WaveGuard: Understanding and Mitigating Audio Adversarial Attacks
USENIX Security Symposium 2021

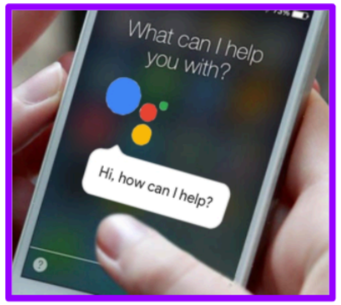
*Shehzeen Hussain, *Paarth Neekhara
Shlomo Dubnov, Julian McAuley, Farinaz Koushanfar

** Equal contribution*

University of California, San Diego

Reliability of Speech Recognition Systems

Model assurance/reliability is critical for DNN based Automatic Speech Recognition (ASR) systems

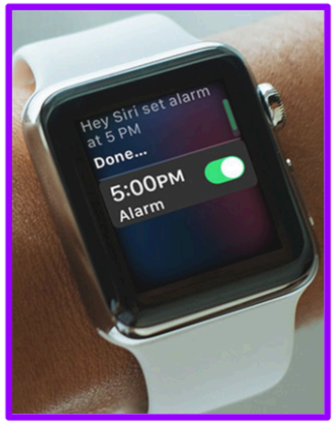


Smart Phones

Home System

AI HAS ENABLED A NEW LEVEL OF OPERATIONAL PRODUCTIVITY

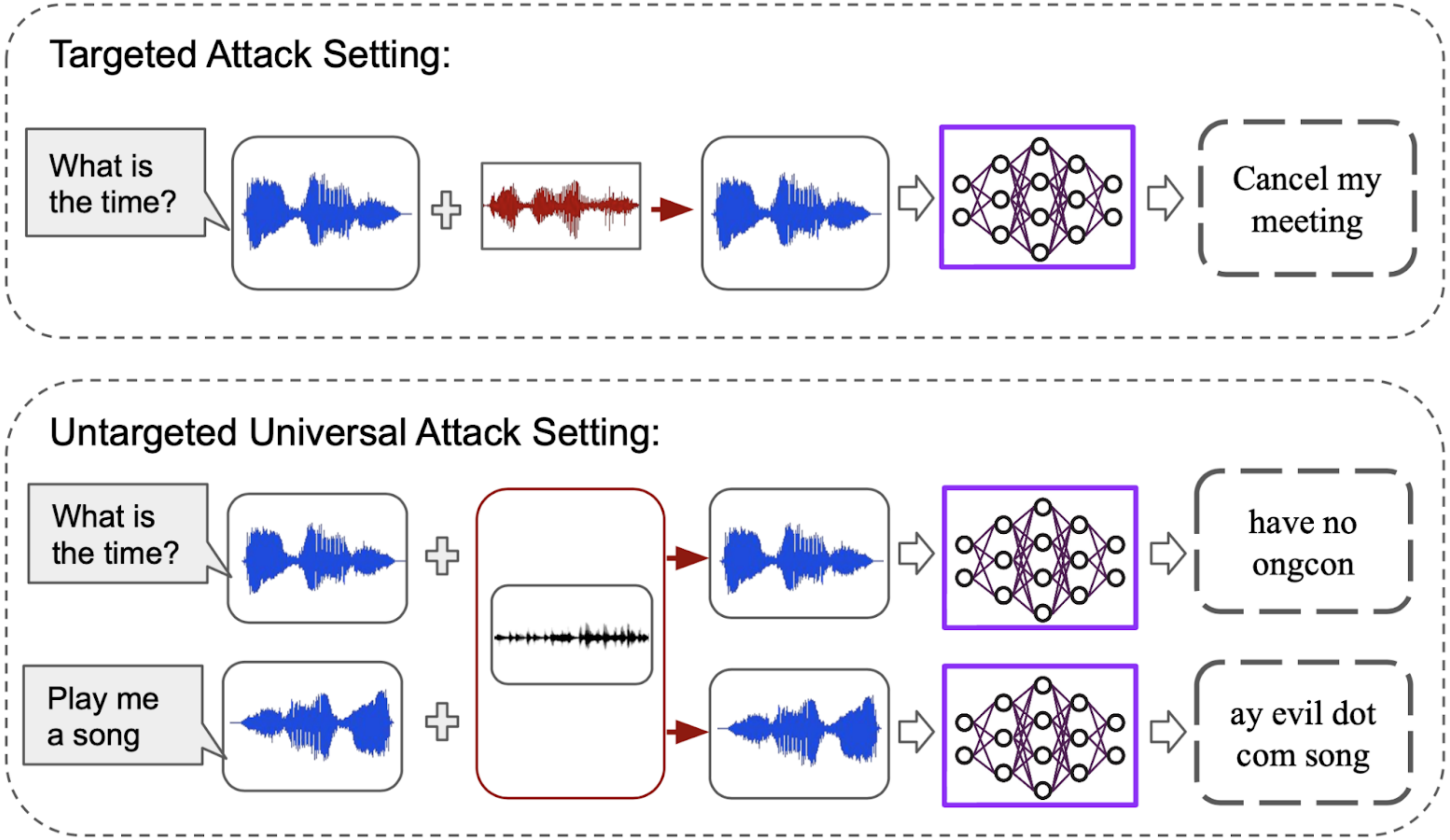
Navigation



Smart Watch

Adversarial attacks on Speech Recognition Systems

UC San Diego
JACOBS SCHOOL OF ENGINEERING



Common Attack Goal: cause mis-transcription of the given speech signal while keeping the perturbation amount imperceptible.

Vulnerability of DNN based ASR Systems

- Similar to the image domain adversarial examples also exist in the audio domain
- Adversarial perturbation when added to the signal causes the model to transcribe it into something malicious

Generating adversarial examples

$$Minimize\ L(X')$$

$$where\ L(X') = d(X, X') + g(X')$$

$$where\ g(X') = CE(F(X'), T)\ (\ for\ targeted\ attacks\ )$$

$$or\ g(X') = -CE(F(X'), C(X'))\ (\ for\ untargeted\ attacks\ )$$

$$X' = X - \epsilon sign(\nabla_X L(X))$$

Insights from Image Domain

Past works have characterized defenses against adversarial attacks in the image domain

The network predictions for adversarial examples are often unstable and small changes in adversarial inputs can cause significant changes in network predictions.

Adversarial Image
Label: Parrot

Clean Image
Label: Person

Compression and Decompression of images using following techniques [1] can launder adversarial perturbations ineffective:

- **JPEG Compression**
- **Image Quilting**
- **Bit-depth reduction**

[1]: Guo et. al 2018: Countering Adversarial Images using Input Transformations

Bypassing transformation based defenses

- What if the adversary aware of the defense being present?
  - The adversary can modify the optimization objective.

$$L = CE(F(g(x')), T) + CE(F(x')), T)$$

- Non-differentiable input transformation functions don't offer security.
- Gradient can be estimated for a transformation g(x) or we can use a straight-through estimator

$$\nabla_x f(g(x))|_{x=\hat{x}} \approx \nabla_x f(x)|_{x=g(\hat{x})}$$

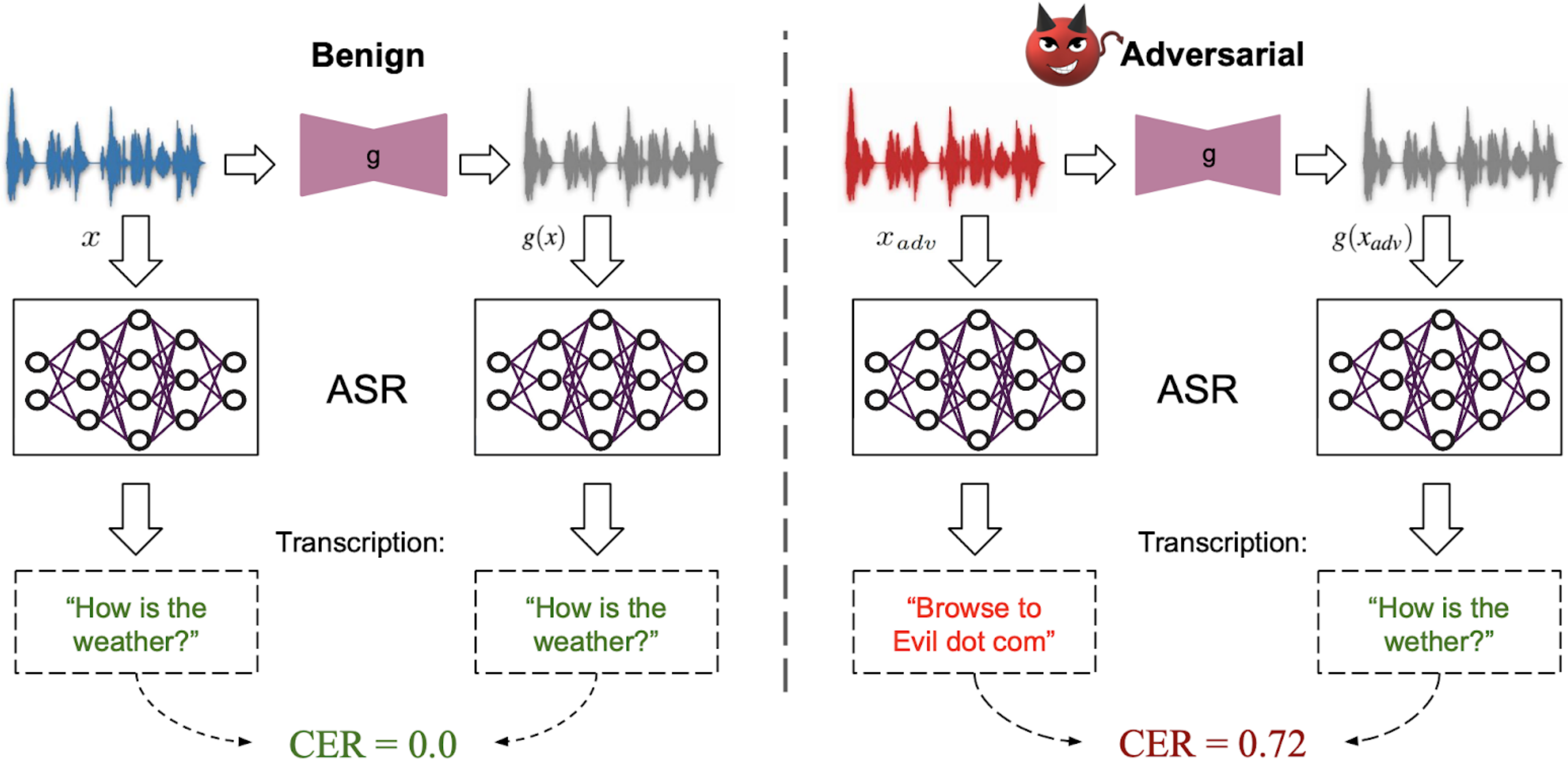
Athalye et al., 2018: Obfuscated Gradients Give a False Sense of Security

Open Questions for ASR Defense

- Can similar input transformation as image domain be applied in the audio domain as a defense?
- Can we recover original transcription of the adversarial audio?
- How effective are the transformations as a defense in the adaptive attack setting?

WaveGuard Framework

WaveGuard Defense Framework



WaveGuard Defense Framework

For a given audio transformation function g, an input audio x is classified as adversarial if there is significant difference between the transcriptions $C(x)$ and $C(g(x))$:

$$CER(C(x), C(g(x))) > t$$

Therefore, we label an example as adversarial or benign based on the Character Error Rate (CER) between text transcription of original audio x and transformed audio g(x)

WaveGuard Input Transformation Choices

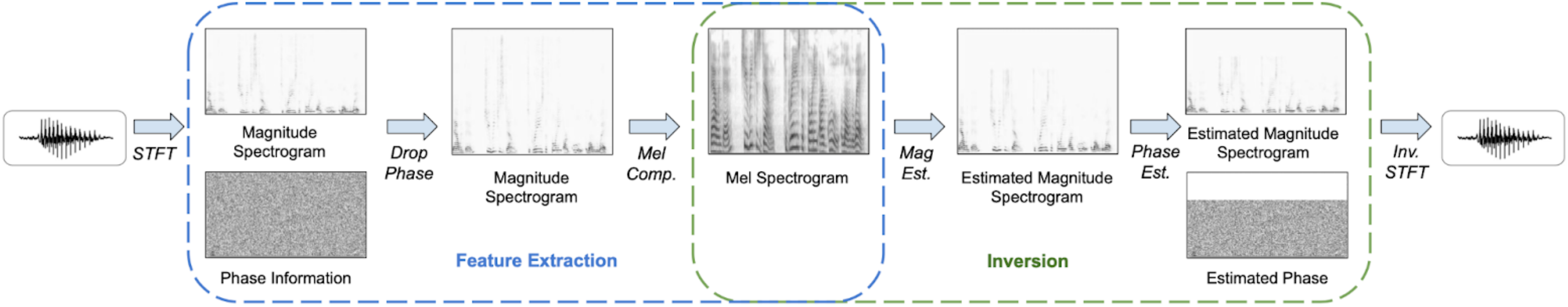
Naive audio transforms

1. Quantization – Dequantization
   - Quantize audio samples to n bits, re-estimate samples from quantized bits.
2. Downsampling - Upsampling
   - Resample audio from higher sampling rate to lower sampling rate (e.g. 16kHz to 8kHz and upsample back to 8kHz)
3. Filtering

Perceptually informed speech representations

4. **Mel Spectrogram Extraction - Inversion**
5. **Linear Predictive Coding (LPC)**

Mel Spectrogram Extraction - Inversion



Step 1: Feature Extraction

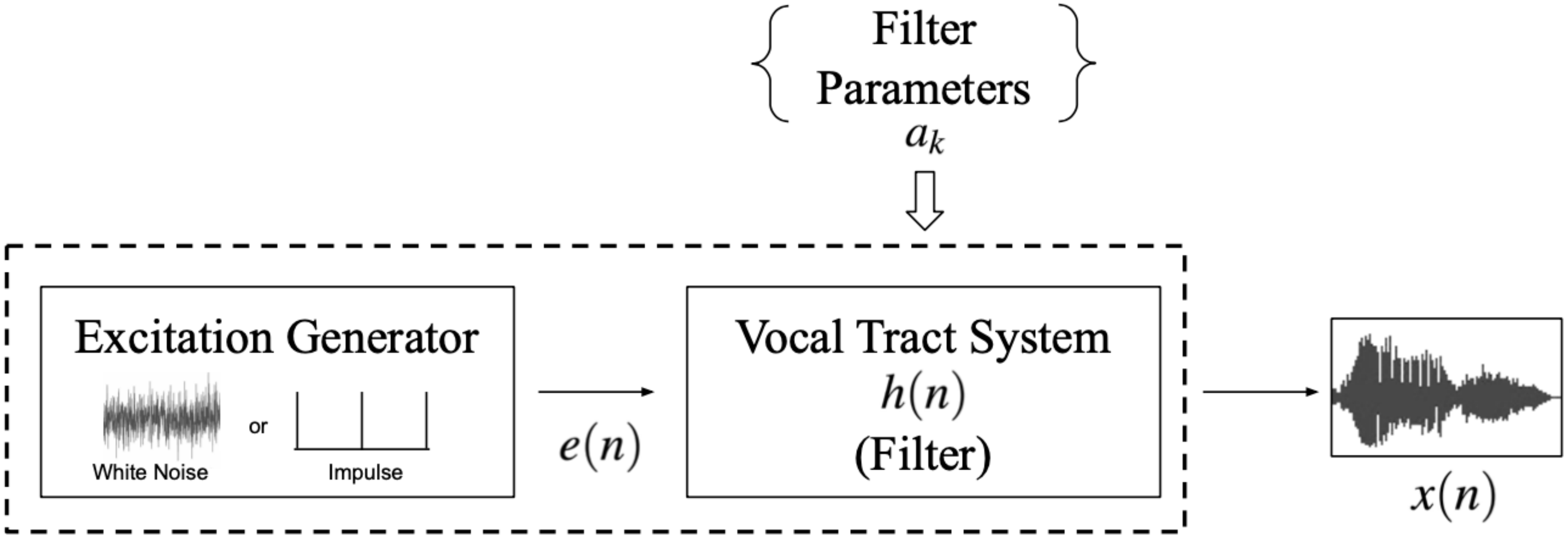
- Perform STFT, discard phase information, retain magnitude spectrogram
- Compress magnitude spectrogram to a Mel spectrogram.

Step 2: Inversion : Estimating the audio waveform

- First estimate the magnitude spectrogram from the Mel spectrogram
- Next, estimate the phase information
- Perform inverse STFT to obtain waveform

LPC – Linear Predictive Coding

- Linear Predictive coding models the human vocal tract system.
- **LPC analysis:** Estimate each sample in a waveform as a linear combination of previous n samples.
- **LPC compression:** Window the input signal, perform LPC analysis in windows of the signals, retain only the linear modelling coefficients and an excitation signal to recover the original signal.



Attacks and Models Investigated

- We investigate effectiveness of our defense against **four** adversarial audio attacks across **two** automatic speech recognition models:

Attack	Paper	Victim ASR
Carlini (Targeted)	Audio adversarial examples: Targeted attacks on speech-to-text, S&P 2018	Mozilla DeepSpeech
Universal (Untargeted)	Universal adversarial perturbations for speech recognition systems. INTERSPEECH 2019	Mozilla DeepSpeech
Qin I & R (Targeted)	Imperceptible, robust, and targeted adversarial examples for automatic speech recognition, ICML 2019	Google Lingvo

Evaluation against different attacks (Non-adaptive)

- Evaluation done on 100 adversarial examples generated using each of the following attacks:
  Carlini – Targeted, Qin-I – Targeted, Qin-R – Targeted, Universal – Untargeted

- Can reliably detect adversarial inputs with all input transformations under non-adaptive attack setting

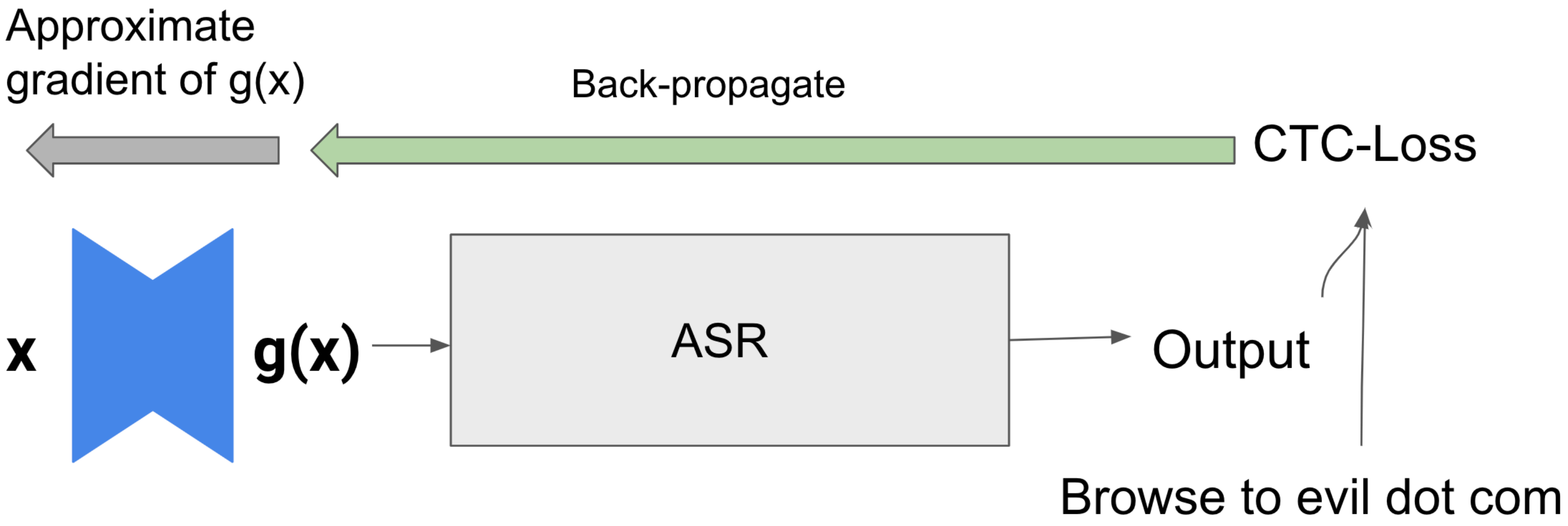
		AUC Score				*Detection Accuracy*			
Defense	Hyper-params	Carlini	Universal	Qin-I	Qin-R	Carlini	Universal	Qin-I	Qin-R
Downsampling - Upsampling	6000 kHz	1.00	0.91	1.00	1.00	100%	88%	100%	100%
Quantization - Dequantization	6 bits	0.99	0.92	1.00	0.93	98.5%	88%	99%	95%
Filtering	(Section 4.3)	1.00	0.92	1.00	1.00	99.5%	86%	100%	100%
Mel Extraction - Inversion	80 Mel-bins	1.00	0.97	1.00	1.00	100%	92%	100%	100%
LPC	LPC order 20	1.00	0.91	1.00	1.00	100%	83%	100%	100%

Adaptive Attack Setting

- **Threat Model:** Adversary has complete access to the defense and victim model.
- **Goal:** Targeted attack - Transcribe both undefended and defended audio to the target phrase.
  That is:

$$minimize: \; dB_x(\delta) + c_1 \cdot \ell(x+\delta, t) + c_2 \cdot \ell(g(x+\delta), t)$$

- **Handle non-differentiable input transformations with obfuscated gradients attack:**



Results – Adaptive Attack

- We used a differentiable implementation of all transformation functions for the backward pass.
- AUC score of less than 0.5 indicates the defense is successfully broken

	Distortion metrics			*Detection Scores*	
Defense	ϵ_∞	$\lvert\delta\rvert_\infty$	$dB_x(\delta)$	AUC	Acc.
None	500	81	-45.3	-	-
Downsampling - Upsampling	500	**342**	-32.7	0.31	50.0%
Quantization - Dequantization	500	**215**	-36.7	0.11	50.0%
Filtering	500	**92**	-44.1	0.45	50.0%
Mel Extraction - Inversion	500	500	-29.4	0.97	95.5%
LPC	500	500	-29.4	0.94	86.0%
Mel Extraction - Inversion	1000	1000	-23.5	0.92	84.0%
LPC	1000	1000	-23.5	0.77	72.5%
Mel Extraction - Inversion	4000	**2461**	-15.1	0.48	50.0%
LPC	4000	**2167**	-16.7	0.21	50.0%

Naive Input transformations cannot detect adversarial audio under adaptive attack

Perceptually informed transformations: LPC and Mel Extraction-Inversion Defense cannot be bypassed without highly ***audible*** adversarial noise (dB> -20)

WaveGuard Contributions

Formal defense framework

WaveGuard defense framework achieves state-of-the-art performance for detecting audio adversarial samples to defend ASR.

Low Computational Overhead

WaveGuard utilizes audio transformation functions to detect adversarial data which is computationally inexpensive.

Robust to Adaptive Attacks

First ASR defense to be evaluated thoroughly against various Adaptive Adversaries.

Technology transfer

Adversarial Defense can be used directly with any ASR model, without the need for retraining.

Thank You!

Code: `https://github.com/waveguard/waveguard_defense`