expressed as algorithm 1.

5 Experiments

5.1 Data Preparation

We perform experiments on AdaIN-VC [Chou *et al.*, 2019] and VQMIVC [Wang *et al.*, 2021], which are the state-of-the-art any-2-any zero-shot voice conversion models.

We use the CSTR VCTK corpus [Veaux *et al.*, 2017] in our experiments, where we randomly select 20 speakers as objects of defense, and each speaker has 100 utterances. For each speaker, a speaker classifier is used to find the most unlikely speaker identity, and 100 utterances of the speaker are selected as the direction of disruption.

In defense stage 1, we use the Adam [Kingma and Ba, 2014] optimizer with learning rate 0.001 for 3000 iterations and initialize $\hat{\delta}$ to a random vector that fits the normal distribution $N(0, 1)$. In defense stage 2, we initialize the adaptive parameter α to 1.0 and perform 1500 iterations.

5.2 Evaluation Metrics

In the performance evaluation of adversarial audio, disruption ability and concealment are equally critical and should be evaluated together. Additionally, the performance of VC takes into account not only the speaker identity of the converted audio, but also the naturalness and completeness.

We propose a comprehensive evaluation metric that combines disruption ability and concealment for adversarial audio on voice conversion. Our metrics are speaker identity, speech

Model	Method	Defend	MCD	ASV	MOS	CER	WER
AdaIN-VC	*rec*	0.00%	0.74	100.00% → 100.00%	4.53 → 3.73	3.01% → 6.50%	8.09% → 15.63%
	emb	57.19%	3.71	76.01% → 19.65%	3.06 → **2.81**	4.55% → 8.26%	10.37% → 18.41%
	fb	64.65%	**3.23**	90.39% → 25.75%	3.52 → 2.98	4.27% → 8.43%	9.58% → 18.54%
	e2e	55.39%	4.10	83.60% → 29.04%	3.64 → 2.93	3.90% → 8.39%	9.87% → 18.30%
	VG	**86.40**%	4.39	**99.17**% → **12.76**%	**3.65** → 3.12	**3.84**% → **9.05**%	**9.17**% → **20.28**%
VQMIVC	*rec*	0.00%	1.01	92.86% → 92.86%	3.98 → 3.60	1.28% → 12.01%	3.18% → 23.90%
	emb	77.94%	4.57	84.79% → 7.78%	3.79 → **3.56**	3.29% → 15.41%	7.95% → **29.90**%
	fb	40.17%	**4.52**	80.06% → 48.32%	**3.82** → 3.58	3.91% → 14.71%	8.07% → 21.11%
	e2e	53.65%	4.64	81.37% → 32.48%	3.62 → 3.58	2.55% → 13.67%	7.17% → 26.70%
	VG	**94.44**%	5.40	**100.00**% → **5.56**%	3.51 → 3.75	**2.52**% → **15.88**%	**7.06**% → 28.90%
AdaIN-VC to VQMIVC	*emb*	60.07%	3.58	76.19% → 20.25%	3.08 → 3.70	4.56% → 15.68%	11.85% → 29.60%
	fb	69.38%	**3.30**	85.93% → 18.63%	3.56 → 3.73	3.58% → 14.58%	9.57% → 27.60%
	e2e	69.84%	3.43	82.22% → 18.25%	3.67 → **3.70**	3.25% → 14.37%	**8.93**% → 27.51%
	VG	**88.83**%	4.68	**97.22**% → **8.39**%	**3.69** → 3.71	**2.90**% → **16.50**%	9.01% → **32.86**%
VQMIVC to AdaIN-VC	*emb*	38.33%	4.60	87.26% → 49.76%	3.82 → 2.82	2.61% → 6.60%	5.73% → 16.48%
	fb	12.88%	**4.52**	89.96% → 78.79%	**3.82** → 3.01	2.81% → 7.52%	5.91% → 17.32%
	e2e	21.59%	4.69	80.83% → 60.08%	3.62 → 2.81	4.17% → 7.54%	9.50% → 16.80%
	VG	**51.84**%	5.27	**98.33**% → **46.49**%	3.50 → **2.65**	**1.76**% → **16.08**%	**4.63**% → **30.69**%

Table 1: Defense performance comparison between the proposed method and the baseline methods. The best results are highlighted in **bold**. *Defend* refers to the defense success rate of adversarial audio, and *ASV* stands for the accuracy of speaker classification. *MCD* is the distortion of defended aduio. The larger the value, the higher the distortion. *MOS* is the predicted Mean Opinion Score with values ranging from $[0, 5]$, where higher values represent better subjective perceived quality. *CER/WER* is the Character/Word Error Rate, and the lower the value is, the more similar the speech content is to the original audio. The data with arrows indicate the change in the corresponding metrics from the defended audio to the converted audio. *rec* refers to the preprocessing and reconstruction steps of the corresponding VC model, and *VG* stands for Voice Guard defense.

naturalness, and speech content. We set a threshold for each index, and an index exceeding the threshold is considered to have changed greatly. Only adversarial audio with high concealment and strong jamming capability is considered to provide an effective defense.

We use GE2E [Wan *et al.*, 2018] based speaker classifier, MOSNet [Lo *et al.*, 2019] speech quality prediction model and Whisper [Radford *et al.*, 2022] speech recognition model to measure the above indicators. The definition of adversarial audio defense can be expressed as follows,

$$F_{imp} = \begin{cases} True & \text{if } \begin{array}{c} S_{df} \text{ is } ASV(W_{df}) \text{ and} \\ MOS(W_{df}) \geq MOS_{th} \text{ and,} \\ ASR(W_{df}) \geq ASR_{th} \end{array} \\ False & \text{otherwise,} \end{cases} \tag{9}$$

$$F_{str} = \begin{cases} True & \text{if } \begin{array}{c} S_{df} \text{ is not } ASV(W_{vc}) \text{ or} \\ MOS(W_{vc}) < MOS_{th} \text{ or,} \\ ASR(W_{vc}) < ASR_{th} \end{array} \\ False & \text{otherwise,} \end{cases} \tag{10}$$

$$F_d = F_{imp} \text{ and } F_{str} \tag{11}$$

where $ASV(\cdot)$ is the speaker classification model, which can extract speaker features from the audio, compare them with the registered speaker database, and return the speaker with the highest feature similarity. $MOS(\cdot)$ is a speech quality prediction model, which can evaluate audio quality. $ASR(\cdot)$ is a speech recognition model that evaluates the completeness of the speech content. MOS_{th} and ASR_{th} are the audio quality threshold and information integrity threshold, respectively. Audio above the threshold is considered acceptable, otherwise it is considered unacceptable. S_{df} is the protected speaker. W_{df} and W_{vc} are the defended audio and the converted audio, respectively. F_d is the flag that determines whether the defense is successful or not.

In order to make a fair comparison between the adversarial examples generated by different methods, we use MEL Cepstral distortion (MCD) to compare the performance of adversarial examples with the same distortion. We calculate the defense success rate $D_{Acc} = N_s/N$ by counting the number of defenses N that appear in a certain distortion interval and the number N_s of successful defense markers F_d.

5.3 Defense Performance Evaluation

We compare the defense performance of Voice Guard with the only previous work, attack-vc [Huang *et al.*, 2021], in white-box and black-box scenarios. In the white-box scenario, we know the structure and all parameters of the voice conversion model, and directly generate adversarial examples of the target voice conversion model. In the black-box scenario, we know nothing about the voice conversion model, and can only transfer the adversarial examples generated in the white-box scenario to this scene for defense. To reduce the impact of VC model performance, we screen out samples that fail to perform VC correctly when $\epsilon = 0$. Thus, when $\epsilon = 0$, the defense success rate of the remaining audio is 0.00%.

Tab.1 shows the performance comparison between the Voice Guard and the baseline under white and black box scenarios. In order to facilitate comparison with the traditional evaluation system, this experiment does not introduce MOS_{th} and ASR_{th}. Instead, it only uses speaker identity as the evaluation criterion for defense success.

Model	Method	Defend	MCD	ASV	MOS	CER	WER
white box	*emb*	33.33%	3.70	76.73% → 19.03%	3.08 → **2.84**	4.63% → 8.61%	10.58% → 19.36%
	fb	44.15%	3.26	89.86% → 26.88%	3.56 → 3.02	3.43% → 8.77%	8.92% → 19.04%
	e2e	48.22%	**3.15**	83.71% → 28.82%	3.68 → 2.95	3.79% → 8.54%	9.23% → 18.57%
	VG w/o P	75.90%	5.86	99.17% → **10.41%**	3.71 → 3.14	3.09% → **9.16%**	7.86% → **19.54%**
	VG	**80.06%**	4.34	**99.17%** → 13.55%	**4.07** → 3.27	**3.05%** → 8.89%	**7.84%** → 18.91%
black box	*emb*	30.79%	3.58	74.70% → 19.81%	3.06 → 3.69	4.73% → **15.42%**	11.83% → 28.74%
	fb	40.85%	3.33	86.16% → 19.01%	3.52 → 3.72	3.53% → 14.75%	9.62% → 27.67%
	e2e	37.13%	**3.15**	80.44% → 18.21%	3.63 → 3.71	3.27% → 14.55%	9.13% → 27.65%
	VG w/o P	72.63%	5.75	97.22% → **12.53%**	3.77 → **3.63**	3.13% → 13.79%	8.93% → 27.30%
	VG	**81.62%**	4.23	**97.22%** → 13.22%	**4.13** → 3.70	**2.31%** → 15.26%	**7.40%** → **30.92%**

Table 2: Comparison of defense performance under more stringent evaluation metrics. The best results are highlighted in **bold**. *VG w/o P* represents the defense method that only advances the perturbation to the time domain and does not introduce the psychoacoustic model.

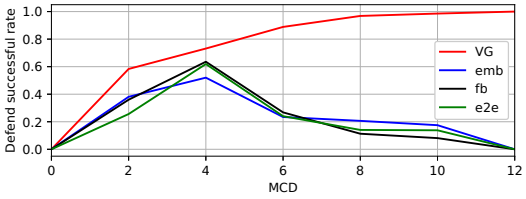


Figure 3: Trend of defense success rate variation with MCD distortion. The defense success rate of our proposed method significantly surpasses the baseline method at all distortion stages.

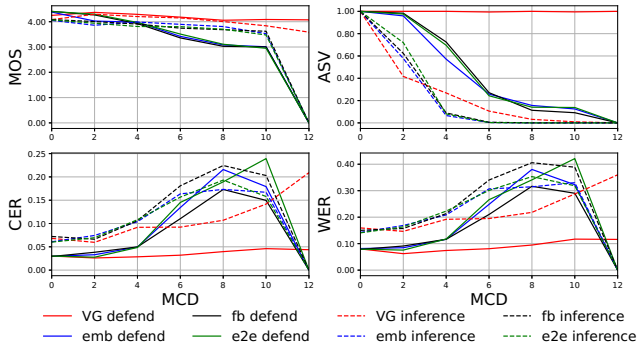


Figure 4: Trend of speech features variation with MCD distortion. The solid and dashed lines with the same color represent the performance of the defended audio and the converted audio of the same defense method under the corresponding metrics.

Obviously, Voice Guard has a significant improvement in the defense success rate. It is well known that any defense method will destroy the original information of the audio to varying degrees. However, compared with the baseline method, on the one hand, Voice Guard destroys the original information of the audio, especially the speaker identity, to a lesser extent. On the other hand, the audio defended by Voice Guard will produce a greater degree of loss in the feature information contained after performing voice conversion. Therefore, it can be considered that the adversarial audio constructed by Voice Guard has a significant improvement in terms of disruption ability and concealment.

It is worth noting that the Voice Guard defense does not perform as well as the feedback defense in the MCD metric. This is because MCD is a distortion calculation method based on the frequency domain features of the audio frequency. As a result, the noise directly added in the frequency domain leads to a lower MCD distortion. In addition, the feedback defense sacrifices efficiency and directly considers the speaker characteristics of the converted audio as the disruption target, so the MCD index performs better. Despite the fact that adding noise in the time domain results in large MCD distortions, the psychoacoustic model prevents the additional noise from adversely affecting audio quality. To further analyze the effectiveness of our proposed method, we study the trend of each indicator under different distortion levels.

Fig.3 shows the trend of the defense success rate of different methods as the distortion degree changes for the white-box scenario of the AdaIN-VC model. It can be seen that the defense success rates of the baseline methods all show a trend of first increasing and then decreasing. This is because when the distortion is low, the strength of the added adversarial noise is low. Thus, it is difficult to change the feature information of the converted audio. When the distortion is too large, the added adversarial noise is too powerful, which seriously affects the quality of adversarial audio and makes it lose its concealment. Only when the distortion degree is in the right position, the adversarial perturbation can change the characteristics of the converted audio without affecting the original audio characteristics. This enables a defense success rate of around 60%.

However, the defense success rate of Voice Guard only has an upward trend but no downward trend, which is due to the fact that we directly add adversarial noise in the time domain in order to avoid cross-domain conversion. Therefore, both the original audio features and the generated adversarial noise are completely preserved in this process. The introduction of the psychoacoustic model further reduces the distortion caused by the addition of adversarial noise. The concealment of adversarial audio is significantly improved. The changes in MOS, ASV, CER and WER indicators in Fig.4 confirm this view. The characteristics of the audio defended by our proposed method (red solid line) change slowly, especially in ASV and MOS indicators, and the red solid line is almost a horizontal line. It is not affected by the increase in distortion. It should be mentioned that when MCD reaches 12, all baseline defended audio is considered as a defense failure, so we denote the corresponding index by 0.

Voice Guard does not reduce the disruption ability of adversarial audio due to the improvement of their concealment. From the trend of MOS and ASV index in Fig.4, it can be seen that the audio defended by Voice Guard still shows large characteristic differences after voice conversion. It can be believed that the Voice Guard improves concealment while ensuring disruption ability, so the defense success rate has been improved by leaps and bounds.

5.4 Ablation Experiment

Since the psychoacoustic model can only be applied to a defense launched from the time domain, we test the performance changes due to perturbations only in the time domain in AdaIN-VC's white and black box scenarios. We add restrictions on MOS and ASR metrics according to Sec.5.2, where $MOS_{th} = 3$ and $ASR_{th} = 10\%$.

By comparing the performance of the baseline in Tab.2 and 1, we find that after considering the subjective perceived quality and the completeness of speech content, the defense performance of the baseline decreases substantially. However, Voice Guard still shows satisfactory defense performance in the face of more stringent evaluation systems.

As can be seen, the defense success rate of the *VG w/o P* method improves considerably. The features before conversion are more similar to the original audio, and the features after conversion have a larger disturbance amplitude. This is because some information of audio will be lost after cross-domain conversion, and high-level features such as speaker identity will be offset. On the other hand, adversarial examples also suffer from loss after experiencing cross-domain transformation, resulting in decreased disruption performance. The process of cross-domain conversion can be eliminated by advancing the perturbation to the time domain, so the concealment and disruption ability of the adversarial audio generated by this method are increased.

The problem is that directly adding adversarial perturbations in the time domain is equivalent to adding more noise sources. This leads to greater MCD distortion and the presence of noise is more easily perceived by humans. Therefore, we introduce psychoacoustic models based on this to hide adversarial noise using high-energy signals in raw audio. The data show that the features before conversion, especially MOS, have been greatly improved, and the concealment of adversarial audio has been further improved.

6 Conclusion

Adding strong and imperceptible noise to audio to defend against voice conversion-based identity tampering is a highly challenging task. In this paper, the Voice Guard defense method is based on improving the disruption ability and concealment of adversarial audio. This is accomplished by advancing the perturbation from frequency domain to time domain and introducing a psychoacoustic model, which greatly improves the success rate of defense. In order to meet the needs of real-world scenarios, we also propose a more stringent evaluation system for adversarial audio, and compare the defense performance of several state-of-the-art zero-shot voice conversion models with baseline methods. The results are encouraging. In future work, we will further improve the robustness of Voicd Guard in black-boxe scenarios.

A Psychoacoustic Model

The computation of the masking threshold consists of 3 steps:

A.1 STEP 1: Identifications of Maskers

The normalized PSD estimate of reasonable maskers must satisfy three constraints. First they need to be local maxima,

$$\bar{P}_x(k) \geq \bar{P}_x(k+1) \text{ and } \bar{P}_x(k) \geq \bar{P}_x(k-1) \tag{12}$$

Secondly, they should be larger than the absolute threshold of hearing (ATH). $ATH(f)$ is approximated by the following frequency dependence function:

$$\begin{aligned} ATH(f) = 3.64(\frac{f}{1000})^{-0.8} + 10^{-3}(\frac{f}{1000})^4 \\ -6.5exp\{-0.6(\frac{f}{1000} - 3.3)^2\} \end{aligned} \tag{13}$$

Finally, they must be the highest within 0.5 Bark of the masked frequency. Bark is a psychoacoustic driven frequency scale, and $b(f)$ represents the Bark scale at frequency f, which is related to frequency as follows:

$$b(f) = 13arctan(\frac{0.76f}{1000}) + 3.5arctan(\frac{f}{7500})^2 \tag{14}$$

A.2 STEP 2: Calculation of Individual Masking Thresholds

After the masks are identified, the masking threshold in the frequency domain needs to be calculated for each mask. Since the spread function of the mask is similar under different Bark, we calculate the masking threshold using the dual-slope diffusion function at the Bark scale,

$$SF[b(i), b(j)] = \begin{cases} 27\Delta b_{ij}, & \text{if } \Delta b_{ij} \leq 0 \\ G(b(i)) \cdot \Delta b_{ij}, & \text{otherwise} \end{cases} \tag{15}$$

where $G(b(i)) = [-27 + 0.37max\{\bar{P}_x(b(i)) - 40, 0\}]$, $\Delta b_{ij} = b(j) - b(i)$, $b(i)$ and $b(j)$ are Bark scales with frequencies $i$and j, respectively. $T[b(i), b(j)]$ represents the masking threshold of frequency j for frequency i, which can be calculated as follows:

$$T[b(i), b(j)] = \bar{P}_x(b(i)) + \Delta_m[b(i)] + SF[b(i), b(j)] \tag{16}$$

$$\Delta_m[b(i)] = -0.6025 - 0.275b(i) \tag{17}$$

A.3 Step 3: Global Masking Threshold

Finally, the final global masking threshold is obtained by stacking the masking threshold and the silencing threshold of each mask in the logarithmic domain,

$$\theta_x(i) = 10log_{10}[10^{\frac{ATH(i)}{10}} + \sum_{j=1}^{N_m} 10^{\frac{T[b(i),b(j)]}{10}}] \tag{18}$$

where N_m is the number of maskers and the calculated θ_x is the frequency masking threshold of the input audio $\hat{x}$.

Acknowledgments

This work was supported in part by National Natural Science Foundation of China NSFC (62072343), and the National Key Research and Development Program of China (2019QY(Y)0206).

References

[Carlini and Wagner, 2017] 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.

[Chou *et al.*, 2019] Ju-chieh Chou, Cheng-chieh Yeh, and Hung-yi Lee. One-shot voice conversion by separating speaker and content representations with instance normalization. *arXiv preprint arXiv:1904.05742*, 2019.

[Huang *et al.*, 2021] Chien-yu Huang, Yist Y Lin, Hung-yi Lee, and Lin-shan Lee. Defending your voice: Adversarial attack on voice conversion. In *2021 IEEE Spoken Language Technology Workshop (SLT)*, pages 552–559. IEEE, 2021.

[Huang *et al.*, 2022] Rongjie Huang, Max W. Y. Lam, Jun Wang, Dan Su, Dong Yu, Yi Ren, and Zhou Zhao. Fastdiff: A fast conditional diffusion model for high-quality speech synthesis. In Lud De Raedt, editor, *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 4157–4163. International Joint Conferences on Artificial Intelligence Organization, 7 2022. Main Track.

[Joshi *et al.*, 2021] Sonal Joshi, Jesús Villalba, Piotr Żelasko, Laureano Moro-Velázquez, and Najim Dehak. Study of pre-processing defenses against adversarial attacks on state-of-the-art speaker recognition systems. *IEEE Transactions on Information Forensics and Security*, 16:4811–4826, 2021.

[Jung *et al.*, 2022] Jee-weon Jung, Hee-Soo Heo, Hemlata Tak, Hye-jin Shim, Joon Son Chung, Bong-Jin Lee, Ha-Jin Yu, and Nicholas Evans. Aasist: Audio anti-spoofing using integrated spectro-temporal graph attention networks. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6367–6371. IEEE, 2022.

[Kassis and Hengartner, 2021] Andre Kassis and Urs Hengartner. Practical attacks on voice spoofing countermeasures. *arXiv preprint arXiv:2107.14642*, 2021.

[Kingma and Ba, 2014] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[Kos *et al.*, 2018] Jernej Kos, Ian Fischer, and Dawn Song. Adversarial examples for generative models. In *2018 ieee security and privacy workshops (spw)*, pages 36–42. IEEE, 2018.

[Lin and Abdulla, 2015] Yiqing Lin and Waleed H Abdulla. Principles of psychoacoustics. In *Audio Watermark*, pages 15–49. Springer, 2015.

[Lo *et al.*, 2019] Chen-Chou Lo, Szu-Wei Fu, Wen-Chin Huang, Xin Wang, Junichi Yamagishi, Yu Tsao, and Hsin-Min Wang. Mosnet: Deep learning based objective assessment for voice conversion. *arXiv preprint arXiv:1904.08352*, 2019.

[Mu and Li, 2022] Feiteng Mu and Wenjie Li. Enhancing text generation via multi-level knowledge aware reasoning. In Lud De Raedt, editor, *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 4310–4316. International Joint Conferences on Artificial Intelligence Organization, 7 2022. Main Track.

[Paszke *et al.*, 2019] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

[Perov *et al.*, 2020] Ivan Perov, Daiheng Gao, Nikolay Chervoniy, Kunlin Liu, Sugasa Marangonda, Chris Umé, Mr Dpfks, Carl Shift Facenheim, Luis RP, Jian Jiang, et al. Deepfacelab: Integrated, flexible and extensible face-swapping framework. *arXiv preprint arXiv:2005.05535*, 2020.

[Qin *et al.*, 2019] 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.

[Radford *et al.*, 2022] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. *arXiv preprint arXiv:2212.04356*, 2022.

[Ruiz *et al.*, 2020] Nataniel Ruiz, Sarah Adel Bargal, and Stan Sclaroff. Disrupting deepfakes: Adversarial attacks against conditional image translation networks and facial manipulation systems. In *European Conference on Computer Vision*, pages 236–251. Springer, 2020.

[Shu *et al.*, 2021] Kai Shu, Yichuan Li, Kaize Ding, and Huan Liu. Fact-enhanced synthetic news generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13825–13833, 2021.

[Sinha *et al.*, 2022] Sanjana Sinha, Sandika Biswas, Ravindra Yadav, and Brojeshwar Bhowmick. Emotion-controllable generalized talking face generation. In Lud De Raedt, editor, *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 1320–1327. International Joint Conferences on Artificial Intelligence Organization, 7 2022. Main Track.

[Van Der Walt *et al.*, 2011] Stefan Van Der Walt, S Chris Colbert, and Gael Varoquaux. The numpy array: a structure for efficient numerical computation. *Computing in science & engineering*, 13(2):22–30, 2011.

[Veaux *et al.*, 2017] Christophe Veaux, Junichi Yamagishi, Kirsten MacDonald, et al. Cstr vctk corpus: English multi-

speaker corpus for cstr voice cloning toolkit. *University of Edinburgh. The Centre for Speech Technology Research (CSTR)*, 2017.

[Wan *et al.*, 2018] Li Wan, Quan Wang, Alan Papir, and Ignacio Lopez Moreno. Generalized end-to-end loss for speaker verification. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4879–4883. IEEE, 2018.

[Wang and Yamagishi, 2021] Xin Wang and Junich Yamagishi. A comparative study on recent neural spoofing countermeasures for synthetic speech detection. *arXiv preprint arXiv:2103.11326*, 2021.

[Wang *et al.*, 2020] Qing Wang, Pengcheng Guo, and Lei Xie. Inaudible adversarial perturbations for targeted attack in speaker recognition. *arXiv preprint arXiv:2005.10637*, 2020.

[Wang *et al.*, 2021] Disong Wang, Liqun Deng, Yu Ting Yeung, Xiao Chen, Xunying Liu, and Helen Meng. Vqmivc: Vector quantization and mutual information-based unsupervised speech representation disentanglement for one-shot voice conversion. *arXiv preprint arXiv:2106.10132*, 2021.

[Wenger *et al.*, 2021] Emily Wenger, Max Bronckers, Christian Cianfarani, Jenna Cryan, Angela Sha, Haitao Zheng, and Ben Y Zhao. " hello, it's me": Deep learning-based speech synthesis attacks in the real world. In *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*, pages 235–251, 2021.

[Xie *et al.*, 2020] Cihang Xie, Mingxing Tan, Boqing Gong, Jiang Wang, Alan L Yuille, and Quoc V Le. Adversarial examples improve image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 819–828, 2020.

[Yamagishi *et al.*, 2021] Junichi Yamagishi, Xin Wang, Massimiliano Todisco, Md Sahidullah, Jose Patino, Andreas Nautsch, Xuechen Liu, Kong Aik Lee, Tomi Kinnunen, Nicholas Evans, et al. Asvspoof 2021: accelerating progress in spoofed and deepfake speech detection. *arXiv preprint arXiv:2109.00537*, 2021.

[Ye *et al.*, 2022] Zhenhui Ye, Zhou Zhao, Yi Ren, and Fei Wu. Syntaspeech: Syntax-aware generative adversarial text-to-speech. In Lud De Raedt, editor, *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 4468–4474. International Joint Conferences on Artificial Intelligence Organization, 7 2022. Main Track.

[Yi *et al.*, 2022] Jiangyan Yi, Ruibo Fu, Jianhua Tao, Shuai Nie, Haoxin Ma, Chenglong Wang, Tao Wang, Zhengkun Tian, Ye Bai, Cunhang Fan, et al. Add 2022: the first audio deep synthesis detection challenge. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 9216–9220. IEEE, 2022.

[Yuan *et al.*, 2019] Xiaoyong Yuan, Pan He, Qile Zhu, and Xiaolin Li. Adversarial examples: Attacks and defenses for deep learning. *IEEE transactions on neural networks and learning systems*, 30(9):2805–2824, 2019.