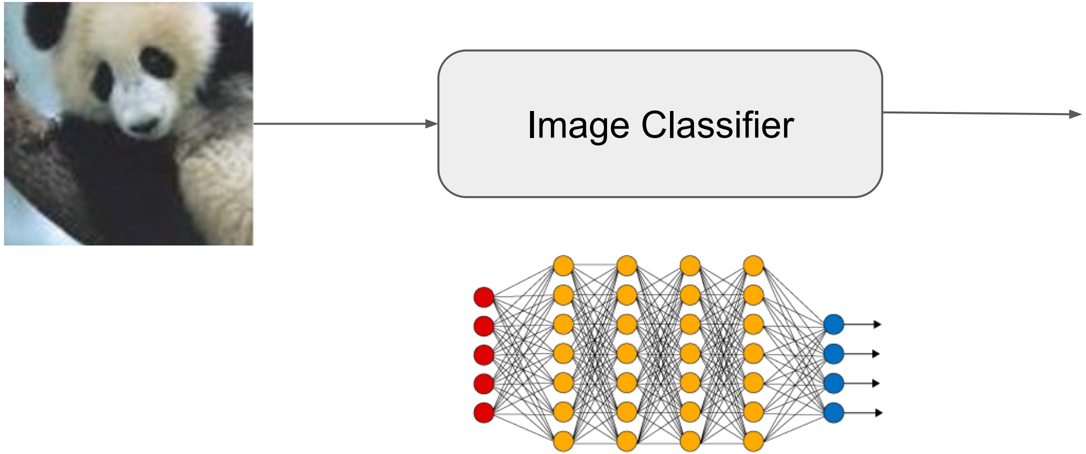
Towards a Robust Defense for Imperceptible Audio Adversarial Examples

MIT PRIMES Computer Science Conference
June 7, 2020

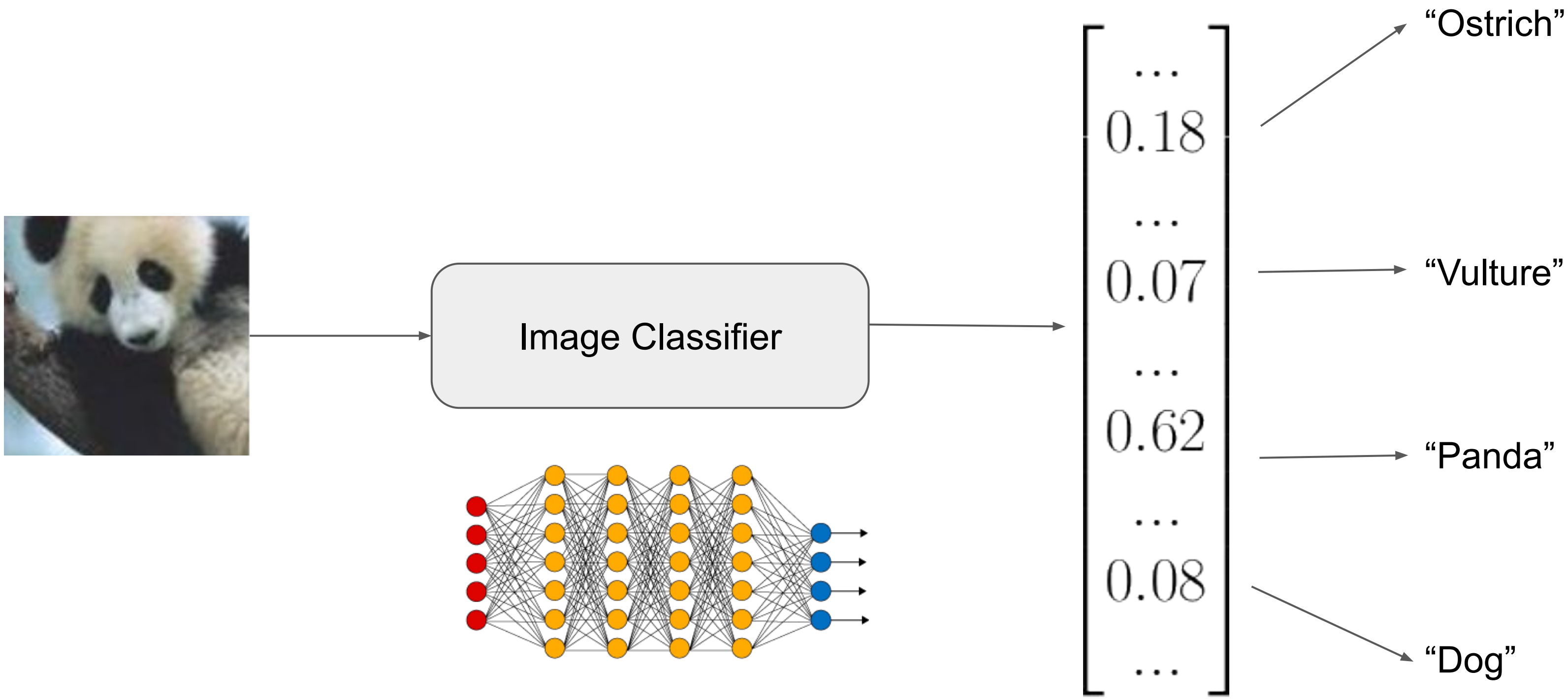
Ethan Mendes
Mentor: Kyle Hogan

Conventional Classification Process



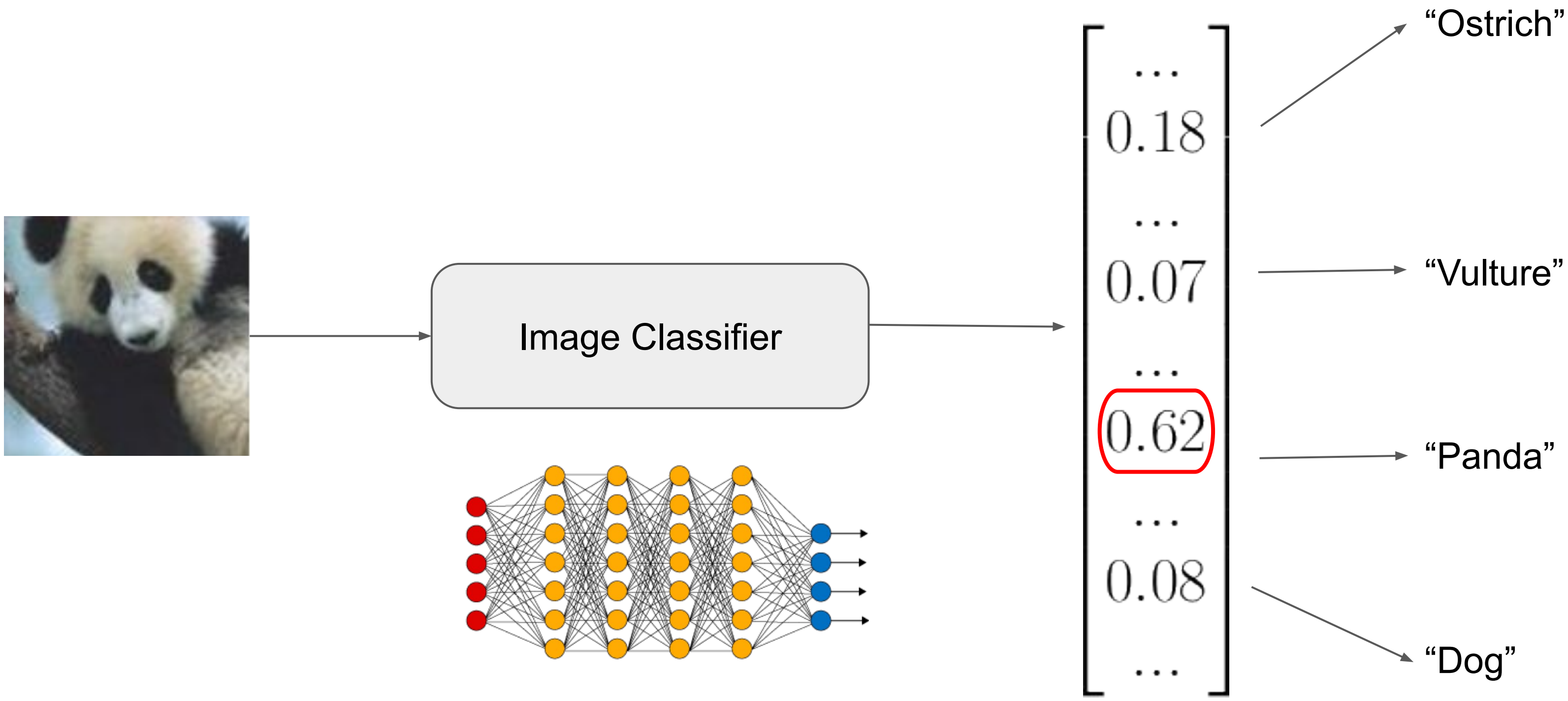
Goodfellow et al.

Conventional Classification Process



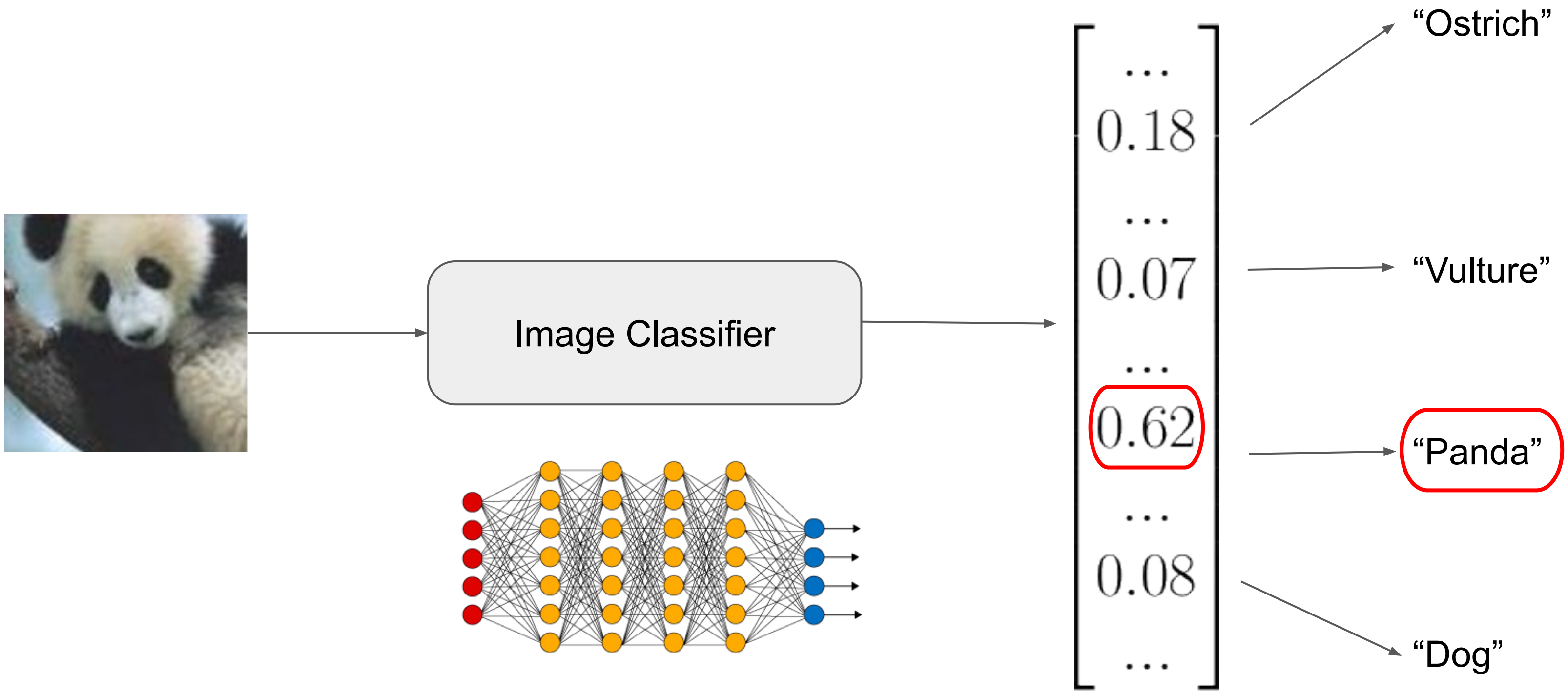
Goodfellow et al.

Conventional Classification Process



Goodfellow et al.

Conventional Classification Process

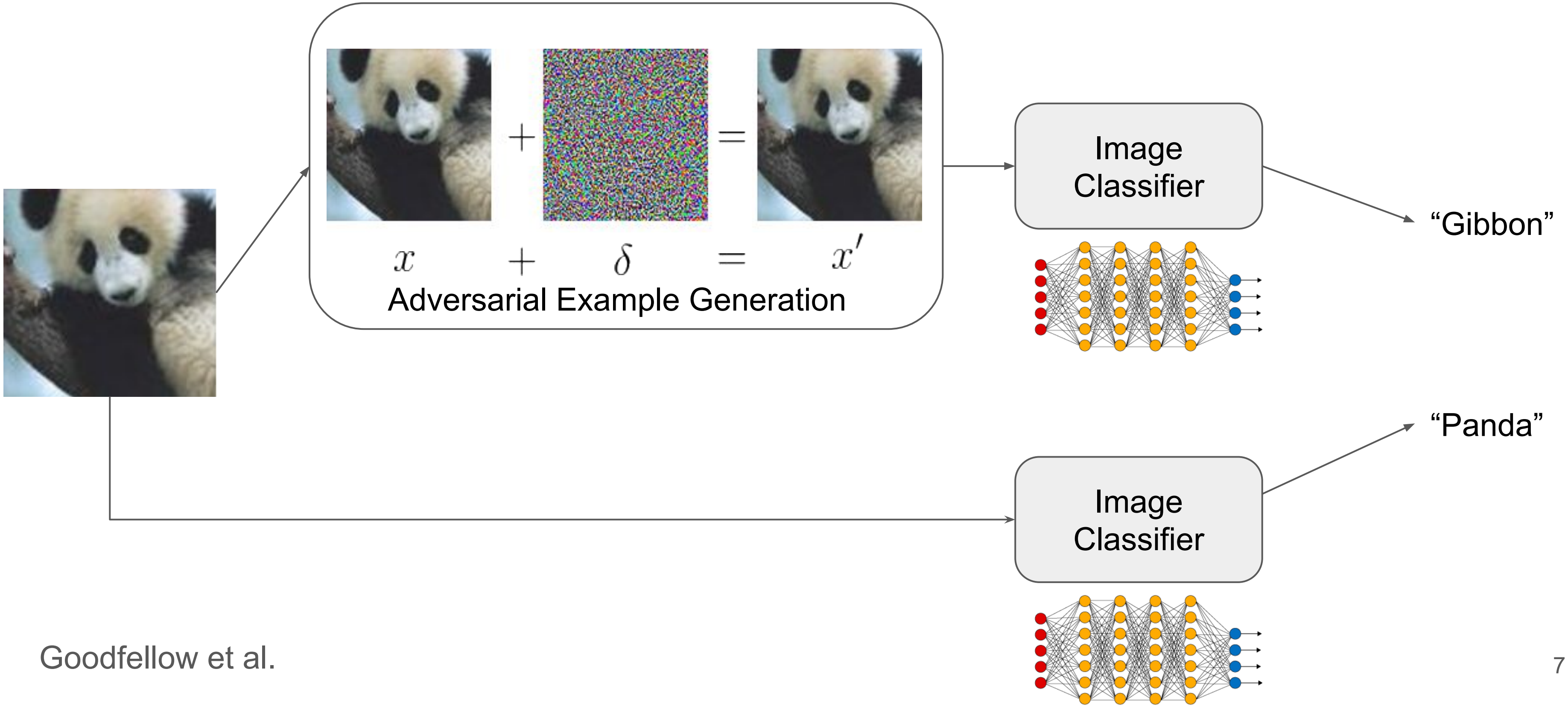


Goodfellow et al.

Adversarial Examples

Athalye et al.

Adversarial Examples



Goodfellow et al.

Consequences of Adversarial Examples

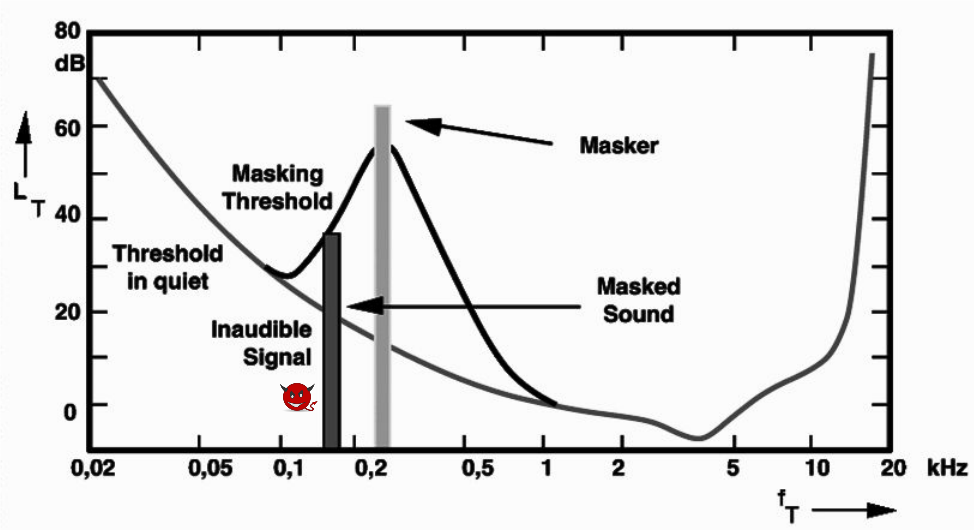
- **Smart Speakers**:
    - Imperceptible audio adversarial examples (AAEs) originating from TV or radio can maliciously interact with smart home devices (turn on lights, unlock doors) without the owner's knowledge

"Alexa, what's the weather?"

Doors Unlocked

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

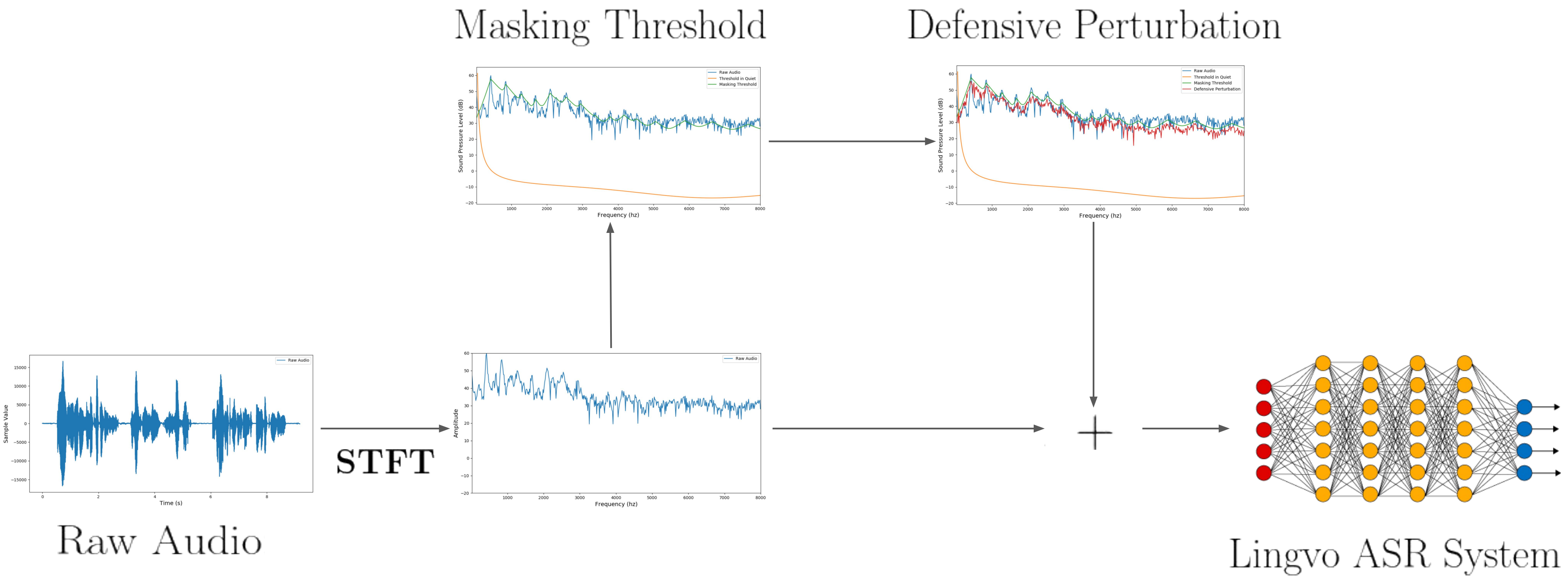


Ex. $l(x, \delta, y) = \underbrace{\boxed{l_{net}(f(x+\delta), y)}}_{\text{Accuracy}} + \underbrace{\boxed{\alpha \cdot l_\theta(x, \delta)}}_{\text{Imperceptibility}}$ (Qin et al.)

Defense Goals

1. Does our defense lower the efficacy of the adversarial examples?
2. Does our defense preserve high accuracy on benign samples?
3. Does our defense revert transcriptions of adversarial examples back to their original transcriptions?

Overview of Defense



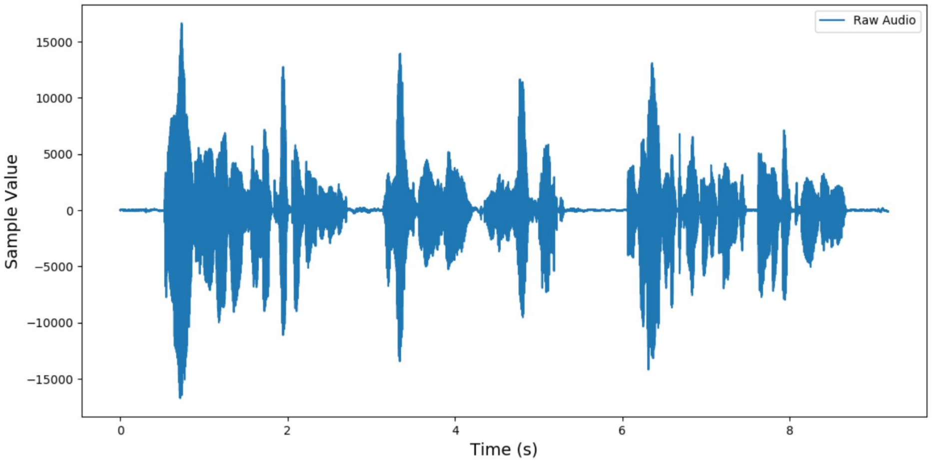
Generating Defense

Raw Audio: $(x(t))$

1. Convert to Frequency Domain: $\mathbf{STFT}(x(t))$

Intended Unperturbed (Benign) Audio Transcription:
"Alexa, what's the weather?"

Intended Adversarially Perturbed Audio Transcription:
"Alexa, open the garage door."

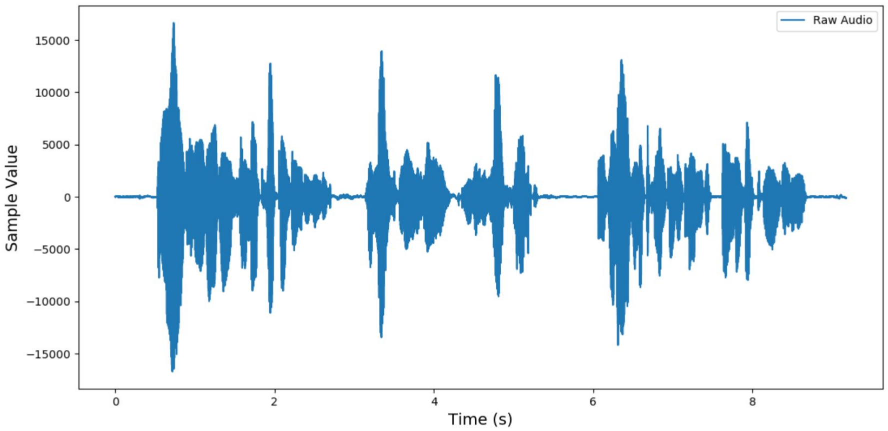


Generating Defense

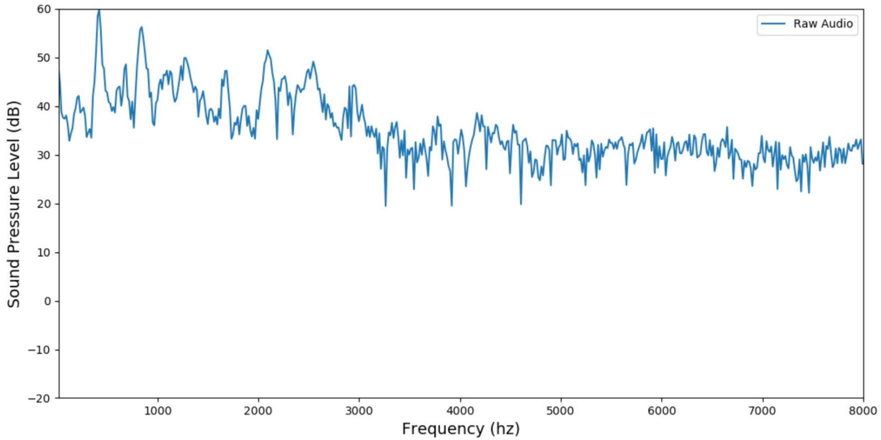
Raw Audio: (x(t))

1. Convert to Frequency Domain: $\mathbf{STFT}(x(t))$

2. Calculate Masking Threshold: $\theta_x(\nu)$



$\mathbf{STFT}(x(t))$

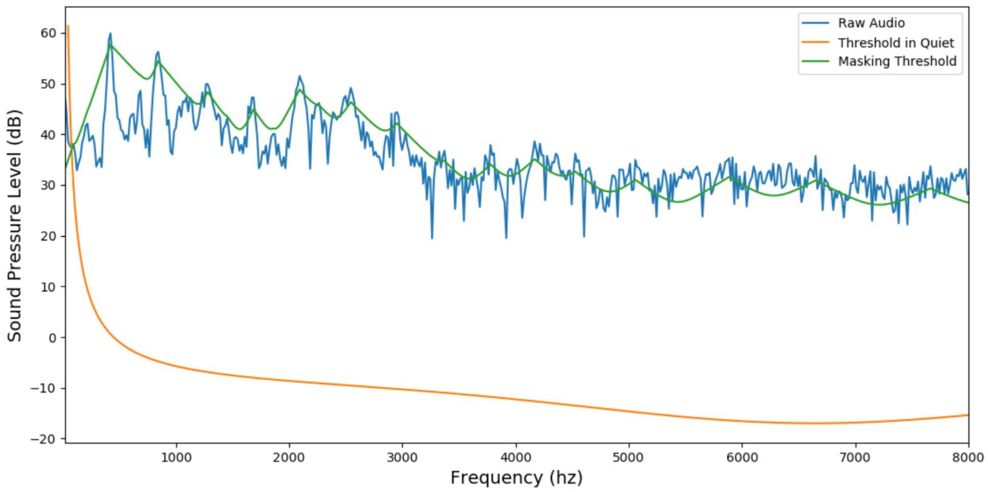


Generating Defense

1. Convert to Frequency Domain: $\mathbf{STFT}(x(t))$

2. Calculate Masking Threshold: $\theta_x(\nu)$

3. Generate Defensive Perturbation: $\delta_D(\nu)$

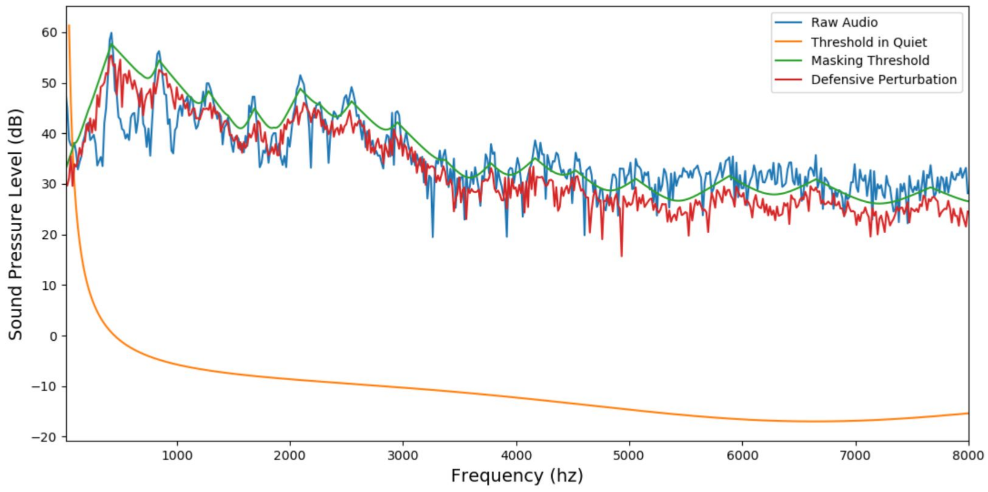


Generating Defense

2. Calculate Masking Threshold: $\theta_x(\nu)$

3. Generate Defensive Perturbation: $\delta_D(\nu)$

4. Generate Input to ASR System: $x(\nu)+\delta_D(\nu)$

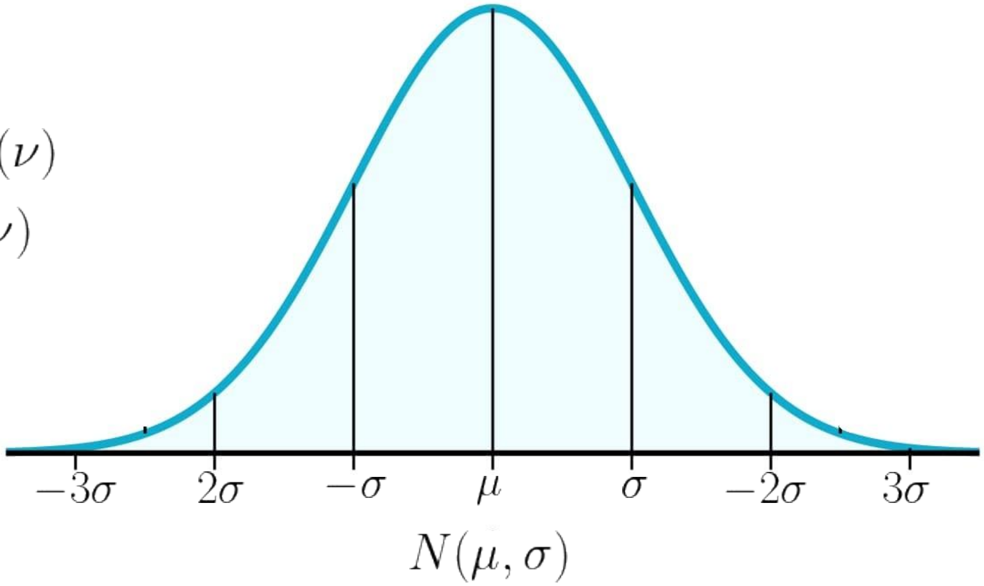


Defensive Perturbation (Definition)

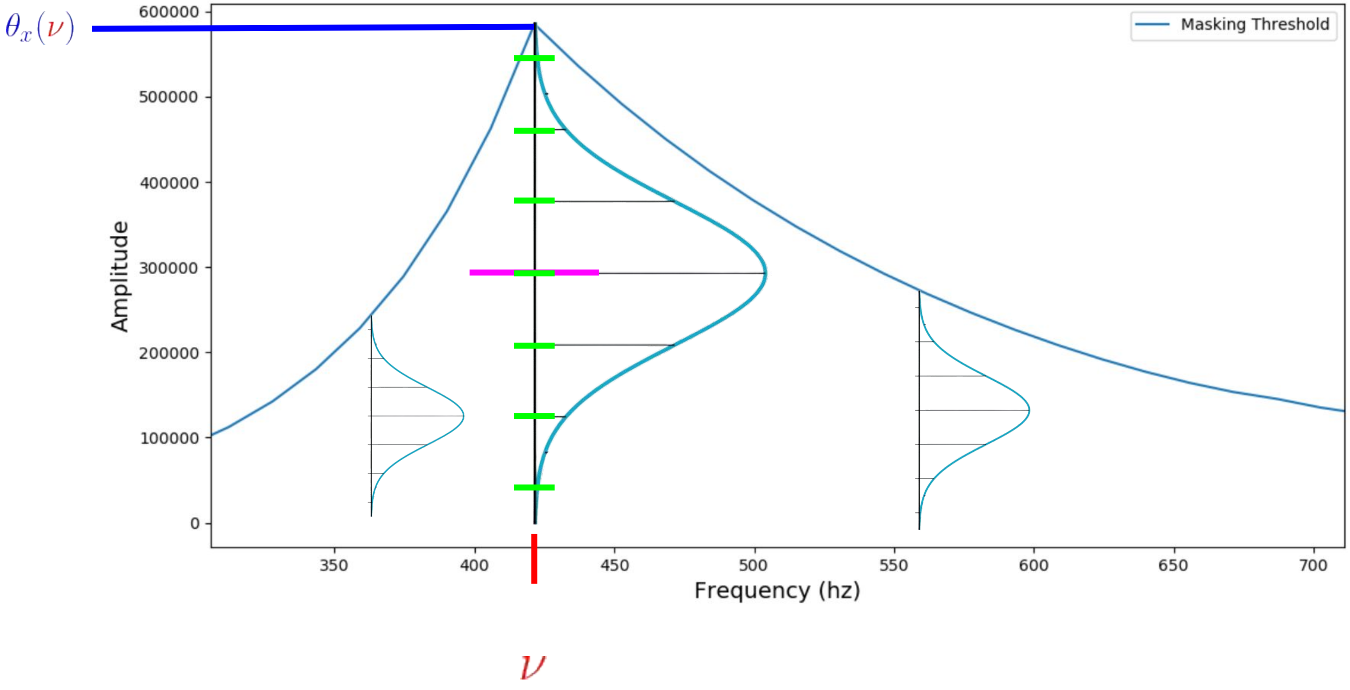
- Sample from a gaussian distribution with a mean and size proportional to the masking threshold

$$\mu = 3k \times \theta_x(\nu)$$

$$\sigma = k \times \theta_x(\nu)$$



Defensive Perturbation (Example)



$$k = \frac{1}{6}:$$

$$\mu = 3k \times \theta_x(\nu) = \frac{\theta_x(\nu)}{2}$$

$$\sigma = k \times \theta_x(\nu) = \frac{\theta_x(\nu)}{6}$$

Testing Metric

- Word Error Rate (WER): difference between intended transcription and the actual transcription wrt. # of substitutions (S), # of deletions (D), # of insertions (I), and # of words in the intended transcription (N):

$$WER = \frac{S + D + I}{N}$$

- WER = 0% → intended transcription matches actual transcription
- WER = 100% → ASR system returns no transcription ($D = N$)
- WER > 100% → intended transcription vastly different from actual transcription

Testing Metric

Example:

Intended: We wanted people to know that we've got something brand new, and essentially this product changes the way that people interact with technology.

Actual: We wanted people to know that how to me where I know and essentially this product changes the way people are rapid technology.

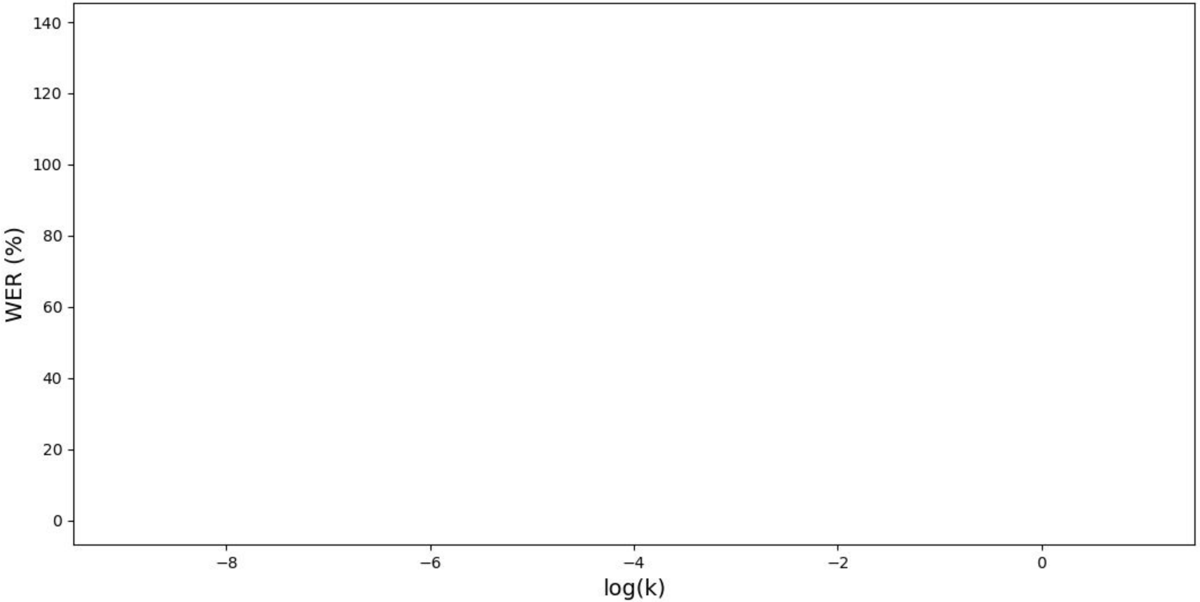
We wanted people to know that how to me where I know and essentially this product changes the way ~~that~~ people are rapid technology.

$$WER = \frac{7+1+1}{23} \approx 39\%$$

Credit: Allison Koo

Results

Attack: Qin et al. ‘19

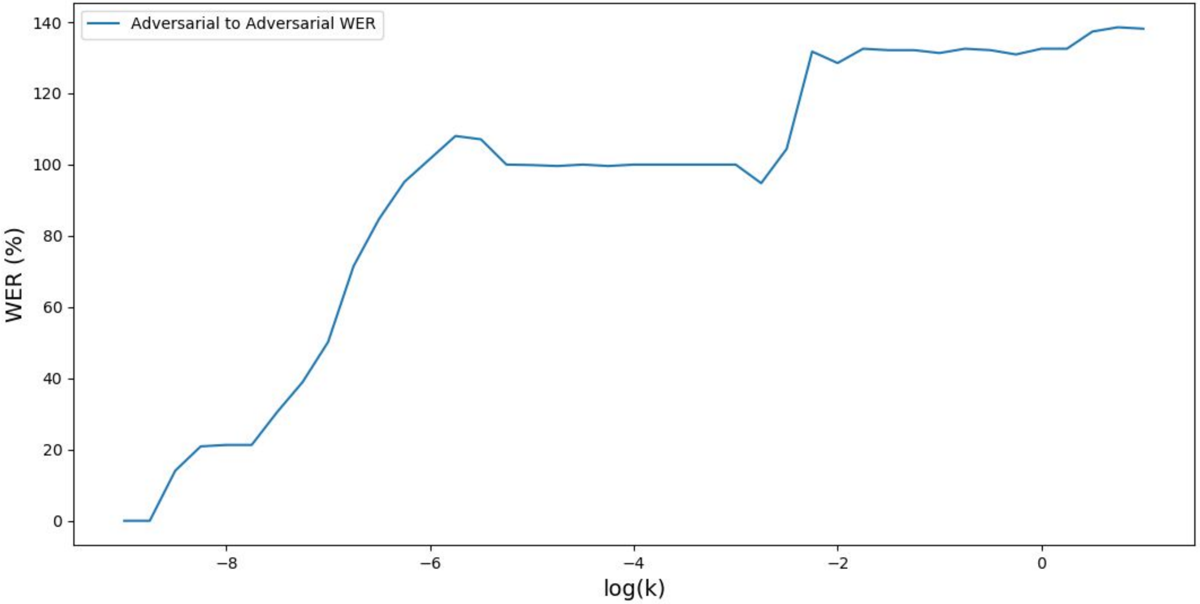


$$\mu = 3k \times \theta_x(\nu)$$

$$\sigma = k \times \theta_x(\nu)$$

Results

Attack: Qin et al. ‘19



$$\mu = 3k \times \theta_x(\nu)$$

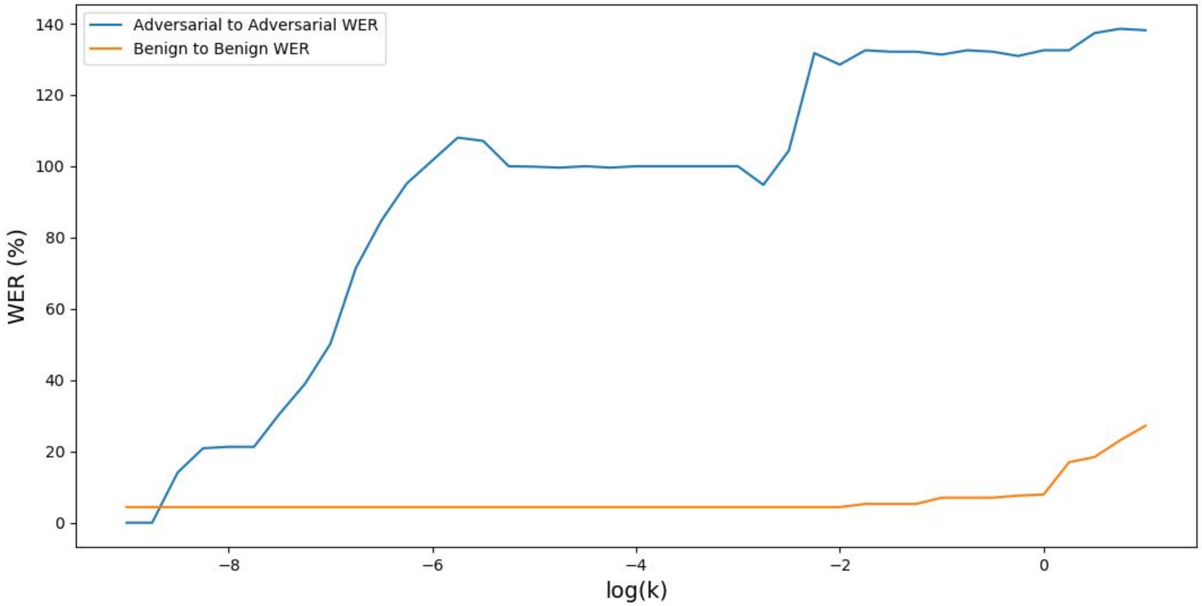
$$\sigma = k \times \theta_x(\nu)$$

We want:

HIGH Adversarial to Adversarial WER

Results

Attack: Qin et al. ‘19



$$\mu = 3k \times \theta_x(\nu)$$

$$\sigma = k \times \theta_x(\nu)$$

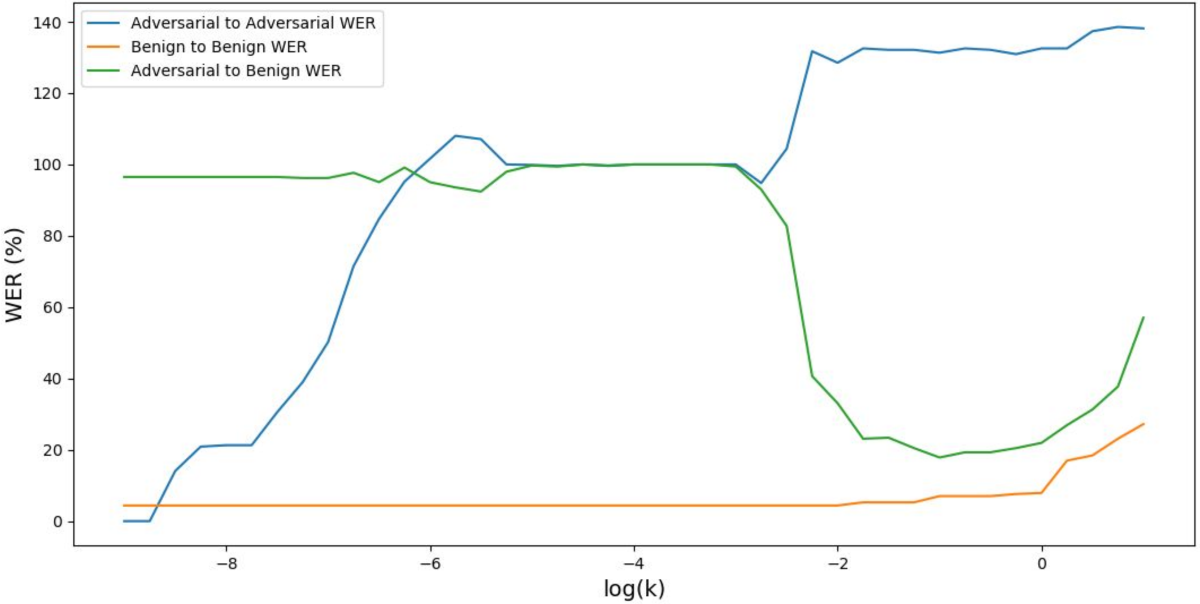
We want:

HIGH Adversarial to Adversarial WER

LOW Benign to Benign WER

Results

Attack: Qin et al. ‘19



We want:

HIGH Adversarial to Adversarial WER

LOW Benign to Benign WER

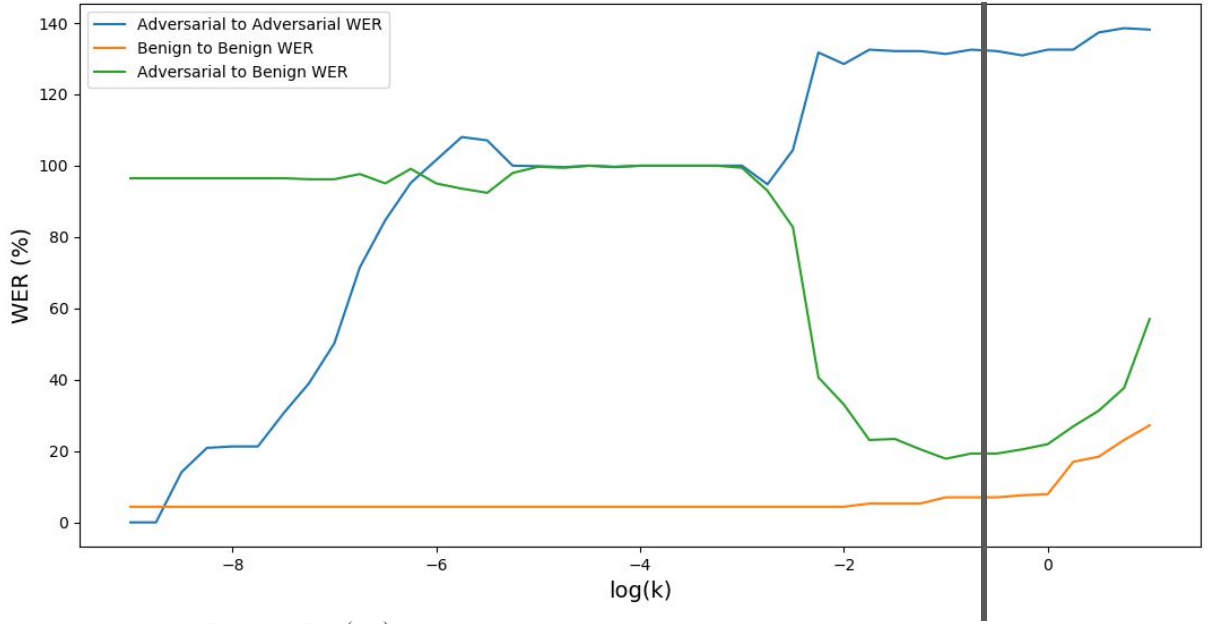
LOW Adversarial to Benign WER

$$\mu = 3k \times \theta_x(\nu)$$

$$\sigma = k \times \theta_x(\nu)$$

Results

Attack: Qin et al. ‘19



$$\mu = 3k \times \theta_x(\nu)$$

$$\sigma = k \times \theta_x(\nu)$$

Theoretically Optimal k (⅙)

We want:

HIGH Adversarial to Adversarial WER

LOW Benign to Benign WER

LOW Adversarial to Benign WER

Future Work

- Train speech recognition classifier with noisy data to achieve increased accuracy (motivated by adversarial training)
- Explore how the findings of this work can be applied to vision (i.e. can we create stronger adversarial examples by considering contrast and shading to hide adversarial perturbation)

Acknowledgements

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

Questions?