INTERSPEECH 2020

Detecting Audio Attacks on ASR Systems with Dropout Uncertainty

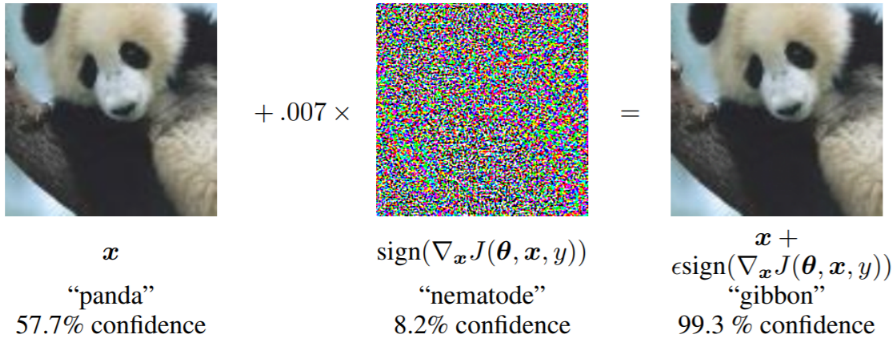
Tejas Jayashankar[1,2,3], Jonathan Le Roux[1], Pierre Moulin[1,2]

[1] Mitsubishi Electric Research Laboratories (MERL)
[2] University of Illinois at Urbana-Champaign (UIUC)
[3] Massachusetts Institute of Technology (MIT)

MITSUBISHI ELECTRIC RESEARCH LABORATORIES (MERL)
Cambridge, Massachusetts, USA
http://www.merl.com

Adversarial Attacks

- Subtly modify a signal such that a system misclassifies it or generates a malicious output
- **Targeted adversarial attack:** classify the signal into a target class or generate a malicious target output
- Two categories of attacks:
  - **Black box:** the adversary is unaware of the internal working/parameters of the system
  - **White box:** the adversary is informed of the internal working/parameters of the system
- Attacks usually created by adding optimized perturbations to an input signal
- Finely-tuned differences accumulate within the network to result in a malicious output



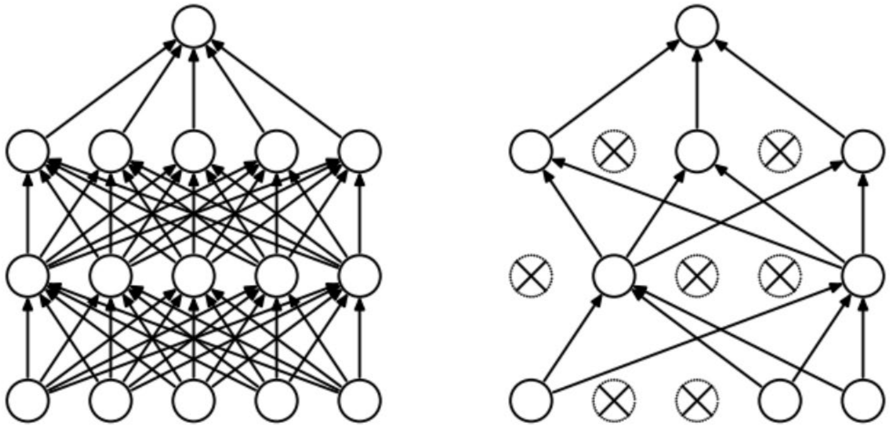
[Goodfellow '14]

Automatic Speech Recognition (ASR)

- Transcribes audio waveforms to text
- Focus on end-to-end ASR: use a neural network to map input features into a sequence of words or characters
- Input features are typically log-mel spectrogram or MFCC features
- Systems are differentiable and can be taken advantage of
- Popular architectures are CTC, CTC-Attention, RNN-T and the Transformer

Dropout

- Regularizer that prevents overfitting [Srivastava '14]
- Allows the neural network to learn multiple different internal realizations for an input-output pair



(a) Standard Neural Net (b) After applying dropout.

- Adversaries often know underlying architecture
- **Idea:** disarm attack by perturbing architecture via a random process

Applications & Motivation

- Voice commands can be modulated on ultrasonic carriers [Guoming '17]
- Can embed a voice command or message in any audio waveform
  - An innocent looking audio file might secretly contain malicious information
- Recently targeted adversarial attacks on ASR systems [Carlini '18, Qin '19]
- Why study adversarial machine learning for ASR?
  - Adversarial training (more robust loss functions)
  - Forgery detection
  - Secure ASR systems

Carlini & Wagner (CW) attack

- Proposed by Carlini and Wagner [Carlini ‘18]
- Input waveform x
- Perturbed waveform $x' = x + \delta$ sounds like input waveform
- Perturbation δ is optimized to make waveform transcribe as target transcription t
  - **Target transcription is “okay Google unlock phone and delete files” in all experiments**

- **Optimization problem:**

$$\min_{\delta} \ell(x+\delta, t)\, s.t.\, dB(\delta) \leq dB(x) - \tau$$

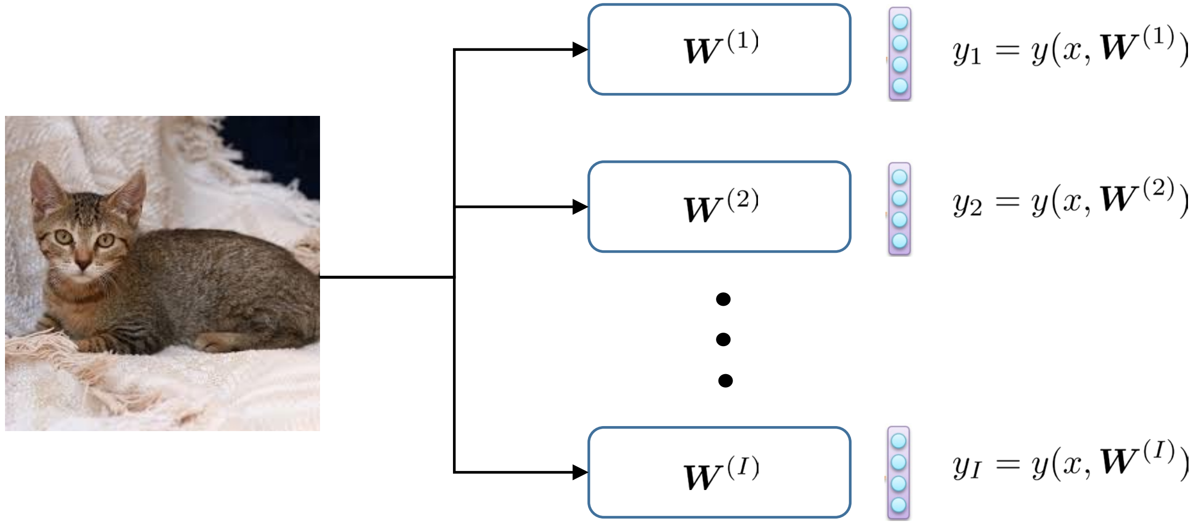
$$dB(x) = 20 \max_{i} \log(|x_i|)$$

- $\ell(\cdot)$ = CTC loss

“Only a minority of literature is written this way” “Okay Google unlock phone and delete files”

Dropout as a Defense

- Feinman *et al.* [Feinman '17] - dropout can be used as an uncertainty estimator in neural networks for image classification

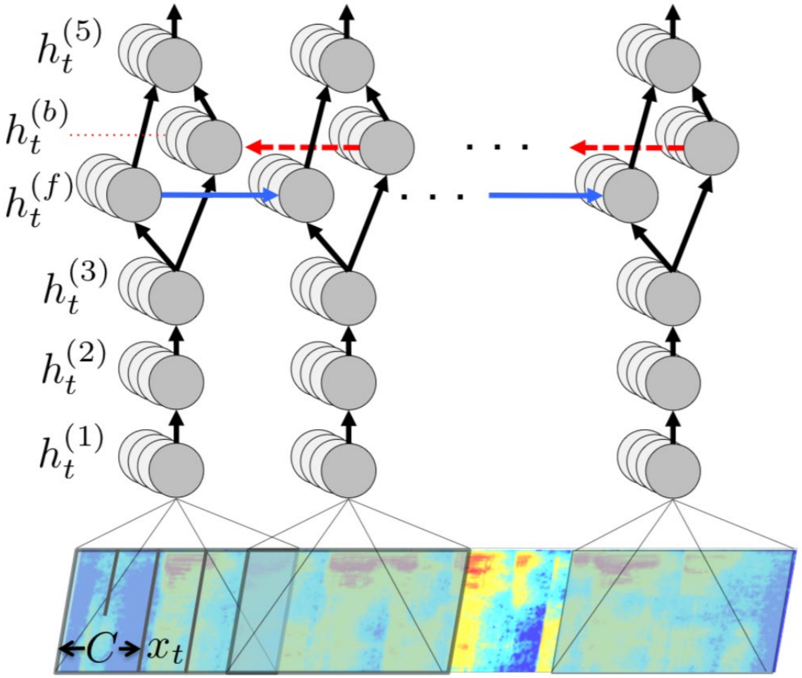


Input Image $\quad$ I Realizations of Dropout $\quad$ Output Class Probability Vectors

- Uncertainty of the network w.r.t. the input x is $U(x) = \frac{1}{I}\sum_{i=1}^{I} \|y_i - \hat{y}\|^2$
  - Average variation between each realization and a reference realization
- Train a threshold classifier on these uncertainties
- **Reasoning:**
  - Legit samples will not show much variation in network output for different dropout realizations
  - Adversarial samples show larger variation and hence the uncertainty is higher

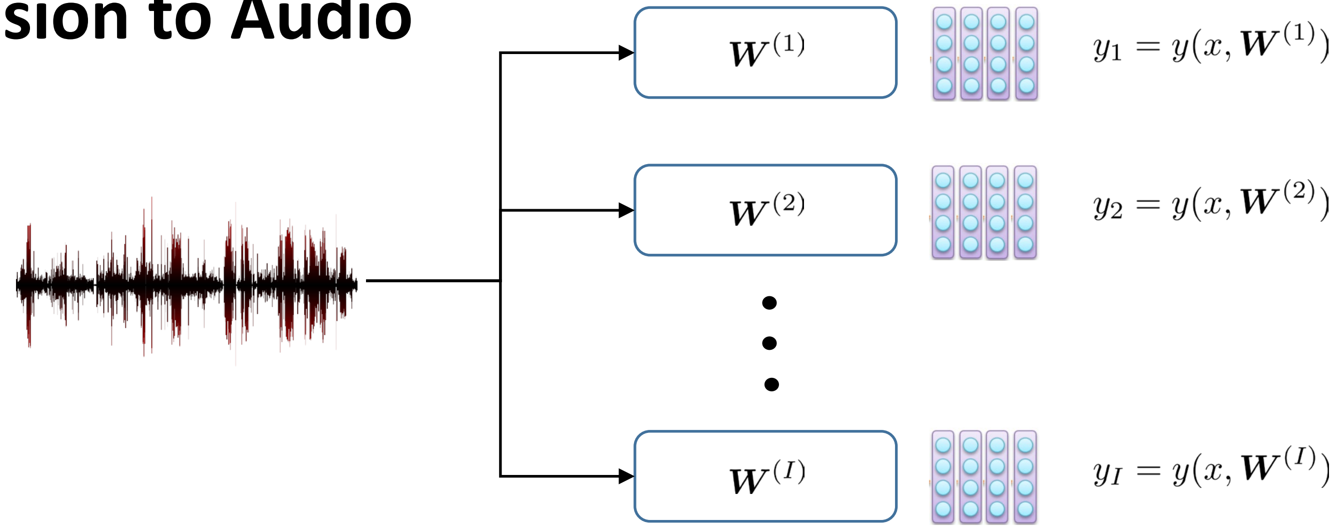
Mozilla DeepSpeech ASR Engine

- Specs: Bi-directional LSTM, trained with dropout rate of 0.05 in all layers except LSTM [Hannun ‘14]
- Operates on log-mel spectrogram



- Trained with CTC (Connectionist Temporal Classification) loss [Hannun ‘17]
  - Aligns the output sequence with the ground truth sequence to calculate the loss

Extension to Audio

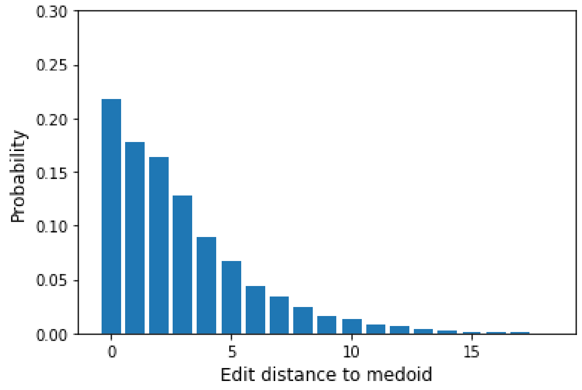


Input Audio Waveform

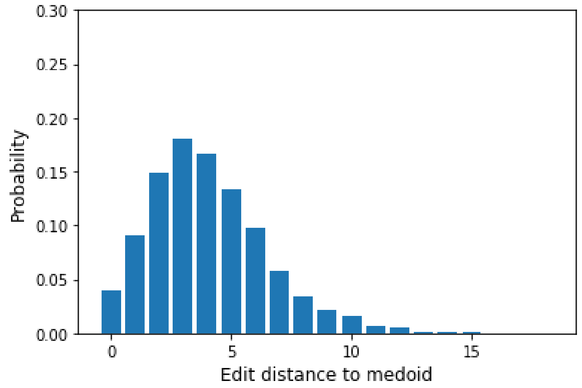
I Realizations of Dropout

CTC Probabilities/Character-level transcriptions

Legit Audio Samples



CW Adversarial Samples, p=0.05



- Construct an uncertainty distribution: $\mathbb{P}(z) = \sum_i \mathbb{1}_{\{d(\hat{y}, y_i) = z\}},\ z \in \mathbb{R}^+$
- Based on the output we use to calculate uncertainty, we have different distance metrics:
  - **CTC probabilities:** L2/Frobenius norm
  - **Character-level transcriptions:** Damerau-Levenshtein distance (Edit distance)
- Can use multiple features from this distribution
  - E.g., the image case used the second moment of the distribution
- For character level transcription compute **medoid transcription:** $\hat{y} = \arg\min_{y \in \{y_1, \ldots, y_I\}} \sum_i d(y, y_i)$
  - Medoid is the element of a set which is the closest to all other elements in the set
- Train a classifier on features extracted from the uncertainty distribution to classify a sample as adversarial or not

Feinman-Like Defense

1. Obtain I = 50 output realizations of the input audio using dropout
2. Here each realization is the output CTC probability tensor of size <num windows> x <alphabet size>
3. Obtain the average CTC probability tensor $\hat{y}$
4. Compute the uncertainty distribution of the input audio waveform and calculate its various moments
5. Denote this distribution as $\mathbb{P}_x^{\mathrm{prob}}$

Our Character-based Defense

1. Obtain I = 50 output realizations of the input audio using dropout
2. Here each realization is an output transcription
3. Obtain the medoid transcription
4. Compute the uncertainty distribution of the input audio waveform and calculate its various moments
5. Denote this distribution as $\mathbb{P}_x^{\mathrm{char}}$

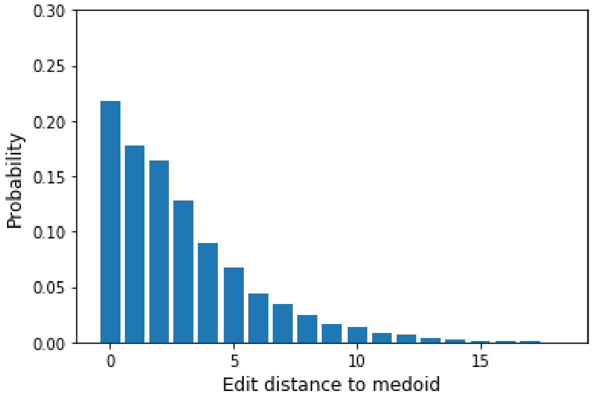
Dropout Robust Attack

- Create attacks robust to default dropout rate of 0.05 used in training the ASR system
- Optimization problem:

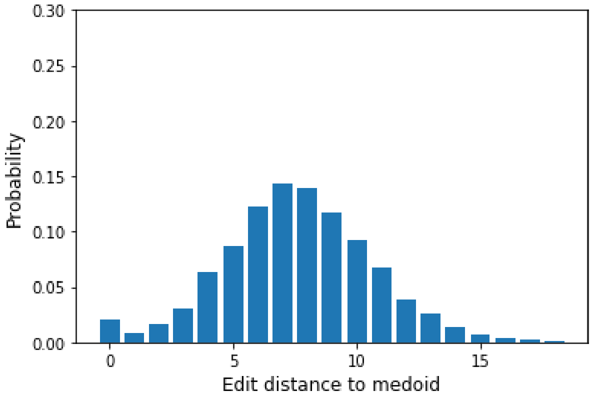
$$\min_{\delta} \ell(x+\delta,t) + \beta \ell_{p_{DR}}(x+\delta,t) \; s.t. \; dB(\delta) \leq dB(x) - \tau$$

- The existing defense using 0.05 won't work
  - **Modify the defense to use 0.1 dropout at inference**
- Experiments show that **successful** attacks cannot be created if the dropout rate used for creating the attack is more than 0.05
  - Likely due to the fact that the native DeepSpeech engine uses dropout of 0.05

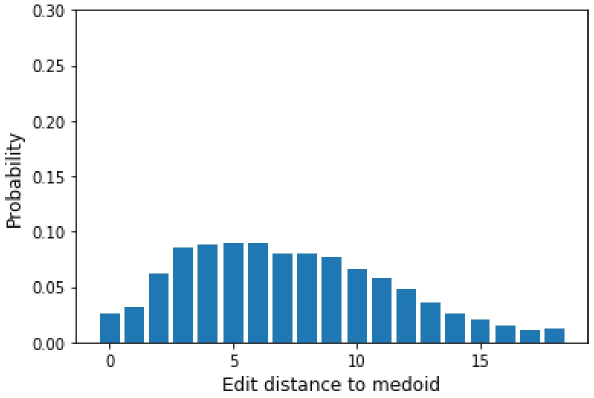
Legit Audio Samples



CW Adversarial Samples



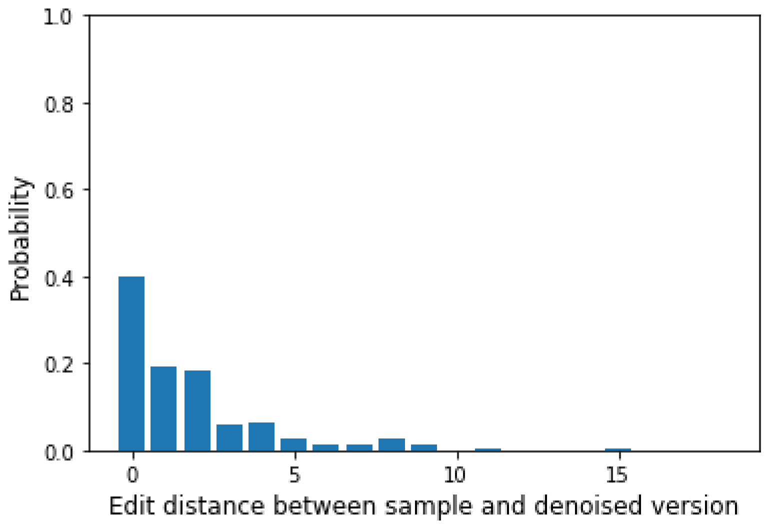
DR Adversarial Samples



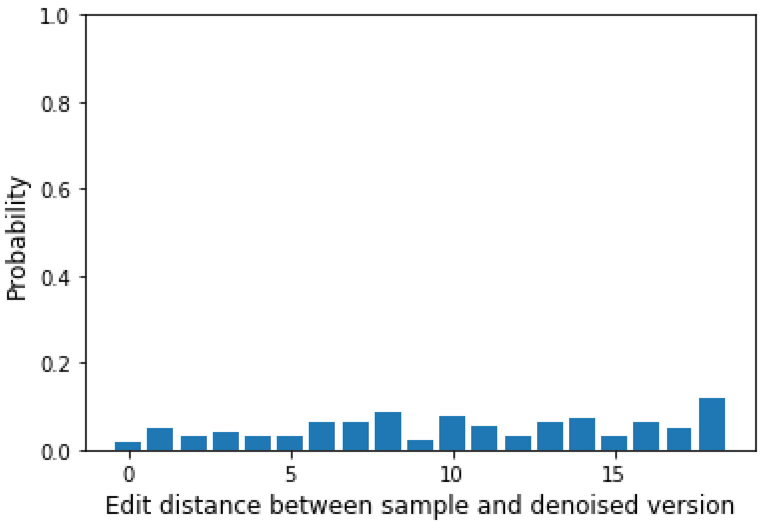
Noise Reduction Robust (NRR) attack

- Perturbation can be partially/completely removed by spectral subtraction or logmmse algorithm
- Denoised CW adversarial samples do not transcribe as target transcription

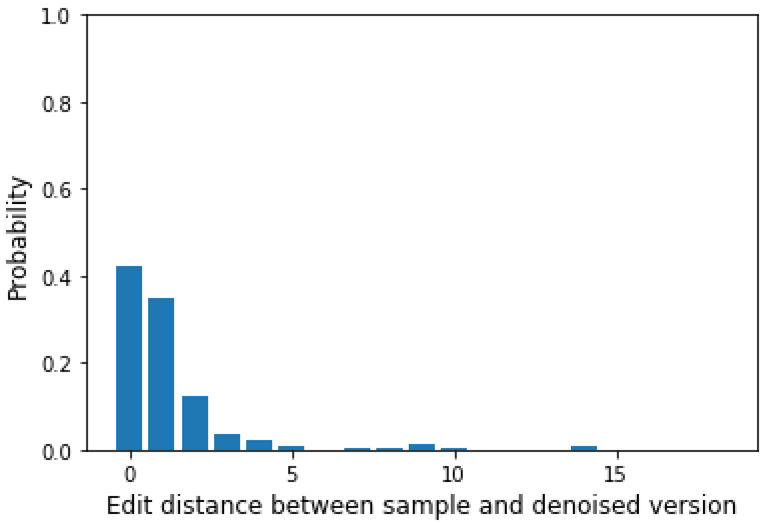
Legit Audio Samples



CW Adversarial Samples



Noise Robust Adversarial Samples



- Backpropagate through the spectral subtraction algorithm
  - Implemented a spectral subtraction algorithm in Tensorflow which can be appended to the computation graph

Imperceptible Audio (IA) attack

- Proposed by Qin *et al.* [Qin '19] for the attention-based Kaldi ASR
- Uses frequency masking
  - Louder frequencies mask out surrounding sounds at lower frequencies
- Power spectral density of the perturbation enforced to fall below the masking threshold of the audio waveform
- Two stage optimization: first generate vanilla CW attack and then optimize perturbation to render imperceptible
- **Optimization problem:**

$$\min_{\delta} \ell(x+\delta, t) + \alpha \sum_{k=0}^{\frac{N}{2}} \max\{p_{\delta}(k) - \theta_k(x), 0\} \; s.t. \; dB(\delta) \leq dB(x) - \tau$$

- Re-implemented attack for the CTC-based Mozilla DeepSpeech ASR system

Urban Sound (US) attack

- Applied the vanilla CW attack to the Urban Sound dataset
- Dataset consists of street, construction and automobile noises
- The aim of the experiment is two-fold:
  - Can the vanilla CW attack be extended to general sounds?
  - Can our defense detect attacks concealed in such recordings?

Experiments

- Extract moments from uncertainty distributions obtained from Feinman-like and our character-based defense
- For Feinman-like defense, train:
  - **DS:** a decision stump on the mean of the distribution (most direct extension of work in [Feinman ‘17])
  - **SVM-4:** a linear SVM on the first four moments of the distribution
  - **Decision Tree:** a decision tree on the first four moments of the distribution
- For our character-based defense, train:
  - **DS:** a decision stump on the mean of the distribution (most direct extension of work in [Feinman ‘17]
  - **SVM-4:** a linear SVM on the first four moments of the distribution
  - **SVM-F:** a linear SVM on the complete discrete distribution
- Compute area under ROC curve for the various classifiers
- All experiments except Urban Sound performed on randomly chosen 500 samples from the Mozilla Common Voice Dataset
- Use 70-30 train-test split in all experiments except for Urban Sound
- All attacks targeted to transcribe as “okay google unlock phone and delete files”
- Compare against recent entropy-based classifier [Däubener ‘20]
  - DS on the entropy of the uncertainty distributions

Results – Detection Accuracy

Table 1: *Detection accuracy [%] on various attacks for the different classifiers.* p *denotes the defense dropout rate.*

		$p = 0.05$	$p = 0.1$				
		CW	CW	DR	NRR	IA	US
$\mathbb{P}_x^{\text{prob}}$	DS	71.7	83.3	82.5	75.5	91.0	90.4
	SVM-4	66.7	80.8	68.0	53.3	68.0	64.4
	DecTree	65.0	80.8	72.0	70.0	73.3	91.8
$\mathbb{P}_x^{\text{char}}$	DS	72.3	**96.5**	81.0	81.0	**92.0**	79.0
	SVM-4	76.7	**96.5**	**88.5**	**88.5**	**92.0**	**93.9**
	SVM-F	74.0	85.8	86.5	87.5	88.3	83.0
Entropy	DS	**80.0**	90.5	88.0	84.2	78.3	79.5

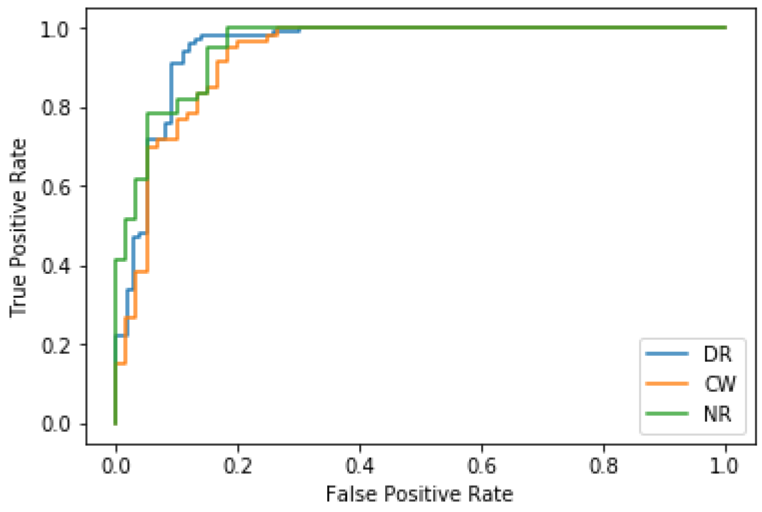
The character-based SVM-4 results in the best detection accuracy across all attacks

Results – Area under ROC curve

Table 2: *AUC score on various attacks for the different classifiers. p denotes the defense dropout rate.*

		$p = 0.05$	$p = 0.1$				
		CW	CW	DR	NRR	IA	US
$\mathbb{P}_x^{\text{prob}}$	DS	0.72	0.85	0.83	0.84	0.82	0.91
	SVM-4	0.84	0.91	0.88	0.89	0.90	**0.98**
	DecTree	0.72	0.85	0.83	0.84	0.82	0.91
$\mathbb{P}_x^{\text{char}}$	DS	0.72	0.82	0.81	0.82	0.73	0.86
	SVM-4	**0.88**	**0.92**	**0.95**	**0.93**	**0.95**	0.94
	SVM-F	0.75	0.91	0.92	0.93	0.94	0.74
Entropy	DS	0.75	0.81	0.88	0.82	0.92	0.74

ROC Curves for Character-based Classifiers



AUC = Area Under ROC Curve

Conclusion

- We have extended the vanilla CW attack to create adversarial attacks that are
  - Dropout robust
  - Denoising robust
  - Capable of being embedded in urban sounds
- We can use simple classifiers to detect an adversarial attack
- Specifically, an SVM-4 trained on the moments of the character-sequence-level distribution results in the best detection accuracy
- Developed a defense that can detect various attacks by leveraging dropout, including attacks crafted using frequency masking (imperceptible audio attack)

Bibliography

[1] Carlini, Nicholas & Wagner, David. (2018). Audio Adversarial Examples: Targeted Attacks on Speech-to-Text. 1-7. 10.1109/SPW.2018.00009.

[2] Däubener, Sina & Schönherr, Lea & Fischer, Asja & Kolossa, Dorothea. (2020). Detecting Adversarial Examples for Speech Recognition via Uncertainty Quantification.

[3] Feinman, Reuben & Curtin, Ryan & Shintre, Saurabh & Gardner, Andrew. (2017). Detecting Adversarial Samples from Artifacts.

[4] Goodfellow, Ian & Shlens, Jonathon & Szegedy, Christian. (2014). Explaining and Harnessing Adversarial Examples. arXiv 1412.6572.

[5] Guoming, Zhang & Yan, Chen & Ji, Xiaoyu & Zhang, Taimin & Zhang, Tianchen & Xu, Wenyuan. (2017). DolphinAtack: Inaudible Voice Commands. 10.1145/3133956.3134052.

[6] Hannun, Awni & Case, Carl & Casper, Jared & Catanzaro, Bryan & Diamos, Greg & Elsen, Erich & Prenger, Ryan & Satheesh, Sanjeev & Sengupta, Shubho & Coates, Adam & Ng, Andrew. (2014). DeepSpeech: Scaling up end-to-end speech recognition.

[7] Hannun, "Sequence Modeling with CTC", Distill, 2017.

[8] Srivastava, Nitish & Hinton, Geoffrey & Krizhevsky, Alex & Sutskever, Ilya & Salakhutdinov, Ruslan. (2014). Dropout: A Simple Way to Prevent Neural Networks from Overfitting. Journal of Machine Learning Research. 15. 1929-1958.

[9] Qin, Yao & Carlini, Nicholas & Goodfellow, Ian & Cottrell, Garrison & Raffel, Colin. (2019). Imperceptible, Robust, and Targeted Adversarial Examples for Automatic Speech Recognition.

MITSUBISHI ELECTRIC

Changes for the Better