DEFENSE AGAINST ADVERSARIAL ATTACKS ON SPOOFING COUNTERMEASURES OF ASV

Haibin Wu[1], Songxiang Liu[2]*, Helen Meng[2], Hung-yi Lee[1]*

[1] Speech Processing and Machine Learning Laboratory, National Taiwan University
[2]Human-Computer Communications Laboratory, The Chinese University of Hong Kong

* Equal contribution.

OUTLINE

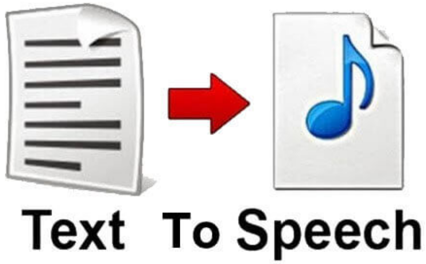
- Motivation
- Adversarial attack
- Defense
- Experiment
- Conclusion

Motivation

Background – Anti-spoofing

- A great number of automatic speaker verification (ASV) models with high accuracy have been proposed.
- However, high-performance ASV may still be attacked by spoofing audios
- These spoofing audios are audios generated by replay, text-to-speech or voice conversion.

audio replay



text to speech

voice conversion

Background – Anti-spoofing

Voice conversion

Anti-spoofing Model

Spoofing

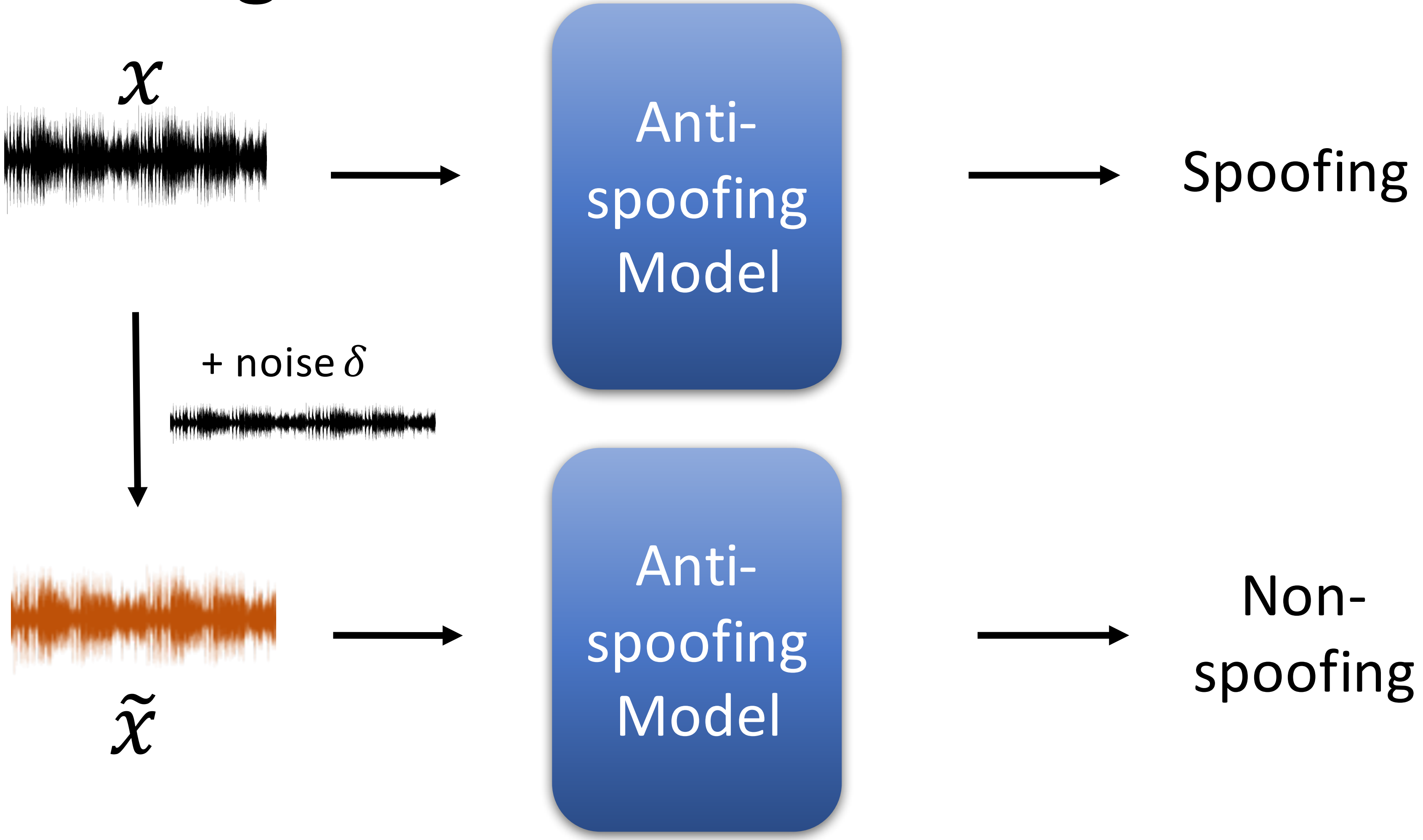
User’s audio

Anti-spoofing Model

Non-spoofing

Anti-spoofing models with high-performance are proposed.

Background – Adversarial Attack

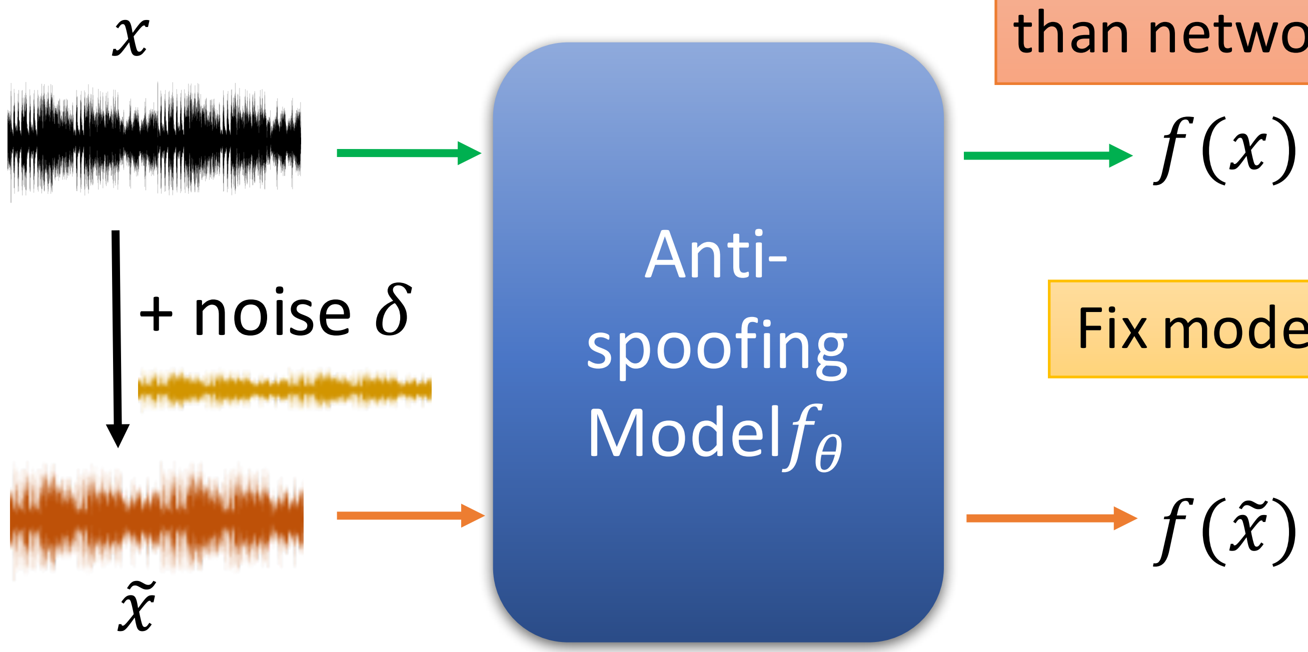


- Address the vulnerability of anti-spoofing systems to adversarial attacks and devise defense methods.

Adversarial attack

Attack – Finding Adversarial Example

Just like training a neural network, but we optimize input x rather than network parameter θ

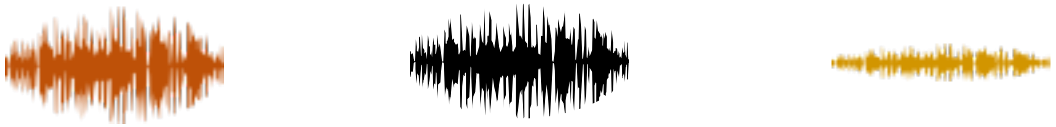


Fix model parameters

Find a suitable δ such that

$$max_{\|\delta\|_\infty \leq \epsilon} Diff\left(f(x), f(\tilde{x})\right)$$

$$\tilde{x} \quad = \quad x \quad + \quad \delta$$



Attack Method: Projected Gradient Descent

$$x^* = arg \max_{\|\delta\|_\infty \leq \epsilon} Diff(f(x), f(\tilde{x}))$$

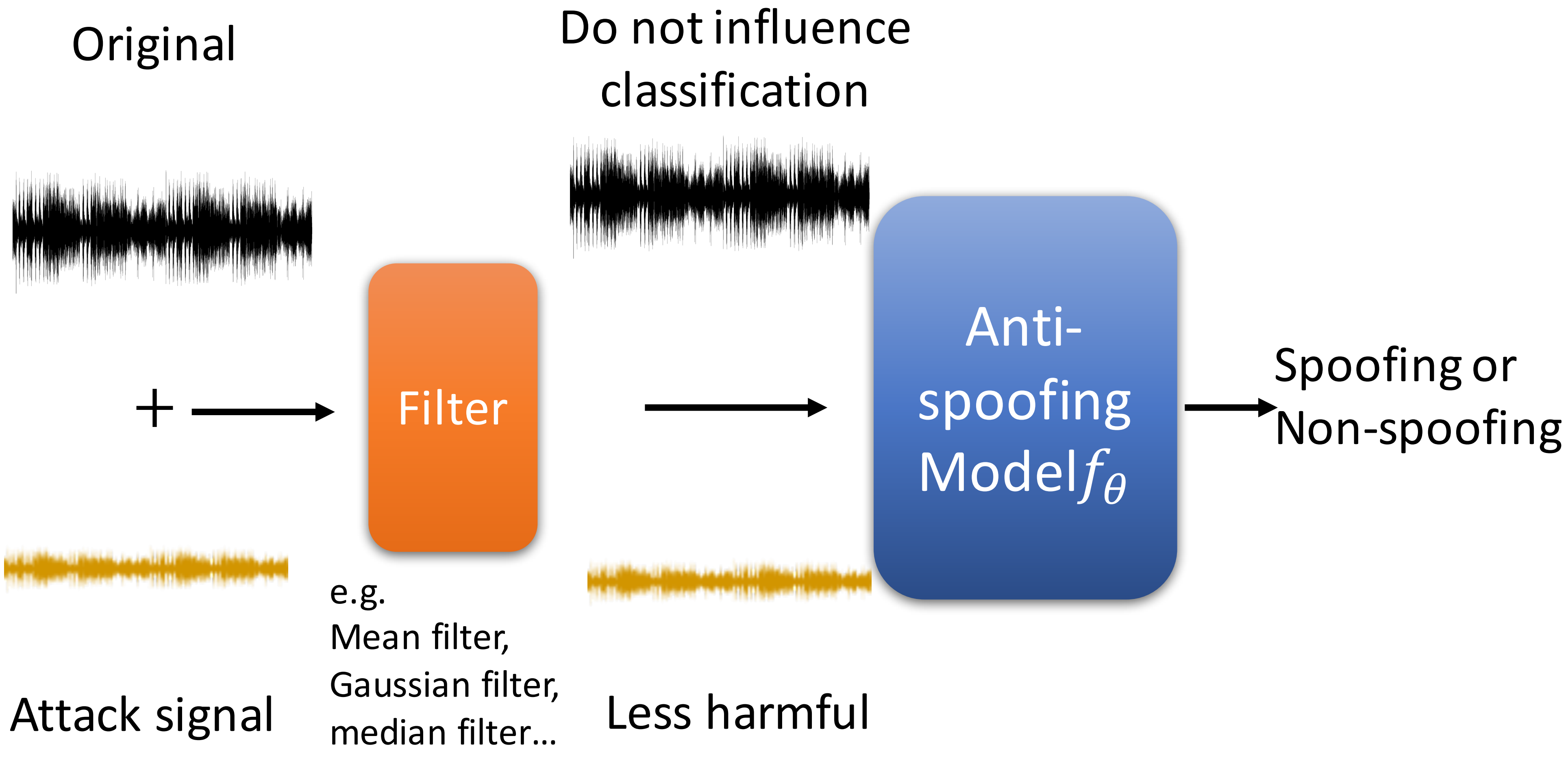
An iterative method. Starting from input $x_0 = x$, then it is iteratively updated as:

$$x_{k+1} = clip(x_k + \alpha \cdot sign\left(\nabla_{x_k} Diff\left(f(x), f(x_k)\right)\right), \quad for\ k = 0, \dots, K-1$$

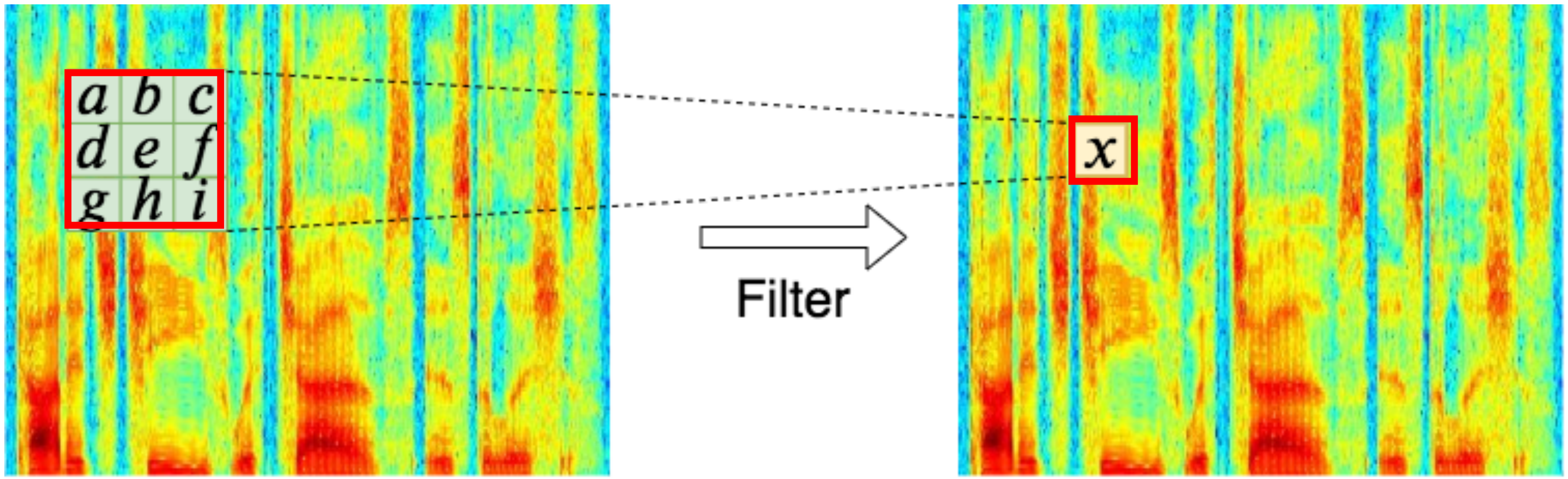
α is step size, K is the iteration number and the $clip(\cdot)$ is the clipping function.

Defense

Defense Method 1: Spatial Smoothing



Defense Method 1: Spatial Smoothing



Mean filter: $x = \frac{1}{9}(a + b + c + d + e + f + g + h + i)$

Gaussian filter: $x = \frac{1}{16}(a + 2b + c + 2d + 4e + 2f + g + 2h + i)$

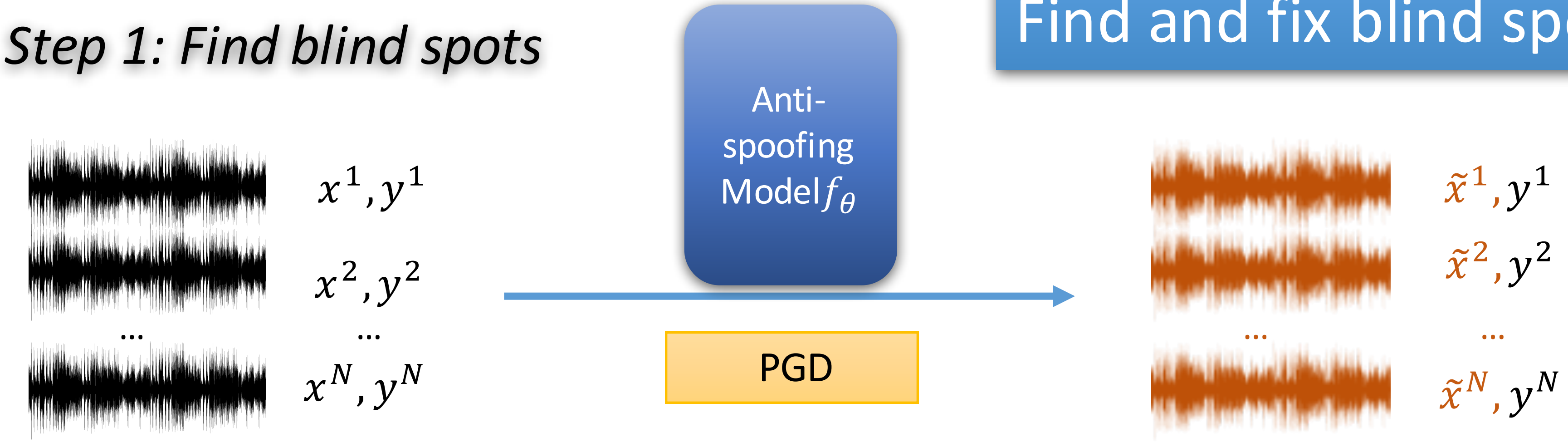
Median filter: $x = the\ median\ of\ (a, b, c, d, e, f, g, h. i)$

Defense Method 2: Adversarial Training

Then we do the two steps iteratively

Idea:
Find and fix blind spots

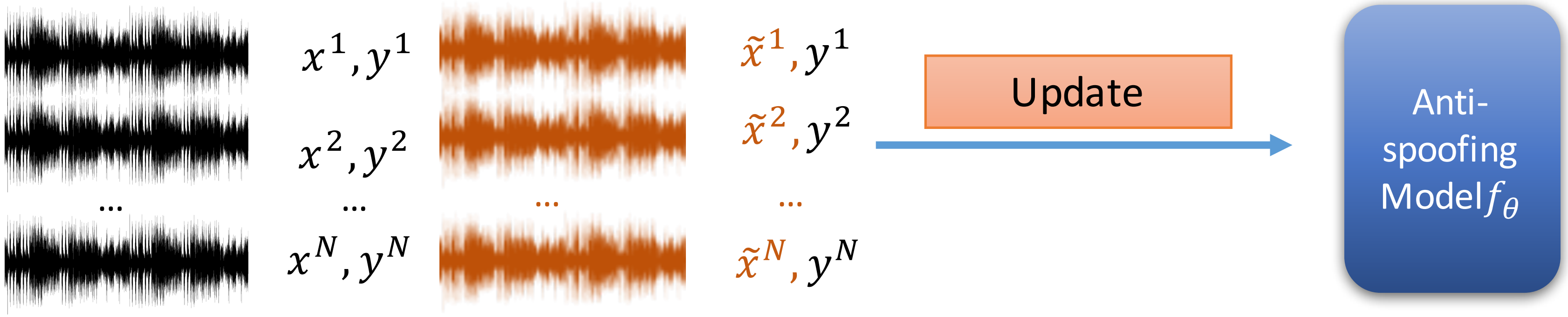
Step 1: Find blind spots



original examples

adversarial examples

Step 2: Fix blind spots



Experiment

Experiment setup

Dataset

LA partition of ASVspoof 2019 challenge which involves synthesized audios from TTS and VC models.

Two different anti-spoofing models

SENet [Lai et al. 2019] and VGG [Zeinali et al. 2018]

Attack method

PGD

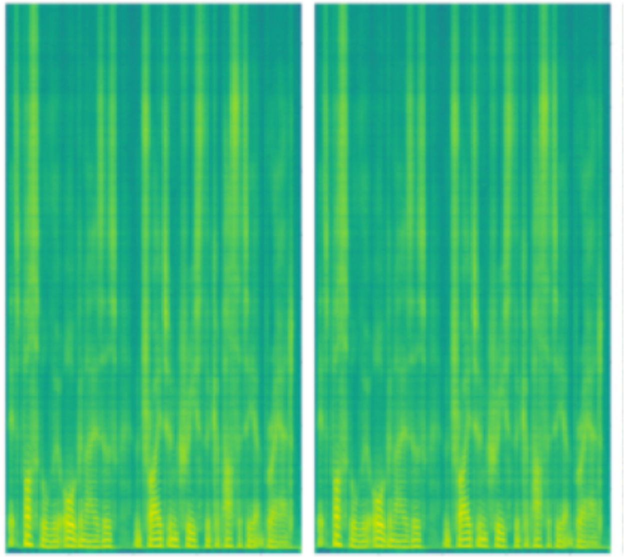
Two defense methods

Spatial smoothing and adversarial training

Experiment result: SENet

Testing accuracies

	Before adversarial training
Normal examples	99.97%
Adversarial examples	48.32%



Normal example adversarial example

- The SENet is subject to adversarial attacks.
- Listeners can not tell the difference between adversarial example and original example.

Experiment result: SENet

Testing accuracies

	Before adversarial training	After adversarial training
Normal examples	99.97%	99.75%
Adversarial examples	48.32%	92.40%
Adversarial examples + median filter	82.00%	93.74%
Adversarial examples + mean filter	82.39%	93.76%
Adversarial examples + Gaussian filter	78.93%	83.72%

- All three kinds of filters have considerable performance in improving the robustness of anti-spoofing models against adversarial examples.
- The improvement of Gaussian filter is much less than the other two filters.

Experiment result: SENet

Testing accuracies

	Before adversarial training	After adversarial training
Normal examples	99.97%	99.75%
Adversarial examples	48.32%	92.40%
Adversarial examples + median filter	82.00%	93.74%
Adversarial examples + mean filter	82.39%	93.76%
Adversarial examples + Gaussian filter	78.93%	83.72%

- Adversarial training improves the robustness of anti-spoofing models.

Experiment result: SENet

Testing accuracies

	Before adversarial training	After adversarial training
Normal examples	99.97%	99.75%
Adversarial examples	48.32%	92.40%
Adversarial examples + median filter	82.00%	93.74%
Adversarial examples + mean filter	82.39%	93.76%
Adversarial examples + Gaussian filter	78.93%	83.72%

- Equipping adversarial training with median filter or mean filter increases the testing accuracy for adversarial examples compared with just using adversarial training.
- While adding Gaussian filter decreases the testing accuracy.

Experiment result: VGG

Testing accuracies

	Before adversarial training	After adversarial training
Normal examples	99.99%	99.99%
Adversarial examples	37.06%	98.60%
Adversarial examples + median filter	92.72%	98.96%
Adversarial examples + mean filter	93.95%	99.24%
Adversarial examples + Gaussian filter	84.39%	87.22%

- We can see a similar phenomenon for VGG

Conclusion

Conclusion

- Both adversarial training and spatial smoothing can make the anti-spoofing models robust enough to counter adversarial attacks.
- More advanced defense methods should be adopted to improve the robustness of anti-spoofing models.