Computer Systems Science & Engineering
DOI: 10.32604/csse.2023.034568
Article

Tech Science Press

Adversarial Examples Protect Your Privacy on Speech Enhancement System

Mingyu Dong, Diqun Yan[*] and Rangding Wang

Department of Information Science and Engineering, Ningbo University, Zhejiang, 315000, China
*Corresponding Author: Diqun Yan. Email: yandiqun@nbu.edu.cn

Received: 20 July 2022; Accepted: 22 September 2022

Abstract: Speech is easily leaked imperceptibly. When people use their phones, the personal voice assistant is constantly listening and waiting to be activated. Private content in speech may be maliciously extracted through automatic speech recognition (ASR) technology by some applications on phone devices. To guarantee that the recognized speech content is accurate, speech enhancement technology is used to denoise the input speech. Speech enhancement technology has developed rapidly along with deep neural networks (DNNs), but adversarial examples can cause DNNs to fail. Considering that the vulnerability of DNN can be used to protect the privacy in speech. In this work, we propose an adversarial method to degrade speech enhancement systems, which can prevent the malicious extraction of private information in speech. Experimental results show that the generated enhanced adversarial examples can be removed most content of the target speech or replaced with target speech content by speech enhancement. The word error rate (WER) between the enhanced original example and enhanced adversarial example recognition result can reach 89.0%. WER of target attack between enhanced adversarial example and target example is low at 33.75%. The adversarial perturbation in the adversarial example can bring much more change than itself. The rate of difference between two enhanced examples and adversarial perturbation can reach more than 1.4430. Meanwhile, the transferability between different speech enhancement models is also investigated. The low transferability of the method can be used to ensure the content in the adversarial example is not damaged, the useful information can be extracted by the friendly ASR. This work can prevent the malicious extraction of speech.

Keywords: Adversarial example; speech enhancement; privacy protection; deep neural network

1 Introduction

Personal voice assistants are increasingly used as interfaces to digital environments [1]. Voice assistants of mobile phones are listening to the specific commands in speech all the time to be wakened up. The privacy in speech may be recorded by the mobile phone while the voice assistant is listening. So, when people talk about something frequently, in no time the application downloaded on the mobile phone will recommend it. The content information in speech is extracted by Automatic Speech Recognition (ASR) technology. To

This work is licensed under a Creative Commons Attribution 4.0 International License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

improve the recognition accuracy, the speech will be enhanced before recognizing [2]. Along with the development of the Deep Neural Network (DNN), technology of Natural Language Processing (NLP) has been progressing, and speech-related technologies are also getting more efficient progress [3,4].

The adversarial example is the deadly weakness of DNN models, which is a technology that can cause the DNN model to output a wrong result. And the generated adversarial example is imperceptible to human hearing [5]. Some classical methods of adversarial examples have developed since being proposed in 2013 [6]. Adversarial examples have always worked on the image domain in recent research, with the image classification task as the target model. As attention became more focused on the audio domain, research on audio adversarial examples grew. Kos et al. [7] proposed an adversarial example to attack the image reconstruction model, which is the first work related to the deep generation model. In the audio domain, Takahashi et al. proposed an adversarial example to attack an audio source separation model that can help protect songs' copyrights from the abuse of separated signals [8]. Huang et al. proposed work of attacks on voice conversion systems [9], which can protect a speaker's private information.

To protect privacy in the speech taped by mobile phones, using adversarial examples to attack the ASR to make the output of the recognition false may work [10]. There are some effective methods to generate a stable adversarial example on ASR [11,12]. But the generated adversarial example can be eliminated by some methods [13], which makes the ASR more robust. Moreover, the high-effective ASR may be equipped with one Speech Enhancement (SE) system to get rid of the environmental noise. Research shows that the denoising process can invalidate the adversarial example. So, the adversarial example against ASR may not work after being denoised by the SE [14]. To solve this problem, in this paper we propose an adversarial method to attack the SE directly, which can make the content of enhanced speech erased. As shown in Fig. 1, before the recognition, the proposed method attacks the speech enhancement. The attacked speech can be called protected speech. Then the protected speech is denoised by the SE, different from the normal example, the enhanced protected speech is lack of most useful content information. And in the mobile processor, the content information extracted from the enhanced protected speech is invalid to the application of the mobile phone. In this way, privacy in speech is protected. The code [1]is available.

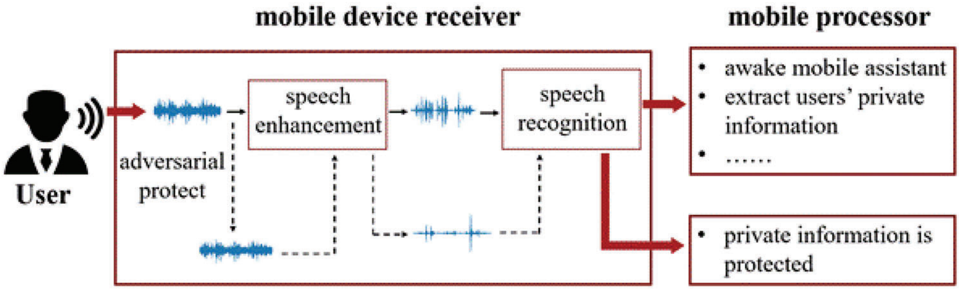


Figure 1: The process of privacy protection in mobile phone terminates

In this paper, the usage of the adversarial example is expanded to protect the privacy of content information in speech. This work designs the attack method to attack the speech enhancement system which is a generation model. As far as we know, there is very little related work at present. The contributions of this paper are as follows:

- In this paper, the adversarial example is creatively extended from ASR to SE, which solves the problem that the adversarial example on ASR is easily affected by SE. By attacking the speech enhancement system, the text content in the speech cannot be extracted maliciously.

[1] https://github.com/DanMerry/Advsarial_SE

- To simulate the real-world scenarios, the speech recognition of different complex scenes and the Signal Noise Ratio (SNR) of the speech has been comprehensively consider in this paper. According to the performance of the adversarial example before and after the attack, the corresponding evaluation metrics are proposed to measure the effect of the adversarial example.
- The privacy in the speech can be protected by the proposed method. The content information in enhanced speech will be erased through SE by the proposed method. Thus, the privacy information of the speaker cannot be extracted.

The rest of this paper is organized as follows. Section II introduces some work related to the proposed method, which is described in Section 3. Section 4 discusses experiments. Section 5 provides the conclusions.

2 Relate Work

Notations used in this paper are summarized in Table 1.

Table 1: Notations and corresponding explanations

Notations	Explanations
x	Input example
t_{adv}	Target of the adversarial example
x^*	Adversarial example
ε	Step of updated adversarial perturbation
∇_x	Gradient of the input
s	Range of adversarial perturbation
r	Adversarial perturbation
k	Confidence of the output
x_{noisy}	Noisy original input example
y_{en}	Enhanced noisy original example
x^*_{noisy}	Adversarial example of the noisy input
y^*_{noisy}	Enhanced adversarial example
gt	Ground truth content of the original example
$loss_{value}$	Value of the loss function

2.1 Gradient-based Adversarial Method

The gradient-based adversarial attack methods need to know the parameters and structure of target victim models. The adversarial perturbation is generated through the gradient calculated from the loss function between target and output. The original example added with the adversarial example can confuse the victim model. The advantage of the gradient-based method is that the adversarial examples can be generated in a short time.

The classic fast gradient sign method (FGSM) [15] is a shortcut algorithm that simply adds a signed gradient to the original example:

$$x^* = x + \varepsilon \cdot sign(\nabla_x \mathcal{L}(f(x),\ t_{adv})) \tag{1}$$

where $\varepsilon f()$ is the target victim model, and $\mathcal{L}(\cdot)$ is the loss function between the predicted result and the target t_{adv}. In the traditional. In the traditional classification task, the output of $f()$ is the classification probability of all categories, and the choice of the loss function is cross-entropy loss. FGSM is a one-step and simple method. But the attack effect of the FGSM is not that satisfactory.

Projected Gradient Descent (PGD) [16] is an iterative method to generate a more effective adversarial example based on FGSM. Compared to the one-step attack, the generated perturbation by PGD is smaller in every step, in which the value of the perturbation will be small and limited to a specific range:

$$x_{t+1}^{*} = \prod_{x+s}\left(x_t^{*} + \varepsilon \cdot sign(\nabla_x \mathcal{L}(f(x),\ t_{adv}))\right) \tag{2}$$

PGD has the advantage that the generated adversarial examples have a higher attack success rate, but the adversarial perturbation in many steps will be magnified, so the attacked example will be greatly changed.

2.2 Optimization-based Adversarial Method

The Carlini and Weanger (C&W) method [17] is a classic optimization-based method. Compared with the gradient-based methods mentioned above, the optimization-based methods have a high attack success rate, meanwhile, the generated adversarial perturbations are more imperceptible. The whole process can be described as follows:

$$r = \frac{1}{2}(tanh(\omega_n) + 1) - x \tag{3}$$

$$\min_{\omega_n} \|r\|_p + c \cdot f(tanh(\omega) + 1)$$

$$where \quad f(x^{*}) = max\left(max\left(Z\left(x_{i:i\neq t}^{*}\right) - Z\left(x_t^{*}\right),\ -k\right)\right) \tag{4}$$

where the ω_n is the generated adversarial in every step of optimization. The optimized object is the original example, several optimization targets will transform the original example into the adversarial example. In the image classification task, the range of every pixel value is [0, 255] in the Red Green Blue (RGB) channel. To optimize the input samples efficiently, the input is transformed to the tanh domain. To ensure the attack success rate of generated adversarial example, the output of the confidence $Z()$ is close to the target category. To ensure the imperceptibility of the adversarial perturbation, the r is required as small as possible, which can be measured by different norms. These two loss functions will guide the optimization direction.

The generated adversarial examples by optimization-based methods are high-quality, but the optimization process requires time. In the audio domain, the optimized objects are long queue audio points, the optimization process will be longer.

2.3 Speech Enhancement System

Speech enhancement systems are the target of adversarial examples, which can transform noisy speech into clear speech [18]. Recently, DNN-based speech enhancement technology has developed quickly. Speech enhancement is an end-to-end task, so the GAN models in DNNs are suitable for this task. So, in the experiments, the advanced Metric Generative Adversarial Net Plus (MetricGAN+) [19] and classical Speech Enhancement Generative Adversarial Net (SEGAN) [20] are chosen as the attack targets. MetricGAN+ is based on MetricGAN [21], and its Perceptual Evaluation of Speech Quality (PESQ) can reach 3.15, which is a state-of-the-art result. SEGAN[2] applies U-Net in speech enhancement, and its PESQ can reach 2.16.

[2] https://github.com/santi_pdp/segan_pytorch

3 Proposed Method

In this section, we expand the adversarial attack method from classification task to generation task. The output of the classification is the probability array of all categories, but the output of the generation task in audio is the long sequence audio sampling value whose length is as same as the input. What effect the adversarial example can do on the SE system depends on the choice of the target, different targets can exhibit different effects. To make the enhanced speech miss important content information, the target can be a clip of the blank speech. In this way, the enhanced speech sounds like one mute clip. To make the enhanced be replaced as one given speech, the target can be one specific speech that may contain some warning. In this way, the content information of enhanced speech is replaced. The noise clip can also be set as the target which will make the enhanced speech sound chaotic. To highlight the effect of an adversarial attack, a silent audio clip is chosen as a target in the following experiments. In speech enhancement models, the enhancement process can be described as:

$$y_{en} = g\left(x_{noisy}\right) \tag{5}$$

where $g()$ is the speech enhancement process that can transform noisy speech to clear speech, x_{noisy} is a noisy example, and y_{en} is an enhanced example.

There is the issue of how to make gradient-based methods work on a speech enhancement model. The proposed adversarial attack methods are based on the above gradient-based methods of FGSM and PGD. The existing methods are designed to attack the classification task, hence the cross-entropy loss function in the original methods is not suitable for the generation task. Mean Squared Error (MSE) or p norm is used as the loss function. The loss function between the predicted output and target results in the one-step FGSM and iterative PGD are replaced by the MSE. MSE loss will better enable the generated adversarial example to achieve the goal. The usage of the MSE loss function in FGSM and PGD can be described as follows:

$$MSE(m,\ n) = \frac{1}{n}\sum_{i=1}^{n}(m_i - n_i)^2 \tag{6}$$

$$x^* = x + \varepsilon \cdot sign(\nabla_x MSE(f(x),\ t_{adv})) \tag{7}$$

$$x^*_{t+1} = \prod_{x+s}\left(x^*_t + \varepsilon \cdot sign(\nabla_x MSE(f(x),\ t_{adv}))\right) \tag{8}$$

In the whole SE model, data is floating in the format of a tensor. According to the loss function, the gradient can be computed through the output of the model and the target. And the loss function guides the direction of the gradient, and the MSE in the loss function guarantees the output of the adversarial example becomes the target. The signed gradient on the first layer is the adversarial perturbation. The adversarial example is the original example added with the adversarial perturbation. Put the generated adversarial example to SE again, the output of the SE model has less loss value computed with the target. In this way, the adversarial example degrades the SE model, and the enhanced adversarial example is close to the target.

Algorithm 1: Optimization-based method

```
Input: Original example $x_{noisy}$, target $t_{adv}$, iterations $itr$
Output: Adversarial example $x^*_{noisy}$
    $loss_{value} \leftarrow 0$
    $x^*_{noisy} \leftarrow x_{noisy}$
```

(Continued)

Algorithm 1 (continued)

```
for $itr$ step do
    $\mathcal{L}_1 \leftarrow MSE\left(x^*_{noisy},\ x^*_{noisy}\right)$
    $\mathcal{L}_2 \leftarrow \left\| g\left(x^*_{noisy}\right) - t_{adv} \right\|_2$
    $\mathcal{L}_3 \leftarrow \mathcal{L}_1 + \alpha \cdot \mathcal{L}_2$
    Update $x^*_{noisy}$ by Adam optimization
    if $loss_{value} > \mathcal{L}_3$ then
        $loss_{value} \leftarrow \mathcal{L}_3$
    else
        $Break$
    end if
end for
return $x^*_{noisy}$
```

How to make the optimization-based (OPT) method used for the speech enhancement model needs to be considered. C&W is an effective attack method. To ensure the generated adversarial example is close to the original example, the MSE loss is used to describe it. To ensure the enhanced adversarial example is close to the target, the norm is used to describe the difference between output and target. The proposed method can be seen in Algorithm 1. In the whole iterative process, these two loss functions will guide the generated adversarial examples to be closer to the demand.

4 Experiments

4.1 Experimental Setup

Voicebank [22] and Demand [23] are chosen as the datasets of clean speech and background noise, respectively. Voicebank contains more than 500 h of recordings from about 500 healthy speakers, and Demand contains 6 indoor noise audios and 12 outdoor noise audios. The open-source pre-trained model in SpeechBrain [24] is used as the speech recognition system. The experimental results are conduct in different SNR values which include −8, −4, 0, 4, and 8. The smaller the SNR, the noisier the speech. And the background noises are in five different noisy scenes which include a busy subway station, a public town square, public transit bus, a private passenger vehicle, and a subway. They are all familiar situations for us that people may have conversations with others. In OPT method, α is det as 0.0001 which makes two loss functions in the same weight.

4.2 Evaluation Metrics

The measurement of the adversarial example on the generation task is different from the classification task. In the classification task, the attack success rate is the main evaluation metric. In the generation task, evaluation could be designed according to the demand of the effect. There are two factors are considered: (a) How much does the adversarial perturbation affect the speech enhancement? (b) How bad does the enhanced adversarial speech sound? This paper uses the Residual Perturbation Rate RPR and the Degree of Enhancement DE to measure the effect of adversarial examples. the x_{noisy}, y_{en}, x^*_{noisy}, and y^*_{en} are defined as the input speech, enhanced original speech, adversarial example, and enhanced adversarial example. And

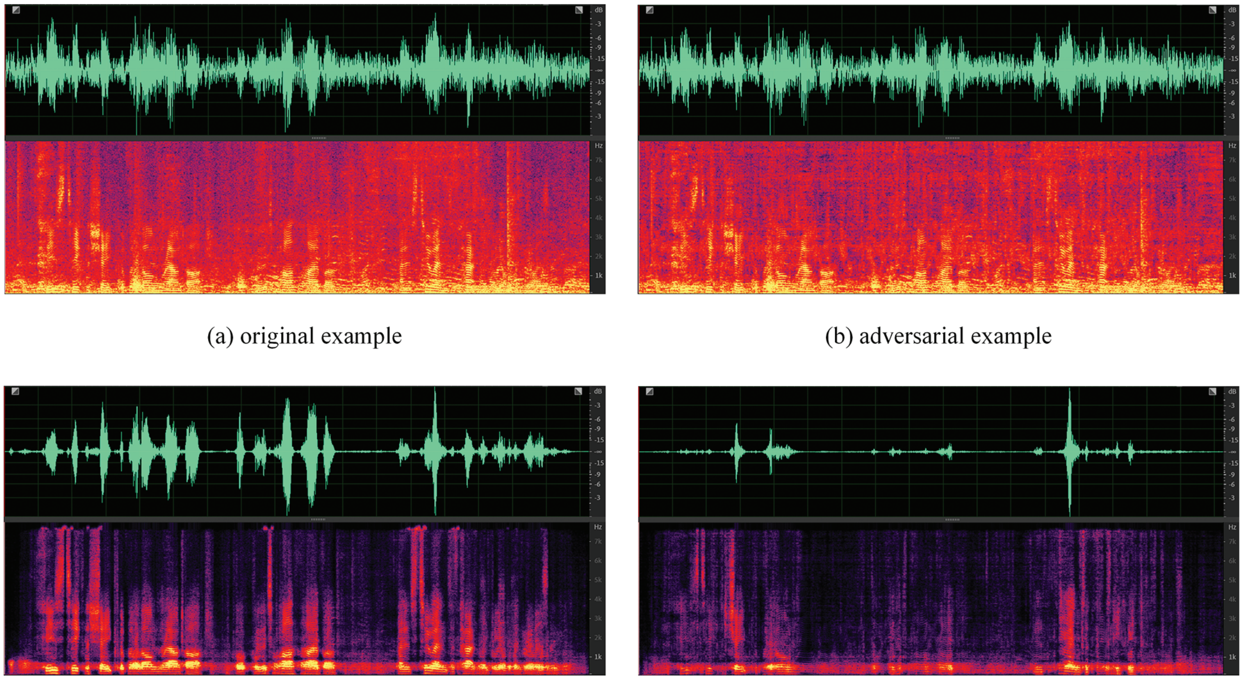
$$RPR = \frac{ln\|y^*_{en} - y_{en}\|_2}{ln\|x^*_{en} - x_{en}\|_2} \tag{9}$$

$$DE = F\left(R\left(y^*_{en}\right),\ gt\right) - F(R(y_{en}),\ gt) \tag{10}$$

where $R()$ is the speech recognition model and $F()$ is the Word Error Rate (WER) of the speech recognition result, and gt is the ground-truth content of the original example. DE measures how much the enhanced adversarial example has changed from the original enhanced example. If the adversarial example works, the recognition system will not output a correct result, so there will be a difference between $F\left(R\left(y^*_{en}\right),\ gt\right)$ and $F(R(y_{en}),\ gt)$. RPR indicates how much the adversarial perturbation affects the enhancement system. In RPR, the adversarial perturbation is the denominator, and the difference between the enhanced original example and the enhanced adversarial example is the numerator. So, if RPR is greater than one, then the higher the RPR, the better the degradation effect on the speech enhancement system, meaning that the method works. If the RPR is less than one, it means that the perturbation has not changed the SE much.

4.3 Results

The effect of the adversarial example protecting speech privacy can be seen in Fig. 2. The adversarial perturbation in the adversarial example is almost imperceptible, but the changes it brings are much more than that. As you can see the speech waveform in Fig. 2 that the waveform peak of the enhanced adversarial example is erased. The peak in the waveform represents the speech content. So, the adversarial example makes much content information in the enhanced adversarial example be erased.



(a) original example

(b) adversarial example

(c) enhanced original example

(d) enhanced adversarial example

Figure 2: Effect of adversarial attack works on speech enhancement model

4.3.1 Privacy Protection Effect

Fig. 3 shows the results of *WER* on speech enhancement processing with different attack methods. In Fig. 3, *WER* is the average value of different examples. The first three lines in legend are the *WER* of original noisy examples and enhanced examples by two different SE. *WER* of original speech decreases with the decrease of *SNR*. The difference between the noisy example and the enhanced example is obvious. So it is necessary to enhance the input example, which can help improve the robustness of speech recognition.

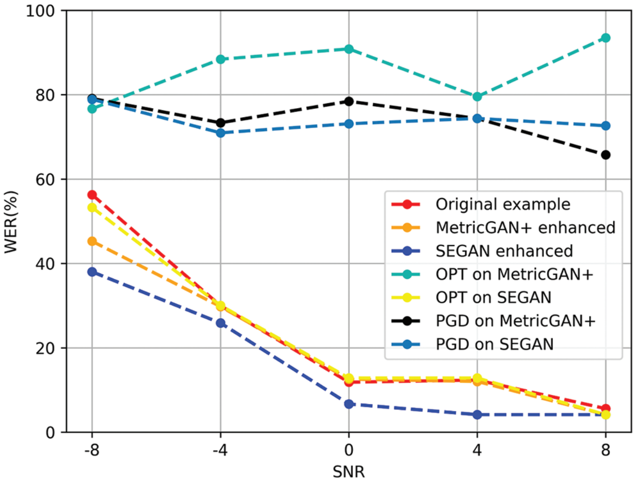


Figure 3: *WER* results of adversarial examples and original example in different *SNR* examples

The *WER* of the enhanced generated adversarial examples is much higher than the enhanced original examples. And no matter the value of *SNR*, the *WER* of enhanced adversarial examples maintains a stable high level. The results show that adversarial examples totally invalidate the speech enhancement system, and the ASR cannot extract the correct information. The attack effect of the FGSM is relatively weak, the result is not shown in the Figure. Both PGD and CW perform well in MeticGAN+. However, the CW does not work in SEGAN. Because the SEGAN is a waveform-to-waveform model, the loss of OPT on the SEGAN converges quickly. The optimization process ends up early, which makes the effect poor.

The performance of degrading on the target SE model is shown in Fig. 4. The choice of the target model is the MetricGAN+ in this experiment and the results are measured by *RPR* and *DE*. The results in Fig. 4 can be concluded that *RPR* on OPT is higher than FGSM, which means the OPT method can generate a greater change than FGSM. And as *SNR* increases, *RPR* increases. However, *DE* decreases as *RPR* increases. The enhanced original examples have a high *WER*, the ASR can recognize more information under the high *SNR*. So, the difference between the *WER* values becomes small. The results show that under different *SNR*s, the OPT method can produce more degradation. Due to the iterative process, OPT can fine-tune a more specific direction to get close to the target in every single step through the loss function. However, FGSM is a one-step method, so the adversarial perturbation may have a rough direction. Thus, the enhanced FGSM adversarial example cannot have the same good effect as OPT. According to the *RPR* and *DE*, the results show that the proposed method can prevent the ASR from extracting the privacy of the speaker by attacking the SE.

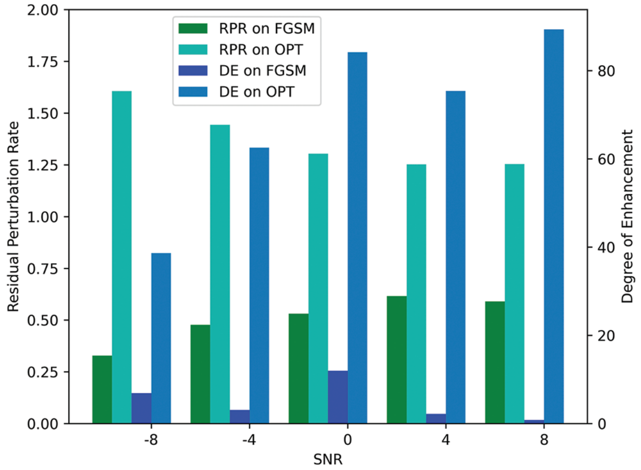


Figure 4: Performance of adversarial examples on proposed evaluation metrics

4.3.2 Target Attack on Speech Enhancement

In this section, the target attack is conduct the target attack on the speech enhancement model. The recognition result of the adversarial example will be replaced with the target content. When the speech is taped by other devices, privacy should not be extracted by the malicious ASR. The proposed method can even replace the speech content with some warnings to remind the speech owner that they are not allowed to use this speech. In the experiment, MetricGAN+ is choose as target model. To achieve the effect of the target attack, the choice of the target is one clean speech with high *SNR* whose speaker identity is different from that of the original example. The chosen attack method is OPT.

Table 2 is the recognition result of the proposed target attack of one specific example. Comparing the ground truth and the recognition results of the enhanced examples, the result shows that ASR has very little *WER*. And the recognition result of the enhanced adversarial recognition (enhanced adv. reco.) is similar to the target example, which means that the target attack can completely replace the victim's speech content information. The target attack in all given *SNR*s is also considered, the *WER* between ground truth of target and recognition results of the enhanced adversarial examples can be low to 33.75%. Consequently, the target attack is practical in the proposed method. It can replace the content information in the speech.

Table 2: Recognition results of target attack

Item	Recognition results
Original ground truth	he claimed his insurance company contested the damages not the restaurant
Target ground truth	we have to look at everything before we make any final decision
Enhanced original reco.	he claimed his insurance company and tested the damages not the restaurant
Enhanced target reco.	we have to look at everything before we make any final decision
Enhanced adv. reco.	we have to look at everything before we make any final decision

4.3.3 Transferability Between Different Speech Enhancement Models

In this section, this work considers the transferability of the adversarial example between different SE models. The expected result is that the adversarial example should not destroy the content information of the speech, which means that the adversarial example should only attack the target SE model. The content information may be useful for some purposes in some cases, so the content information should be extracted as much as possible by other friendly SE models. Speech is leaked easily, and stopping the behavior of malicious extracting from the leaked speech is unrealistic. Thus, how to protect the leaked personal information from the original source is executable. The transferability of adversarial examples is a significant factor that should be considered. In this experiment, adversarial examples are generated on the source model and the PGD is chosen as the attack method. The generated adversarial examples are tested in the target model and another model. The MetricGAN+ and SEGAN are chosen to be the target model and another model, respectively.

Table 3 is the results of the transferability of the adversarial example on different SE models. According to the *RPR* of the adversarial examples generated from MetricGAN+, the value of *RPR* where the adversarial example works on itself is much larger than it works on the SEGAN. The *RPR* on itself is 1.2192 which is greater than it on SEGAN 0.8575, which declares that adversarial examples generated from MetricGan+ can degrade itself more than other models. And the *DE* also reduces when the adversarial example works on the SEGAN. From the table, the *DE* of adversarial examples work on itself is 58.4087 is far greater than it works on SEGAN 19.5005, which shows that the adversarial examples generated from MetricGan+ can erase more content information than on another model. It can be concluded that the transferability of the adversarial example generated by MetricGAN+ is weak. According to the results of the adversarial example generated by SEGAN works on different models, the *RPR* also reduces, but the *DE* gets a little higher. Because the SEGAN is the waveform-to-waveform system, the perturbation is added to the wave directly. The MetricGAN+ can denoise more than SEGAN, so the difference between the adversarial example and the enhanced example denoised by MetricGAN+ is larger than that work in the SEGAN itself. From the results, the effect of the proposed adversarial example is better working on the target model than other models. Weak transferability can be used to protect the speech from being destroyed, other models are allowed to extract the information in speech, and the target model cannot use this speech.

Table 3: Transferability results of adversarial examples between different models

Source	Target	*DE*	*RPR*
MetricGAN+	MetricGAN+	58.4087	1.2192
MetricGAN+	SEGAN	19.5005	0.8575
SEGAN	SEGAN	53.2389	1.0758
SEGAN	MetricGAN+	41.8050	1.3470

5 Conclusion

In this paper, we proposed an adversarial attack method that works on a speech enhancement system to prevent the malicious extraction of leaked speech without destroying the original examples. The adversarial example can invalidate the speech enhancement system, which can prevent privacy in speech from being maliciously extracted. The enhanced speech will be transformed into the preset target speech whose content information cannot be extracted correctly. Experimental results show that the enhanced adversarial examples can result in *WER* of the recognition results greater than 89%, and the adversarial perturbation in the adversarial example is almost imperceptible. The target attack can make one enhanced adversarial example to be recognized as the target phase, the average *WER* of the target phase can be low to 33.75%. We also considered the transferability between different speech enhancement models. This work expands the usage of the adversarial example to protect privacy in speech.

In future work, we intend to improve the transferability of the adversarial example to more models. In the meantime, the friendly model should be protected. The mainstream speech enhancement models are unknown to us, the adversarial example should attack the applied wildly models successfully which is a total black-box situation. To accomplish the requirements of transferability, one protected speech enhancement model and several target models will be set in the future experiments.

Funding Statement: This work was supported by the National Natural Science Foundation of China (Grant No. 61300055), Zhejiang Natural Science Foundation (Grant No. LY20F020010), Ningbo Science and Technology Innovation Project (Grant No. 2022Z075), Ningbo Natural Science Foundation (Grant No. 202003N4089) and K.C. Wong Magna Fund in Ningbo University.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] P. Cheng and U. Roedig, "Personal voice assistant security and privacy—a survey," in *Proc. of the IEEE*, vol. 110, no. 4, pp. 476–507, 2022.

[2] F. Weninger, H. Erdogan, S. Watanabe, E. Vincent, J. Le Roux *et al.,* "Speech enhancement with LSTM recurrent neural networks and its application to noise-robust ASR," in *Int. Conf. on Latent Variable Analysis and Signal Separation of the Springer*, Liberec, Czech Republic, pp. 91–99, 2015.

[3] Y. Jia, Y. Zhang, R. Weiss, Q. Wang and J. Shen *et al.,* "Transfer learning from speaker verification to multispeaker text-to-speech synthesis," *Advances in Neural Information Processing Systems*, vol. 31, pp. 4485–4495, 2018.

[4] A. B. Nassif, I. Shahin, I. Attili, M. Azzeh and K. Shaalan, "Speech recognition using deep neural networks: A systematic review," *Access of the IEEE*, vol. 7, no. 7, pp. 19143–19165, 2019.

[5] D. Wang, R. Wang, L. Dong, D. Yan and X. Zhang *et al.,* "Adversarial examples attack and countermeasure for speech recognition system: A survey," in *Int. Conf. on Security and Privacy in Digital Economy of the Springer*, Singapore, pp. 443–468, 2020.

[6] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna and D. Erhan *et al.,* "Intriguing properties of neural networks," in *Int. Conf. on Learning Representations*, Banff, AB, Canada, 2014.

[7] J. Kos, I. Fischer and D. Song, "Adversarial examples for generative models," in *Security and Privacy Workshops of the IEEE*, San Francisco, California, USA, pp. 36–42, 2018.

[8] N. Takahashi, S. Inoue and Y. Mitsufuji, "Adversarial attacks on audio source separation," in *Int. Conf. on Acoustics, Speech, and Signal Processing of the IEEE*, Toronto, Ontario, Canada, pp. 521–525, 2021.

[9] C. Y. Huang, Y. Y. Lin, H. Y. Lee and L. S. Lee, "Defending your voice: Adversarial attack on voice conversion," in *Spoken Language Technology Workshop of the IEEE*, Shenzhen, China, pp. 552–559, 2021.

[10] Y. Qin, N. Carlini, G. Cottrell, I. Goodfellow and C. Raffel, "Imperceptible, robust, and targeted adversarial examples for automatic speech recognition," in *Int. Conf. on Machine Learning*, PMLR, Vancouver, BC, Canada, pp. 5231–5240, 2019.

[11] N. Carlini and D. Wagner, "Audio adversarial examples: Targeted attacks on speech-to-text," in *Security and Privacy Workshops of the IEEE*, San Francisco, USA, pp. 1–7, 2018.

[12] H. Kwon, Y. Kim, H. Yoon and D. Choi, "Selective audio adversarial example in evasion attack on speech recognition system," *Transactions on Information Forensics and Security of the IEEE*, vol. 15, pp. 526–538, 2019.

[13] K. Tamura, O. Akitada, and H. Shuichi, "Novel defense method against audio adversarial example for speech-to-text transcription neural networks," in *Int. Workshop on Computational Intelligence and Applications of the IEEE*, Hiroshima, Japan, 2019.

[14] S. Joshi, S. Kataria, Y. Shao, P. Zelasko and J. Villalba *et al.,* "Defense against adversarial attacks on hybrid speech recognition using joint adversarial fine-tuning with denoiser," arXiv preprint, arXiv: 2204, 03851, 2022.

[15] J. Goodfellow, J. Shlens and C. Szegedy, "Explaining and harnessing adversarial examples," arXiv preprint, arXiv: 1412.6572, 2014.

[16] A. Madry, A. Makelov, L. Schmidt, D. Tsipras and A. Vladu, “Towards deep learning models resistant to adversarial attacks,” arXiv preprint, arXiv: 1706.06083, 2017.

[17] N. Carlini and D. Wagner, “Towards evaluating the robustness of neural networks,” in *2017 Symp. on Security and Privacy of the IEEE*, California, CA, USA, pp. 39–57, 2017.

[18] N. Das, S. Chakraborty, J. Chaki, N. Padhy and N. Dey, “Fundamentals, present and future perspectives of speech enhancement,” in *Int. Journal of Speech Technology*, vol. 24, no. 4, pp. 883–901, 2021.

[19] S. W. Fu, C. Yu, T. A. Hsieh, P. Plantinga and M. Ravanelli *et al.,* “Metricgan+: An improved version of metricgan for speech enhancement,” arXiv preprint, arXiv: 2104.03538, 2021.

[20] S. Pascual, A. Bonafonte and J. Serra, “SEGAN: Speech enhancement generative adversarial network,” in *Annual Conf. of the Int. Speech Communication Association*, Stockholm, Sweden, 2017.

[21] S. W. Fu, C. F. Liao, Y. Tsao and S. D. Lin, “Metricgan: Generative adversarial networks based black-box metric scores optimization for speech enhancement,” in *Int. Conf. on Machine Learning PMLR*, California, CA, USA, pp. 2031–2041, 2019.

[22] C. Veaux and J. Yamagishi and S. King, “The voice bank corpus: Design, collection and data analysis of a large regional accent speech database,” in *2013 Int. Conf. Oriental Held Jointly with 2013 Conf. on Asian Spoken Language Research and Evaluation (O-COCOSDA/CASLRE)*, Gurgaon, India, IEEE, pp. 1–4, 2013.

[23] J. Thiemann, N. Ito and E. Vincent, “DEMAND: A collection of multi-channel recordings of acoustic noise in diverse environments,” in *Proc. of Meetings on Acoustics*, Montreal, Canada, pp. 1–6, 2013.

[24] M. Ravanelli, T. Parcollet, P. Plantinga, A. Rouhe and S. Cornell *et al.,* “SpeechBrain: A general-purpose speech toolkit,” arXiv preprint, arXiv: 2106.04624, 2021.