Protect User' Privacy from Voice Separation via Time-Delay-Robust Adversarial Attack

Anonymous CVPR 2022 submission

Paper ID 0000

Abstract

Deep voice separation networks have achieved amazing performance on separating certain human voices from noise. However, these methods may lead to personal privacy problems, like employing eavesdropping on the conversation. In this paper, we proposed an adversarial-attack-based method to protect our privacy. By adding a tiny perturbation on mixed audio, we can significantly reduce the performance of voice separation models. With less than five percent of perturbation, we can reduce the quality of voice separation models from 2.34 SDR to 0.61 SDR in a black-box scenario. Even with 0.5-second time-delay on attacks, the quality of voice separation is still less than 1.57 SDR, which means our method is more applicable in the real world.

1. Introduction

Along with the tremendous success of neural networks in voice separation and the popularity of smart speakers [4, 1, 2, 11], the risks of this technology being used to eavesdrop on users' privacy cannot be ignored.

In this work, we focus on preventing user speech from being separated from audio. We expect to achieve that by adding a very small disturbance to the audio, which will hinder the voice separation model, and the disturbance will not interfere with normal communication. Formally, we consider the following problem: Mixing the speech of a particular user with that of someone else and then adding a smile disturbance to the mixed audio, preventing that particular user's speech from being singled out. The current leading methodology is based on predicting a mask for the spectrogram, thus we directly adding our perturbation to the spectrogram of the audio.

Inspired by the amazing success of adversarial attacks in visual models [3, 6, 7], we generate the perturbation on spectrogram using PGD [6] as well. However, different from the image adversarial attack, in reality, adding adversarial attacks to audio may have the problem of the time delay, so we get a compatible audio adversarial attack by optimizing all possible time delays simultaneously.

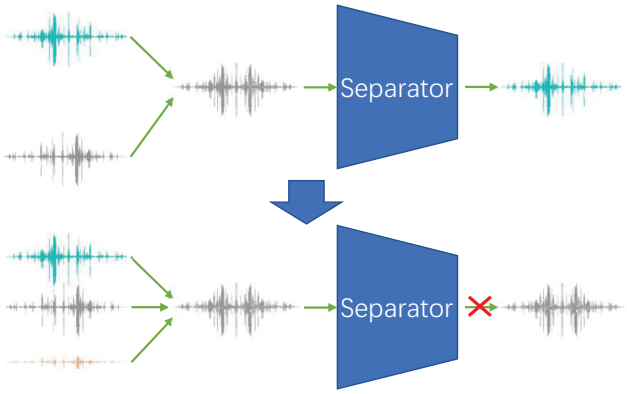


Figure 1. By adding a smile disturbance to the mixed audio, we can prevent users' privacy from being eavesdropped by the separation model.

In summary, this paper makes following contributions:

- Firstly, we are the first to propose and solve the privacy problem by voice separation models in white-box, gray-box, and black-box settings.
- Secondly, we provide a way to solve the time delay problem of deploying the perturbation in the real world.

2. Backgrounds and Related Work

2.1. Problem Setting

Consider a neural network $f : \mathbf{R}^T \to \mathbf{R}^T$, and an input data distribution D with multivariate random audio variables $x, y \in \mathbf{R}^n$ sampled from D. Voice separation problem wants to separate audio y from $x+y$ with some external information α about x. Formally, the problem need to maximize a metric M which measure the performance of the

neural network, that is,

$$\max_{f}(E_{x,y\in D}[M(x, f(x+y;\alpha))]).$$

Our adversarial attack method wants to add a small perturbation $\delta(x+y)$ on input $x+y$, to minimize the performance of voice separation neural networks, that is

$$\min_{|\delta|_\infty<\epsilon} \max_{f}(E_{x,y\in D}[M(x, f(x+y;\alpha))]).$$

2.2. Related Voice Separation Methods

In recent years, deep learning models were used in voice separation task [10], and they show their benefits comparing with classical methods. The first deep-learning-based voice separation model was proposed in 2012 [12, 13, 14], it performs better in feature extracting. A deep autoencoder [5] was proposed in the next year, and it was the first mapping-based method. VoiceFilter [11] uses some reference signal to separate a certain voice from audio. In 2019, a deep audio prior [9] was proposed, and no training data is needed during the separation.

2.3. Related Adversarial-Attack Methods

In computer vision problems, there are lots of adversarial-attack methods, many of which can reduce the capability of existing machine-learning-based models. It was pointed out that adversarial examples exist in deep neural networks [8]. Then in the next year, an easy adversarial-attack algorithm FGSM [3] was proposed, it uses the sign of gradient to maximize the effect of small perturbation. This algorithm was further developed to become PGD [6], which takes several steps of FGSM to get better performance on attacking.

3. Methods

In this section, we'll discuss the pipeline of generating audio perturbations.

Let $x \in R^{C\times T}$ be the spectrogram of target user audio, $y \in R^{C\times T}$ be the spectrogram of mixed audio of other users, and α be the selection information of target user. A voice separation model $f(x+y;\alpha) \to x'$ is a function that gives an approximation x' of x from mixed audio $x+y$. The adversarial perturbation $\delta \in R^{C\times T}$ is a perturbation that $|\delta|_\infty < \epsilon$ and maximize the difference between $f(x+y+\delta;\alpha)$ and x.

White-box Attack Typically voice separation models predict a mask for the target person, thus we try to break down the model by making the predicted mask null. Those voice separation methods that do not use the mask are not included in this paper.

Formally our optimization target is

$$\arg\max_{|\delta|_\infty<\epsilon} |f(x+y+\delta;\alpha)|_2^2,$$

where an effective δ can be found by using PGD [6].

Gray-box Attack When we do not know the target user information $\alpha \in R^d$, we use a random vector $\alpha' \in R^d$ that each dimension is independently drawn from a normal distribution with mean μ and variance σ, where μ and σ are the statistic mean and variance of α in training data.

In this case, the optimization target is

$$\arg\max_{|\delta|_\infty<\epsilon} |f(x+y+\delta;\alpha')|_2^2.$$

Black-box Attack For the case of a black-box attack, we use another trained voice separation network f' as the target network, and which the attack could generalize to other networks.

Thus, the optimization target in this case is

$$\arg\max_{|\delta|_\infty<\epsilon} |f'(x+y+\delta;\alpha')|_2^2.$$

Latency-aware black-box attack In the real world, because the delay of sound propagation is decided by physical distance, we cannot guarantee that the adversarial audio will be accurately recorded by the recording equipment. Thus, we assume that the audio will be delayed by time $d \sim U_{[0,d_{max}]}$.

In this case, we optimize the δ with

$$\arg\max_{|\delta|_\infty<\epsilon} E_d\left[|f'(x+y+\delta^{(d)};\alpha')|_2^2\right],$$

where $\delta^{(d)}$ refer to the audio δ delayed by d and padded with 0.

Time-Invariant black-box attack

We further consider the time-invariant black-box attack, which is more easily to be realized in real world. In this case, $\forall_j \delta_{i,j} = \delta_{i,0}$. This δ can be found with the same method as in the black-box scenario.

4. Results and Discussions

Method	SDR
Baseline	2.34
White-Box	-3.75
Gray-Box	-2.67
Black-Box	0.61
Time-Shifting	1.57

Table 1. This table shows SDR of baseline and SDR after all kinds of attacks

In this section, we will show our result on attacking voice separation models. We will introduce the performance of

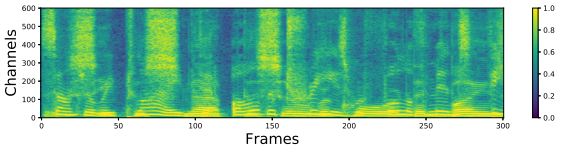
our adversarial attack, including white-box attack, gray-box attack, and black-box attack. We will also show our result on time-shifting attack. We will further do ablation study on the target of our PGD algorithm attack.

4.1. Experiment Settings

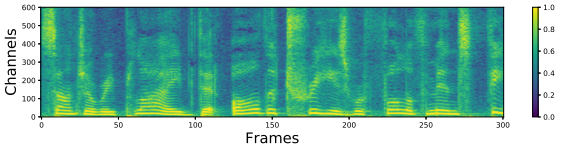
All experiments in this section are using VoiceFilter as our target of attack. The dataset is LibriSpeech, and *train-clean-100* is training set, *dev-clean* is test set. The attack is based on the model trained for 17000 iterations.

4.2. Baseline

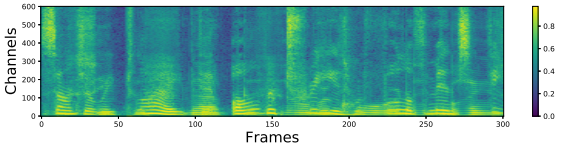
The baseline shows the separation result of VoiceFilter. As shown in Fig 2, after separation, VoiceFilter outputs audio that has a spectrogram almost the same as the target spectrogram. The SDR of the baseline is 2.34 as shown in table 1



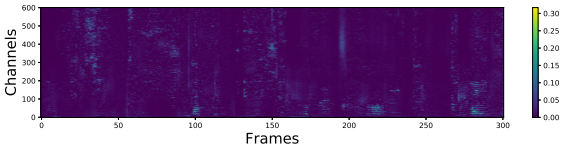
(a) Mixed audio spectrogram



(b) Target audio spectrogram



(c) Voice separation model output

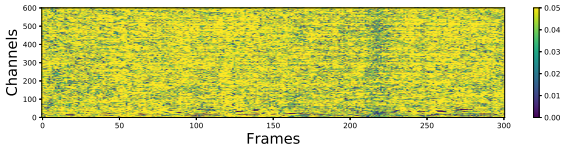


(d) Output difference between (b) and (c)

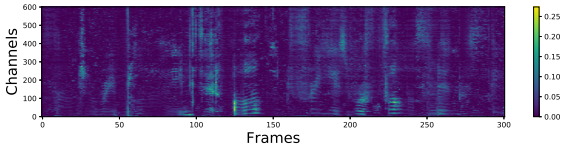
Figure 2. In each figure, the x-coordinate represents for time frame, each frame represents 10ms, and the y-coordinate represents for frequency channel. Fig 2(a) shows the spectrogram of origin mixed audio. Fig 2(b) shows the spectrogram of targeted audio, that is, the ground-truth. Fig 2(c) shows the spectrogram of audio outputted by VoiceFilter, which is quite similar to the target one. Fig 2(d) shows the error between target and output.

4.3. White, Gray and Black Box Attack

In our adversarial attack methods, we step by step remove the input to form a black-box attack. As shown in table 1, the white-box attack destroy the performance and reduce SDR to a negative value of -3.75. After removing the input voice feature, the attack can still reduce SDR to -2.67. In the final black-box attack method, SDR is higher than the previous two attacks, but it's also a successful attack. Fig 3 shows the result of black-box attack, with a small perturbation show in Fig 3(a), the output differs a lot.



(a) Perturbation by black-box attack



(b) Output difference with this attack

Figure 3. In each figure, the x-coordinate represents for time frame, each frame represents 10ms, and the y-coordinate represents for frequency channel. This is a result of the black-box attack. Fig 3 shows the perturbation made by the attack, and Fig 3(b) shows the error between the output and the target.

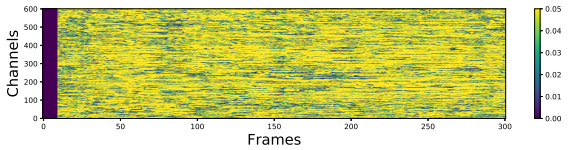
4.4. Time-Shifting Attack

In the time-shifting attack, the model uses signals that are 10 frames earlier to attack the present signal. This attack can decrease SDR from 2.34 to 1.57, which verifies the effectiveness of this attack. As shown in Fig 4, time-shifting can be easily found.

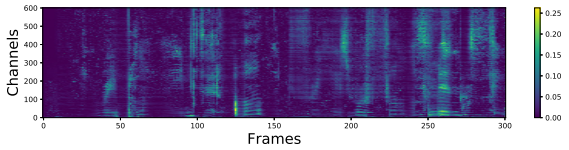
4.5. Ablation Study and Discussion

When applying PGD, we need to optimize certain metrics of given audio, we will choose some perturbation carefully such that the audio differs the most in that metric. As a result, different targets of PGD may result in different performances. In this experiment, we tried different targets and compare their performance. More specifically, different PGD attacks with the target to make the output mask all 0/0.5/1 are tested and compared, the result is shown in table 2. Surprisingly, the target that changes the output mask more zero-like has the best attacking performance.

To make a conjecture, in PGD with a zero-like target, the perturbation is chosen to hide all the signals. As a result, features that help VoiceFilter to separate may be hidden,



(a) Perturbation by time-shifting attack



(b) Output difference with this attack

Figure 4. In each figure, the x-coordinate represents for time frame, each frame represents 10ms, and the y-coordinate represents for frequency channel. This is a result of the time-shifting attack. Fig 4(a) shows the perturbation made by the adversarial attack. Fig 4(b) shows the error between the output and the target

and the outputted audio has a large difference comparing with the target audio.

target	SDR
one-like	0.28
half-like	0.02
zero-like	-1.38

Table 2. This table shows SDR of different attack target

5. Conclusions

In this paper, we proposed an adversarial-attack-based method to protect our privacy. By adding a tiny perturbation on mixed audio, we can significantly reduce the performance of voice separation models.

In this paper, we notice that user privacy could eavesdrop with voice separation models, thus we proposed a simple method to protect user privacy from such risks. We focus on voice separation models that use spectrogram and add perturbation to the spectrogram of mixed audio to obstruct separation models. We give experiments to show how our method works on normal attack scenarios and scenarios with time delay.

Although our method can solve some time delay cases, it is still far from practical application. In future work, we will try to predict the next audio and add perturbation in real-time. And this work only focuses on one type of voice separation model, other types of voice separation models need to be considered in future work.

6. Boarder Impacts

The method of blocking voice separation models may be maliciously used to block some voice separation model that is normally needed. But we still think that such an approach has more advantages than disadvantages for society.

References

[1] Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman. The conversation: Deep audio-visual speech enhancement. In B. Yegnanarayana, editor, *Interspeech 2018, 19th Annual Conference of the International Speech Communication Association, Hyderabad, India, 2-6 September 2018*, pages 3244–3248. ISCA, 2018. 1

[2] Ariel Ephrat, Inbar Mosseri, Oran Lang, Tali Dekel, Kevin Wilson, Avinatan Hassidim, William T. Freeman, and Michael Rubinstein. Looking to listen at the cocktail party: a speaker-independent audio-visual model for speech separation. *ACM Trans. Graph.*, 37(4):112:1–112:11, 2018. 1

[3] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. 1, 2

[4] John R. Hershey, Zhuo Chen, Jonathan Le Roux, and Shinji Watanabe. Deep clustering: Discriminative embeddings for segmentation and separation. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2016, Shanghai, China, March 20-25, 2016*, pages 31–35. IEEE, 2016. 1

[5] Xugang Lu, Yu Tsao, Shigeki Matsuda, and Chiori Hori. Speech enhancement based on deep denoising autoencoder. In Frédéric Bimbot, Christophe Cerisara, Cécile Fougeron, Guillaume Gravier, Lori Lamel, François Pellegrino, and Pascal Perrier, editors, *INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association, Lyon, France, August 25-29, 2013*, pages 436–440. ISCA, 2013. 2

[6] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. 1, 2

[7] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. Deepfool: A simple and accurate method to fool deep neural networks. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 2574–2582. IEEE Computer Society, 2016. 1

[8] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In Yoshua Bengio and Yann LeCun, editors, *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014. 2

[9] Yapeng Tian, Chenliang Xu, and Dingzeyu Li. Deep audio prior. *CoRR*, abs/1912.10292, 2019. 2

[10] DeLiang Wang and Jitong Chen. Supervised speech separation based on deep learning: An overview. *IEEE ACM Trans. Audio Speech Lang. Process.*, 26(10):1702–1726, 2018. 2

[11] Quan Wang, Hannah Muckenhirn, Kevin W. Wilson, Prashant Sridhar, Zelin Wu, John R. Hershey, Rif A. Saurous, Ron J. Weiss, Ye Jia, and Ignacio Lopez-Moreno. Voicefilter: Targeted voice separation by speaker-conditioned spectrogram masking. In Gernot Kubin and Zdravko Kacic, editors, *Interspeech 2019, 20th Annual Conference of the International Speech Communication Association, Graz, Austria, 15-19 September 2019*, pages 2728–2732. ISCA, 2019. 1, 2

[12] Yuxuan Wang and DeLiang Wang. Boosting classification based speech separation using temporal dynamics. In *INTERSPEECH 2012, 13th Annual Conference of the International Speech Communication Association, Portland, Oregon, USA, September 9-13, 2012*, pages 1528–1531. ISCA, 2012. 2

[13] Yuxuan Wang and DeLiang Wang. Cocktail party processing via structured prediction. In Peter L. Bartlett, Fernando C. N. Pereira, Christopher J. C. Burges, Léon Bottou, and Kilian Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012. Proceedings of a meeting held December 3-6, 2012, Lake Tahoe, Nevada, United States*, pages 224–232, 2012. 2

[14] Yuxuan Wang and DeLiang Wang. Towards scaling up classification-based speech separation. *IEEE Trans. Speech Audio Process.*, 21(7):1381–1390, 2013. 2