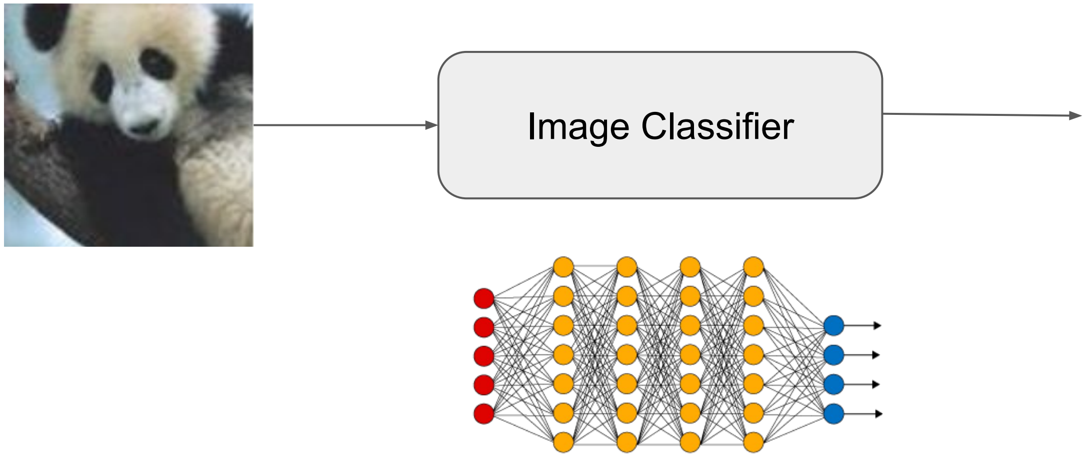
Towards a Certified Defense for Audio Adversarial Examples

MIT PRIMES Computer Science Conference
October 20, 2019

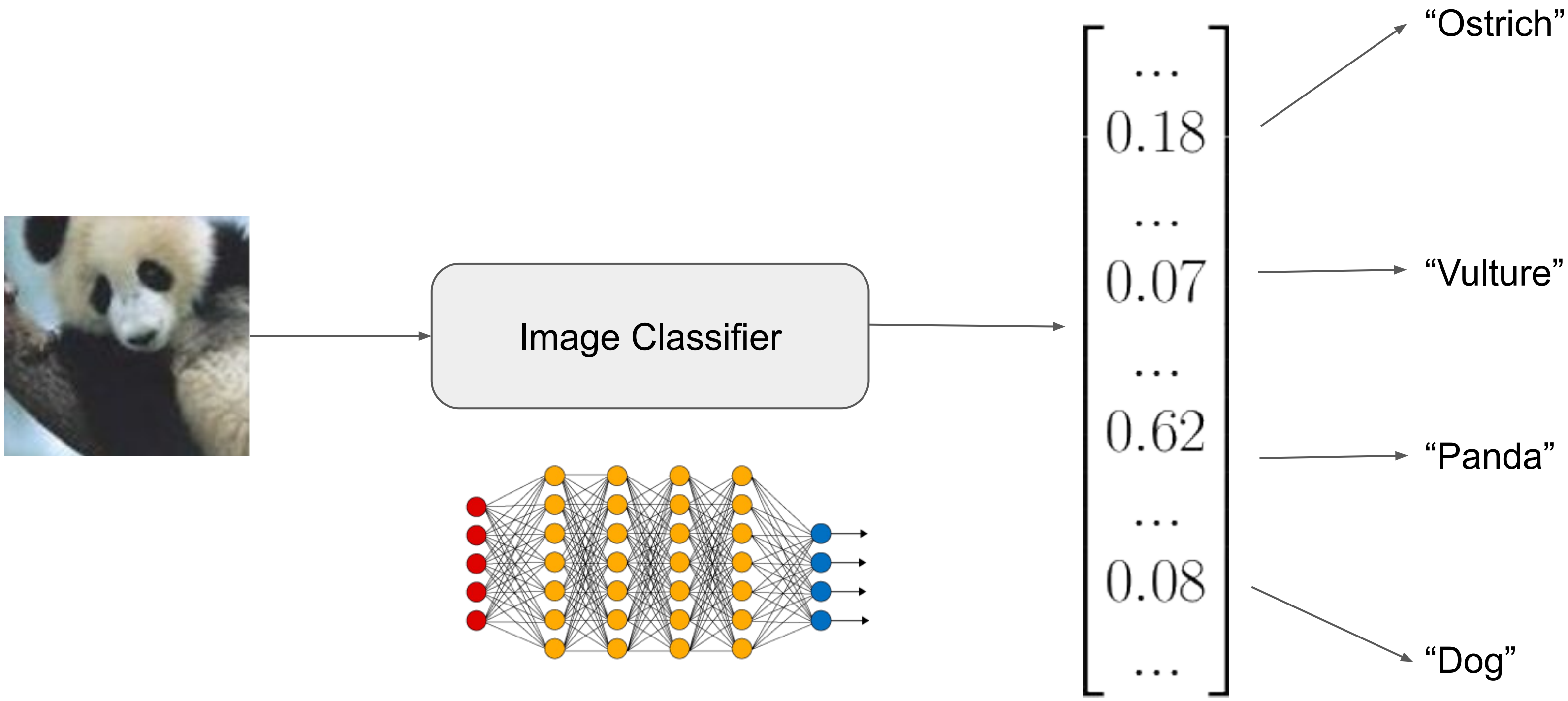
By: Ethan Mendes
Mentor: Kyle Hogan

Conventional Classification Process



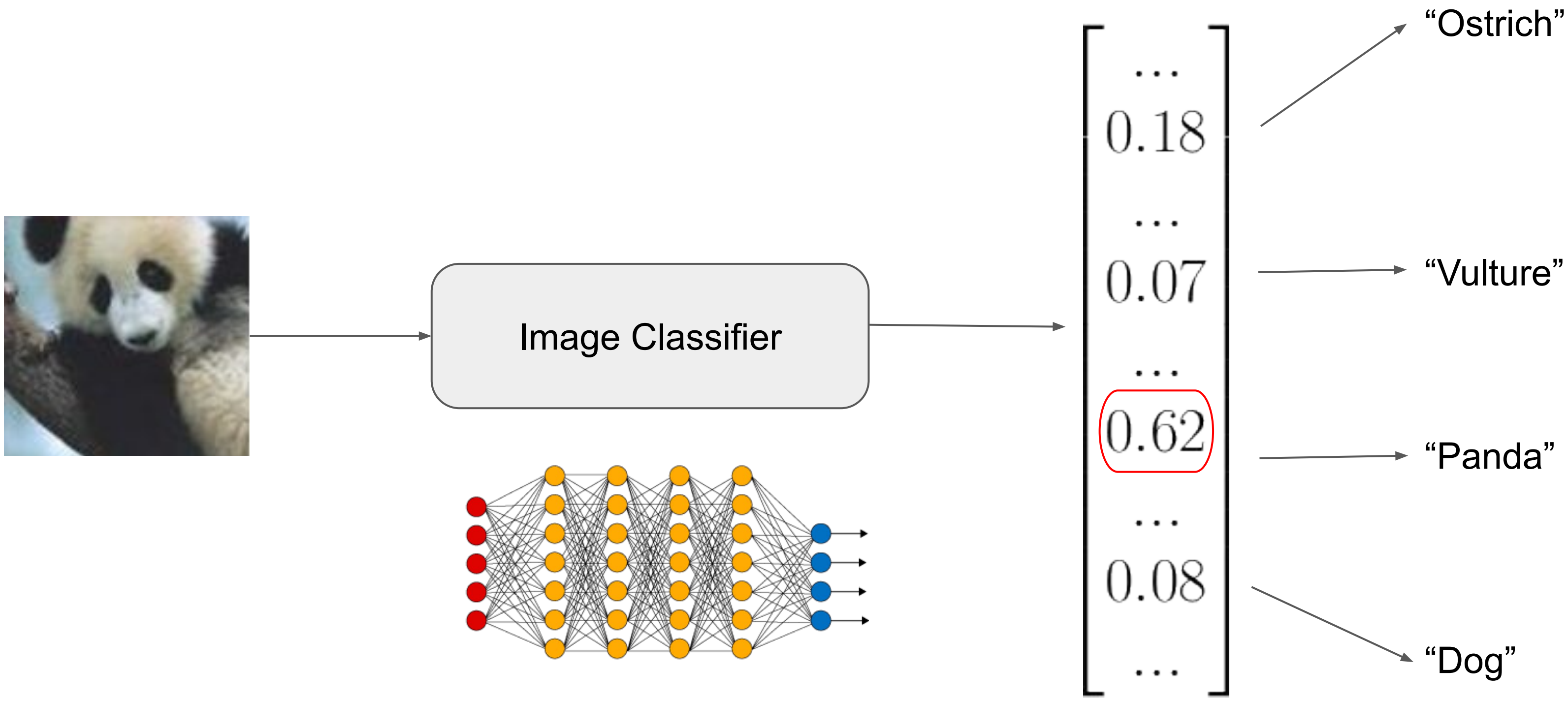
Goodfellow et al.

Conventional Classification Process



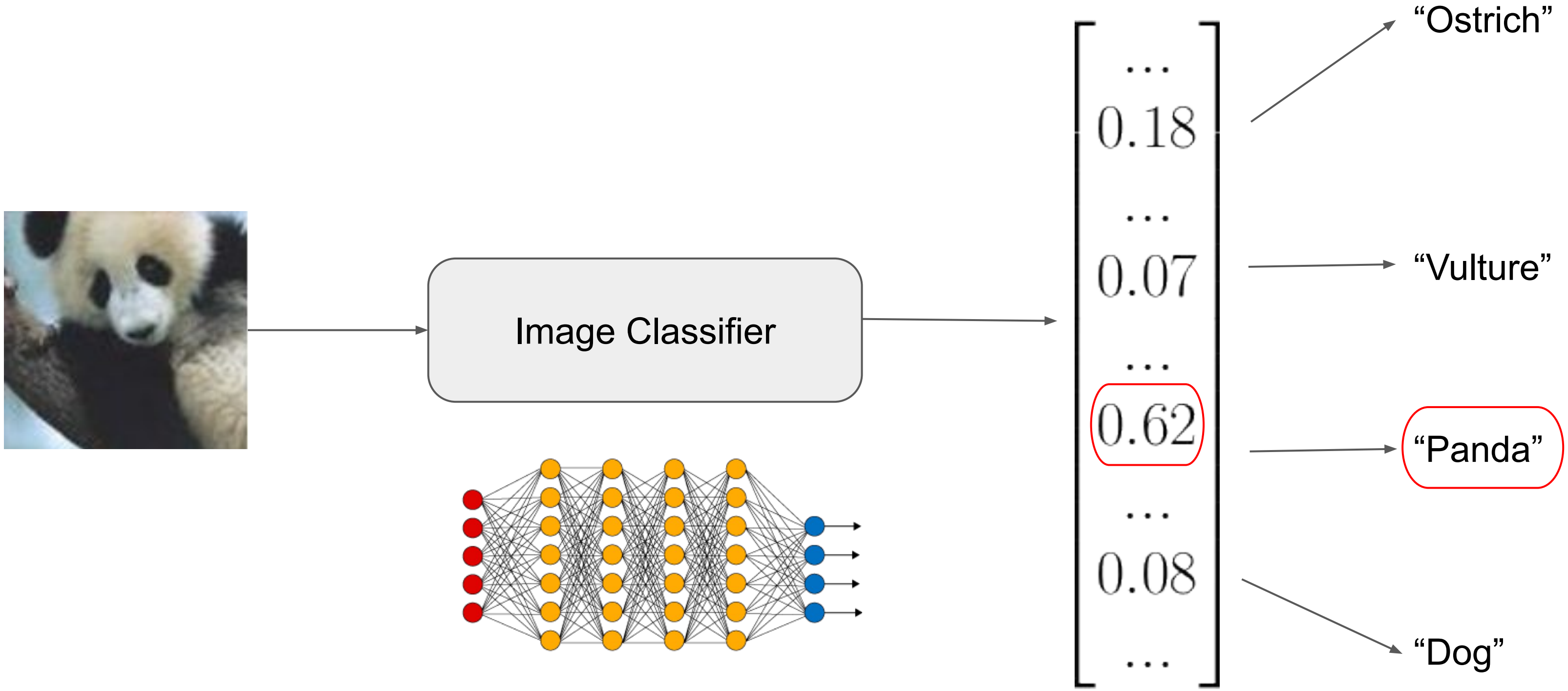
Goodfellow et al.

Conventional Classification Process



Goodfellow et al.

Conventional Classification Process



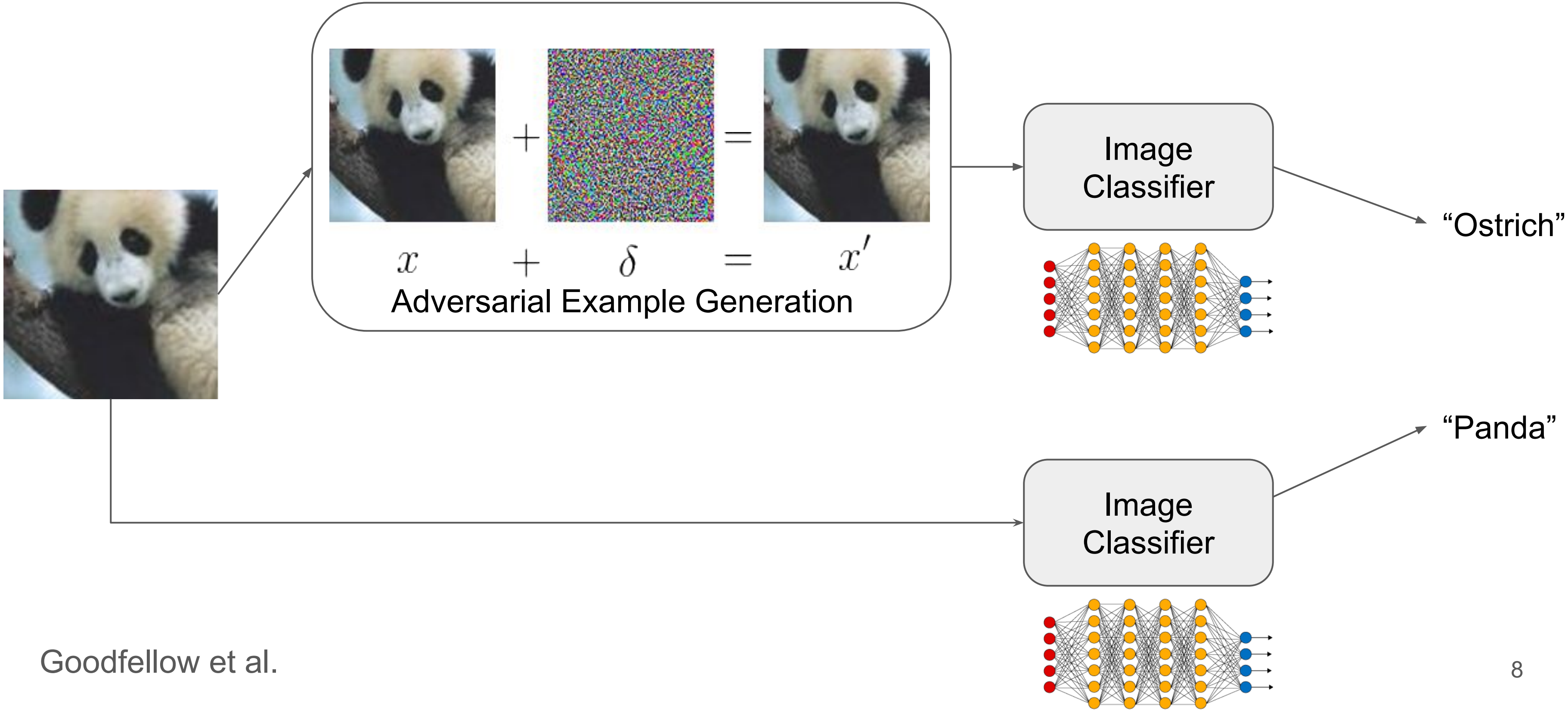
Goodfellow et al.

Adversarial Example Background

Adversarial Examples

Athalye et al.

Adversarial Examples



Goodfellow et al.

Consequences of Adversarial Examples

- **Self Driving Cars**:
  - Accidents can result from the signs with stickers or grafiti which cause false classifications

- **Smart Speakers**:
  - Audio adversarial examples originating from TV or radio can maliciously interact with smart home devices (turn on lights, unlock doors) without the owner’s knowledge

p-norm

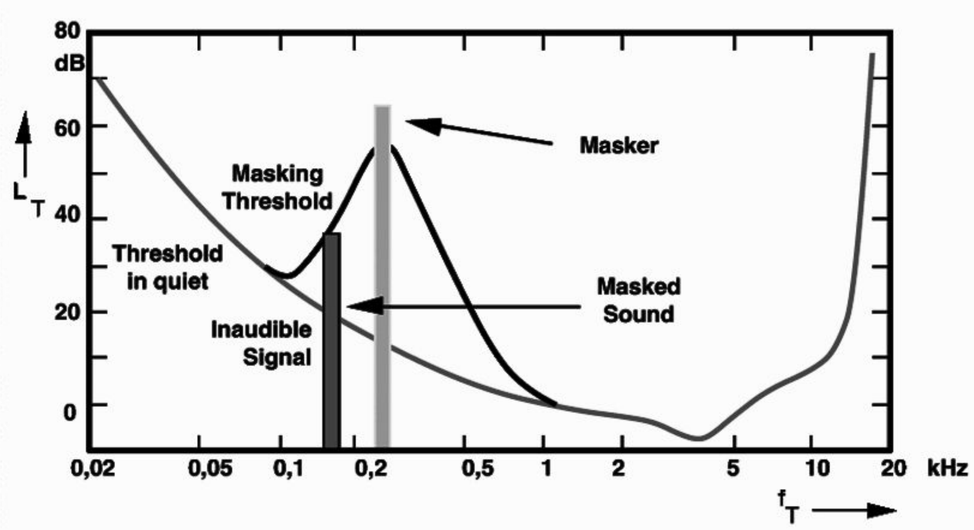
- Constrains the amount of noise that an attacker adds
- For $1 \leq p < \infty$, $||a||_p = (\sum_{i=1}^{n} a_i{}^p)^{1/p}$
- Some special norms
  - Hamming Distance: 0-norm
  - Euclidean Distance: 2-norm
  - Max-norm: ∞-norm
- **These constraints do not work for audio**

0-norm attack[1]

[1] Eykholt et al.

Imperceptible Audio Adversarial Examples

- Attackers create imperceptible adversarial examples by utilizing **auditory masking** (frequency masking)
- Minimize cost functions that take into account imperceptibility and accuracy
- These are usually iterative attacks



Ex. $l(x, \delta, y) = \boxed{l_{net}(f(x+\delta), y)} + \boxed{\alpha \cdot l_\theta(x, \delta)}$ (Carlini et al.)

Accuracy

Imperceptibility

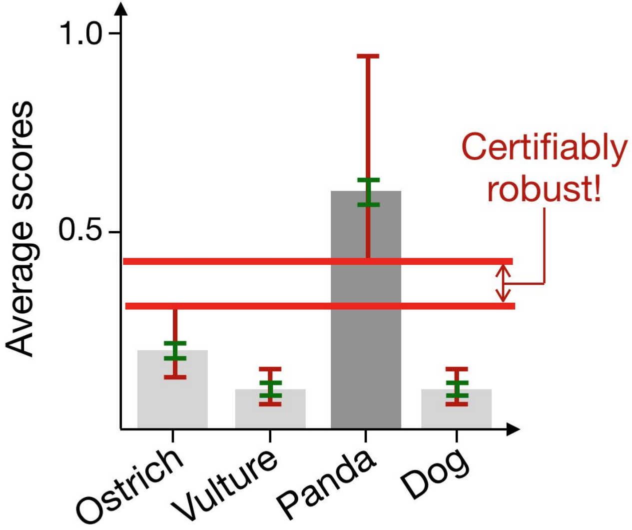
Current Defenses

- Employ MP3 compression and other techniques to remove all noise below the masking threshold
    - Classifier is not trained on this type of filtered data → low accuracy (especially on benign inputs)
    - Filtering removes important information→ even retraining classification network results in low accuracy
    - No provable guarantees

Certified Robustness

Certified Robustness

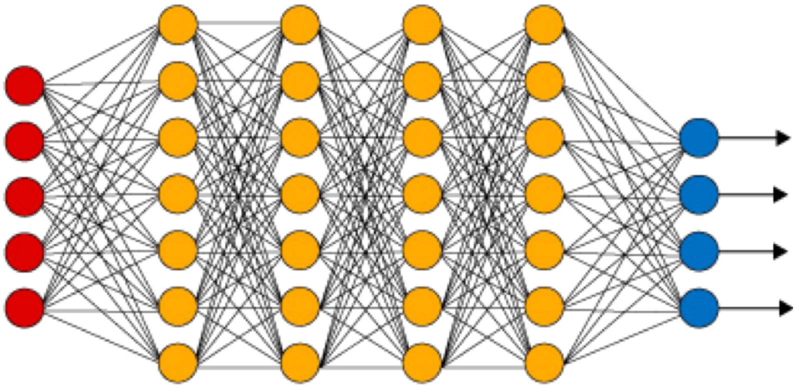
- Provides guarantees of robustness of a defense against bounded attacks using probability theory and statistics for certification



Lecuyer et al.

Certified Robustness via Randomized Smoothing

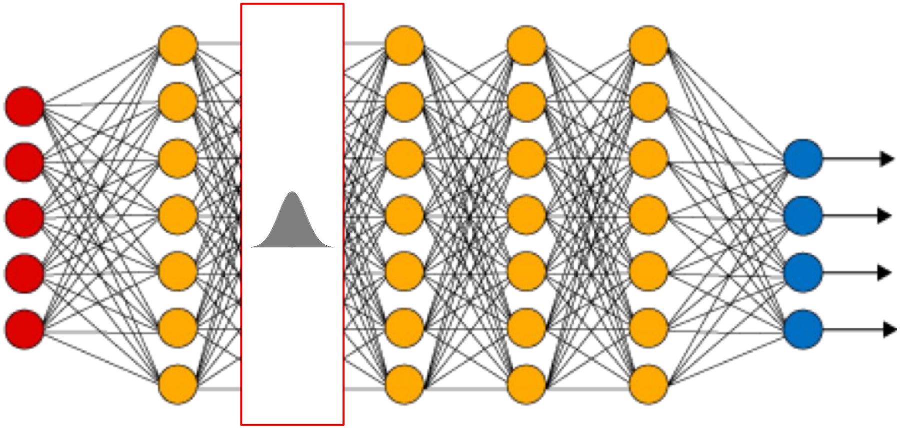
- Add perturbations to the input that exceed the norm-bounded perturbation of the attacker - nullify the adversarial perturbation up to a certain magnitude
- Add a noise layer in the classifier that randomly samples from gaussian or laplacian distributions



Lecuyer et al.

Certified Robustness via Randomized Smoothing

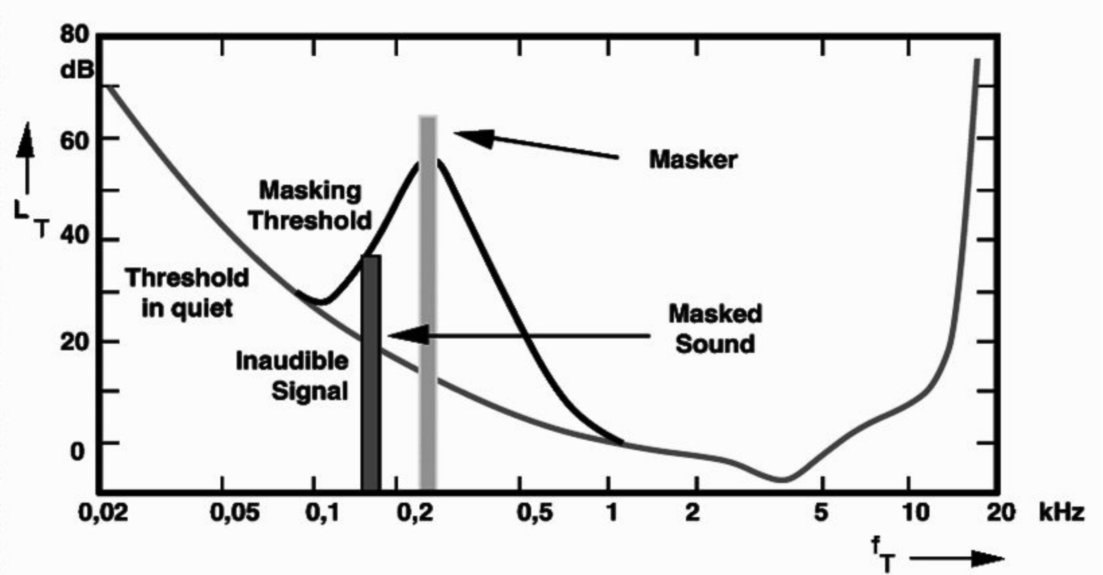
- Add perturbations to the input that exceed the norm-bounded perturbation of the attacker - nullify the adversarial perturbation up to a certain magnitude
- Add a noise layer in the classifier that randomly samples from gaussian or laplacian distributions



Lecuyer et al.

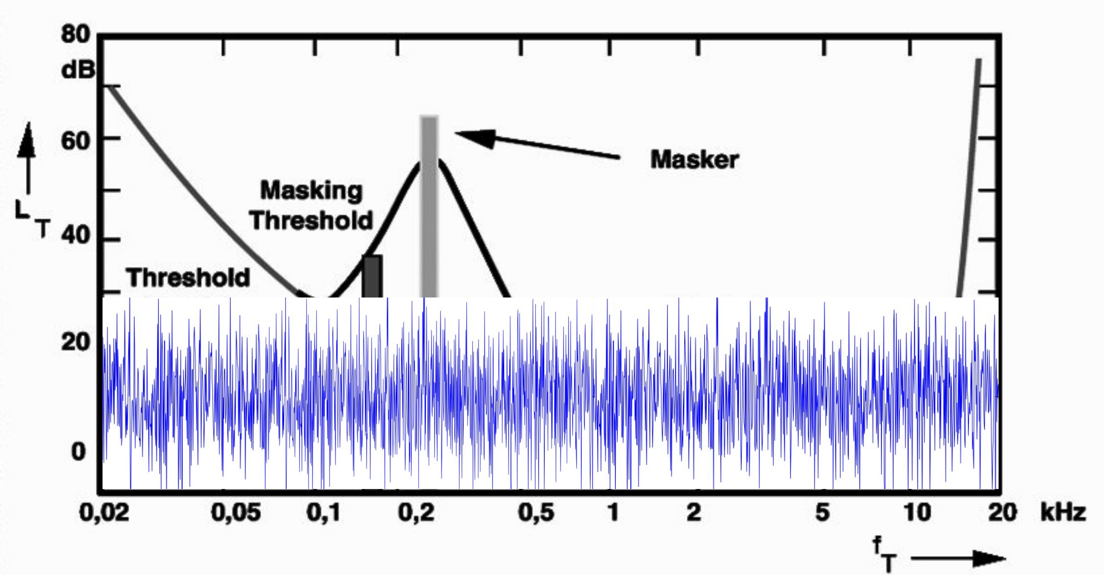
Applying Randomized Smoothing to Audio

- Only works with norm-bounded attacks (images) → imperceptible audio adversarial examples are not norm-bounded
- Noise will not be added to the correct parts of the audio (under the masking threshold)



Applying Randomized Smoothing to Audio

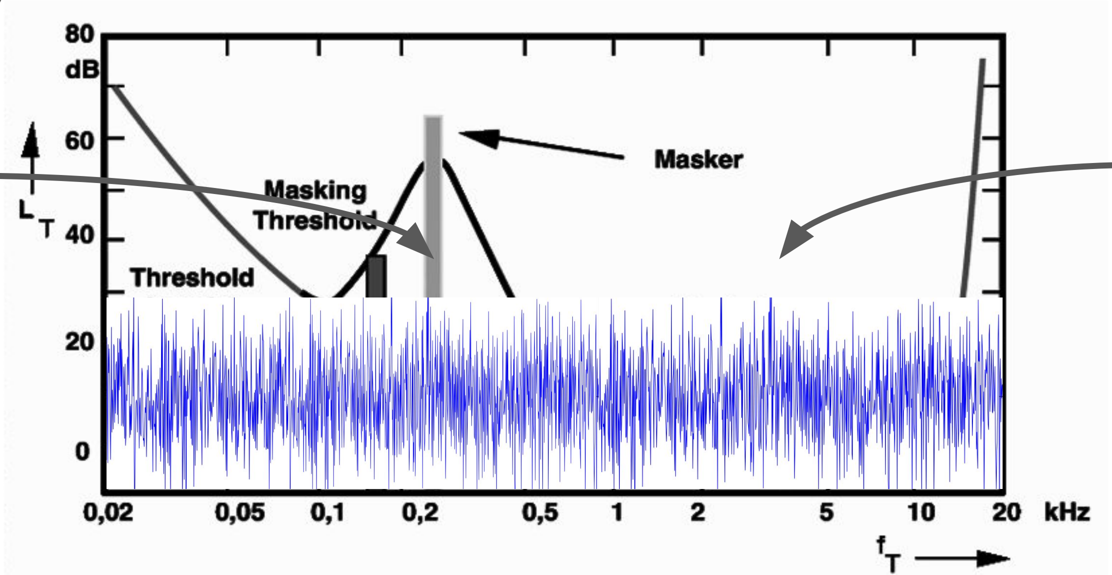
- Only works with norm-bounded attacks (images) → imperceptible audio adversarial examples are not norm-bounded
- Noise will not be added to the correct parts of the audio (under the masking threshold)



Applying Randomized Smoothing to Audio

- Only works with norm-bounded attacks (images) → imperceptible audio adversarial examples are not norm-bounded
- Noise will not be added to the correct parts of the audio (under the masking threshold)

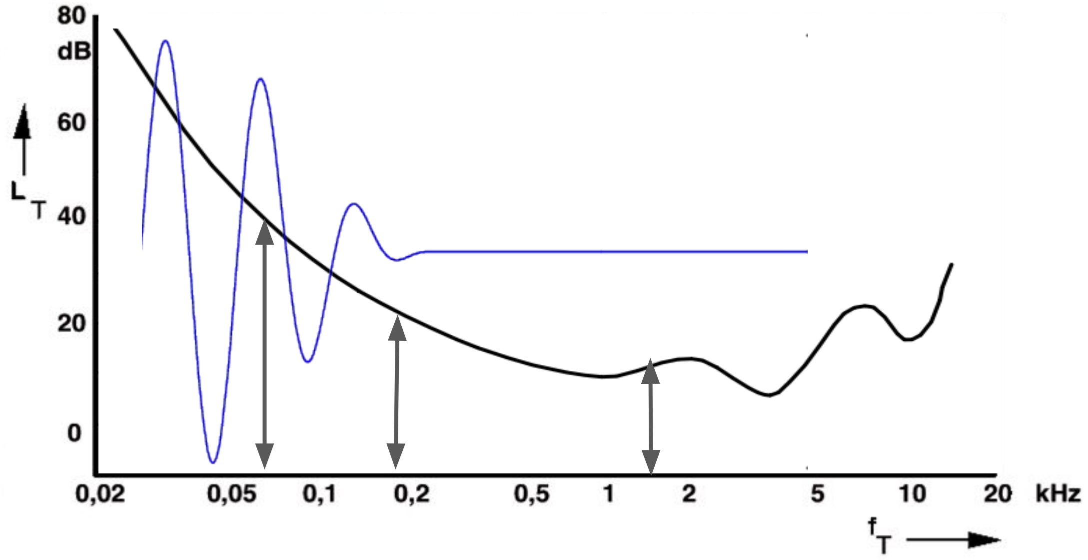
Not enough noise



Too much noise

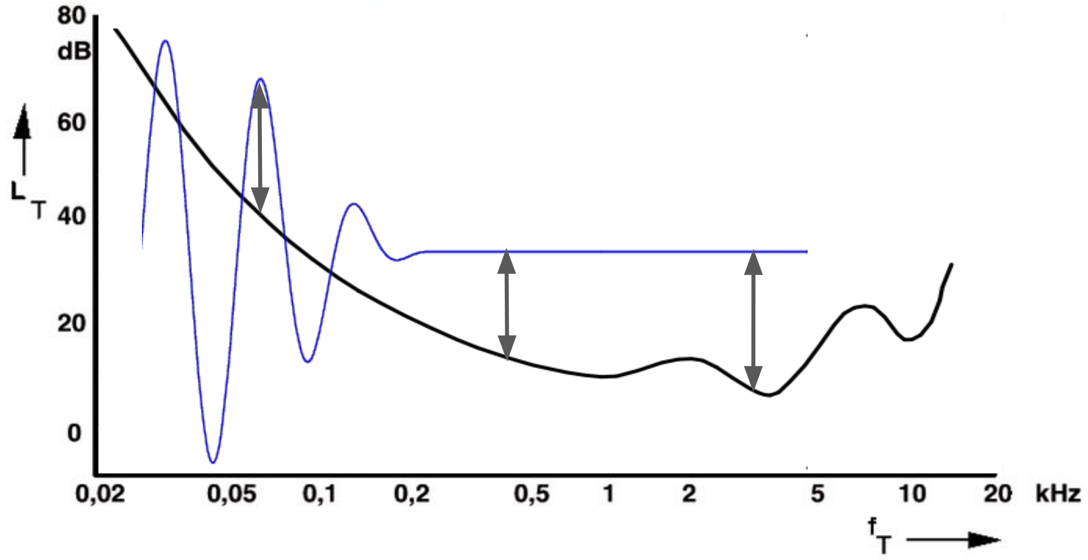
Perturbation Quantification for Audio

- We constrain randomness added in the same way in which attackers add perturbations
  - give scores based threshold and how much attacker exceeds threshold
- Constraint gives basis of how much sound to add to each frequency band



Perturbation Quantification for Audio

- We constrain randomness added in the same way in which attackers add perturbations
  - give scores based threshold and how much attacker exceeds threshold
- Constraint gives basis of how much sound to add to each frequency band



Future Work

- Find a concise mathematical bounding for imperceptible audio adversarial attacks
- Formally prove that proposed method to quantify sound can be used to create certified defenses
- Implement defense and calculate accuracy on both benign and adversarial audio

Acknowledgements

- MIT PRIMES for this incredible opportunity
- My Mentor: Kyle Hogan
- My Parents

Questions?