Interspeech 2022
18-22 September 2022, Incheon, Korea

Defense against Adversarial Attacks on Hybrid Speech Recognition using Joint Adversarial Fine-tuning with Denoiser

Sonal Joshi, Saurabh Kataria , Yiwen Shao*,*
Piotr Żelasko, Jesús Villalba, Sanjeev Khudanpur, Najim Dehak

Center for Language and Speech Processing, Johns Hopkins University, Baltimore, MD
{sjoshi12,skatari1,yshao18,pzelask2,jvillal7,khudanpur,ndehak3}@jhu.edu

Abstract

Adversarial attacks are a threat to automatic speech recognition (ASR) systems, and it becomes imperative to propose defenses to protect them. In this paper, we perform experiments to show that K2 conformer hybrid ASR is strongly affected by white-box adversarial attacks. We propose three defenses–denoiser pre-processor, adversarially fine-tuning ASR model, and adversarially fine-tuning joint model of ASR and denoiser. Our evaluation shows denoiser pre-processor (trained on offline adversarial examples) fails to defend against adaptive white-box attacks. However, adversarially fine-tuning the denoiser using a tandem model of denoiser and ASR offers more robustness. We evaluate two variants of this defense–one updating parameters of both models and the second keeping ASR frozen. The joint model offers a mean absolute decrease of 19.3% ground truth (GT) WER with reference to baseline against fast gradient sign method (FGSM) attacks with different L_∞ norms. The joint model with frozen ASR parameters gives the best defense against projected gradient descent (PGD) with 7 iterations, yielding a mean absolute increase of 22.3% GT WER with reference to baseline; and against PGD with 500 iterations, yielding a mean absolute decrease of 45.08% GT WER and an increase of 68.05% adversarial target WER.

Index Terms: speech recognition, adversarial attacks and defenses, adversarial training, robustness, speech enhancement

1. Introduction

In today's world, voice-based smart assistants are ubiquitous–be it using phones, dedicated home assistants like Alexa, Google Home, Apple Pod, or as call center agents [1]. A core technology behind these assistants is automatic speech recognition (ASR) whose goal is to transcribe speech to text. Recent work [2, 3, 4] has shown that ASR systems are vulnerable to adversarial inputs, which contain specially crafted mostly human inaudible noise. Considering the threat of these adversarial attacks, it becomes of foremost importance to propose defensive countermeasures to protect ASRs. These countermeasures broadly fall into three categories–pre-processing, stochastic, and adversarial training. Pre-processing defenses intend to remove the adversarial noise from the signal before passing it into the machine learning system [5, 6, 7]. Stochastic defenses introduce randomness into the model. Thus, the model used to craft the adversarial sample differs slightly (but stochastically) from the model used to evaluate the sample, reducing attack effectiveness. Randomized smoothing is the most common stochastic defense [8]. Finally, adversarial training tries to make the model inherently robust by training on dynamically generated adversarial examples [9, 10]. The major contributions of this work are highlighted below:

- We evaluate the robustness of a strong baseline K2 Conformer hybrid ASR model against white-box attacks, i.e. when the adversary is aware of parameters of the model. To the best of our knowledge, this is the first work that fully utilizes a differentiable hybrid ASR model to study adversarial robustness.
- We propose four defenses–a pre-processing time-domain denoiser, adversarial fine-tuning and two variants of joint adversarial training of a pre-processing denoiser with ASR model.
- We evaluate these defenses against strong adaptive white-box attacks, i.e. when the adversary is aware of parameters of defense model along with those of ASR.

The rest of the paper is as follows. In Section 2, we introduce adversarial attacks on ASR. In Section 3, we describe defenses; followed by experiments and results in Section 4 and 5 respectively.

2. Adversarial attacks on ASR

An ASR system can be considered as a function $\hat{\mathbf{y}} = f(\mathbf{x}, \theta)$, which predicts a sequence of words $\hat{\mathbf{y}}$ given an audio waveform $\mathbf{x}$. f is a statistical model described by a set of parameters θ. ASR systems are known to be vulnerable to adversarial attacks [2]. The attacker adds a small perturbation to the benign signal to alter the prediction of the system. Depending on the goals of the attacker, we find different attack modalities. When an adversarial example fools the ASR into predicting a particular target phrase that an adversary desires, it is called a *targeted* adversarial attack. *Untargeted* attacks, by comparison, simply induce transcription errors and are not of as much concern in the ASR context [3]. Table 1 explains the concept of targeted attack via an example. Suppose we have a speech signal $\mathbf{x}$. Without any attack, a human will transcribe $\mathbf{x}$ as **actual** = `This is the human ASR output`. Now using $\mathbf{x}$ as input to the ASR model, the output transcription by the ASR is denoted as **benign** = `Thus is the real ASR output`. Suppose, the adversary wants the ASR model to predict the target phrase denoted by **target** = `Transfer $1000 from my account`. To achieve this goal, he/she crafts an adversarial example $\mathbf{x}'$ such that when $\mathbf{x}'$ is given as input to the ASR, the model produces the output transcription **adversarial** = `Transfer a sand from my account`. One can find Word Error Rate (WER) between the pairs of the transcriptions as shown in Table 2. Let WER between reference (ref) and hypothesis (hyp) be denoted by $WER(ref, hyp)$. $Benign = WER(actual, benign)$ denotes the Benign

*equal contribution.

This work is supported by DARPA projects: GARD (www.darpa.mil/program/guaranteeing-ai-robustness-against-deception) and RED (www.darpa.mil/program/reverse-engineering-of-deceptions)

Copyright © 2022 ISCA

10.21437/Interspeech.2022-10977

Table 1: *Table describing types of transcripts with shorthand symbol used and example*

Description	Shorthand	Under attack?	Example
What is the actual (human transcribed) content of speech signal?	actual	✗	This is the human ASR output
What is the text that the ASR system predicts ?	benign	✗	Thus is the real ASR output
What is the text that the adversary wants to achieve?	target	✓	Transfer $1000 from my account
What is the text that is actually predicted by ASR after the attack?	adversarial	✓	Transfer is sand from my account

Table 2: *Table describing different Word Error Rate (WER) metrics used for evaluation of successful defense*

Description	WER type	Formula WER($< ref >, < hyp >$)	Defense success
Does the defense harm the un-attacked system?	Benign ground truth	Benign = WER(actual,benign)	↓
Did the attacker succeed in denial of service?	Adversarial ground truth	GT = WER(actual,adversarial)	↓
Did the adversary make the system recognize what he/she wants?	Adversarial target	TGT=WER(target,adversarial)	↑

ground truth WER, which measures the performance of the ASR system in non-attack conditions. A good ASR system, and hence defense, should have Benign WERs as low as possible, indicated by ↓ in the column *Defense success*. $GT = WER(actual, adversarial)$ is called the adversarial ground truth WER. GT is a performance indicator of how much the attacker succeeded in denial-of-service, i.e. introducing untargeted spelling errors. A higher value indicates as successful untargeted attack, while a low value is characteristic of a robust ASR. $TGT = WER(target, adversarial)$ denotes the adversarial target WER and is a performance indicator whether the adversary was successful in getting the ASR to predict the chosen target phrase. While the adversary wants TGT WER to as low as possible (meaning, target phrase to be recognized perfectly), an ideal defense will make it be as high as possible. This is indicated by ↑ in the column *Defense success*.

Attack Algorithms: An adversarial example is computed as $\mathbf{x}' = \mathbf{x} + \delta$ where is $\mathbf{x}$ is a benign signal and δ is a small adversarial perturbation. Many attack algorithms in the literature [4] propose different ways to compute δ. In this work, we consider FGSM [11] and Projected Gradient Descent (PGD) attacks [9]. For targeted attacks, PGD optimizes delta by gradient descent iterations that minimizes the ASR loss L between the target phrase selected by the attacker $\mathbf{y}^{\text{target}}$ and the adversarial transcript predicted by ASR. Thus, for iteration $i+1$,

$$\delta_{\mathbf{i+1}} = \text{clip}_{\varepsilon}(\delta_i - \alpha \text{sign}(\nabla L(f(\mathbf{x}, \theta), \mathbf{y}^{\text{target}})) \ . \quad (1)$$

Throughout this paper, PGD-i indicates the number of iterations used for PGD attack (e.g.: PGD-7 means 7 iterations). At every iteration, the clip function (projection) ensures that $\|\delta\|_\infty \leq \varepsilon$, keeping the attack imperceptible. We choose the learning rate α at every iteration is one fifth of the max-norm. FGSM is a single iteration version of PGD with step $\alpha = \varepsilon$. While attacking any system (with or without defense), we assume that it fully white-box and adaptive, meaning the adversary knows not only the speech recognition model parameters but also the defense. This is the worst-case scenario to evaluate robustness, a commonly expected norm for evaluation of defenses, as it exposes the system's weakest links [12].

3. Defenses

3.1. Randomized smoothing

Randomized smoothing is a stochastic defense that adds random, normally-distributed noise with standard deviation σ to the input. This additive normal noise tries to mask the gradients that are essential in computing adversarial examples. If σ is too high, the benign accuracy reduces and hence it is vital to find σ that offers robustness without reducing the accuracy. Previous work on speaker identification defenses [13] show that this defense can be easily combined with other defenses and may offer additional protection against high L_∞ attacks.

3.2. Adversarial fine-tuning of ASR model

This defense is a variant of adversarial training [9]. Instead of full adversarial training, which leads to convergence issues in ASR, we propose to bootstrap from a pre-trained ASR (which is trained using clean/benign examples as normally done) and then fine-tune using the model using PGD adversarial attacks. We call this model *ADV-FINETUNE-ASR*. For an ASR model denoted by $f(\cdot, \theta)$, where θ are the model parameters, adversarial training is done by minimizing the loss function given by

$$\theta^* = \arg\min_{\theta} \mathrm{E}_{(\mathbf{x},\mathbf{y})\sim\mathcal{D}} \left[J(f(\mathbf{x} + \delta^*, \theta), \mathbf{y}) \right] \ , \quad (2)$$

where J is the ASR loss function (lattice-free MMI in our case), $\mathcal{D}$ is the set of training audio-transcript pairs $(\mathbf{x}, \mathbf{y})$, and $\delta^* = \arg\max_{\delta, \|\delta\|_\infty \leq \varepsilon} J(f(\mathbf{x} + \delta, \theta), \mathbf{y})$ is adversarial perturbation optimized by PGD iterations.

3.3. Denoiser

The pre-processing denoiser defense maps adversarial signals to benign. The denoiser was trained using deep regression approach [14] in time-domain. Training objective function $\mathcal{L}_{\text{sup}}$ is Multi-Resolution Short-Time Fourier Transform (MRSTFT) auxiliary loss.

$$\mathcal{L}_{\text{sup}} = \mathbb{E}_{\mathbf{x},\mathbf{x}'\sim P_{\mathcal{B},\mathcal{A}}} [\sum_{m=1}^{M} \mathcal{L}_{\text{sup}}^{(m)}(\mathbf{x}, \mathbf{x}')] \ , \quad (3)$$

$$\mathcal{L}_{\text{sup}}^{(m)}(\mathbf{x}, \mathbf{x}') = \mathcal{L}_{\text{sc}}^{(m)}(\mathbf{x}, g(\mathbf{x}', \boldsymbol{\phi})) + \mathcal{L}_{\text{mag}}^{(m)}(\mathbf{x}, g(\mathbf{x}', \boldsymbol{\phi}))), \quad (4)$$

$$\mathcal{L}_{\text{sc}}^{(m)}(\mathbf{x}, \hat{\mathbf{x}}) = \frac{\| |\,\text{STFT}^{(m)}(\mathbf{x})| - |\,\text{STFT}^{(m)}(\hat{\mathbf{x}})| \|_F}{\| |\,\text{STFT}^{(m)}(\mathbf{x})| \|_F}, \quad (5)$$

$$\mathcal{L}_{\text{mag}}^{(m)}(\mathbf{x}, \hat{\mathbf{x}}) = \frac{1}{N} \| \log |\,\text{STFT}^{(m)}(\mathbf{x})| - \log |\,\text{STFT}^{(m)}(\hat{\mathbf{x}})| \|_1 \ ; \quad (6)$$

Table 3: *Ground Truth (GT) Word error rate (%) (↓) for K2 ASR systems under FGSM and PGD-7 attacks and defenses. RSσ stands for randomized smoothing with a σ parameter.*

System	**Benign**	**FGSM Attack**					**PGD-7 Attack**				
L_∞ (max-norm)		0.0001	0.001	0.01	0.1	0.2	0.0001	0.001	0.01	0.1	0.2
Baseline (full LibriSpeech test-clean)	4.34	5.12	7.13	10.62	70.27	90.80	6.29	22.38	63.50	95.24	97.68
Baseline (reduced LibriSpeech test-clean)	**4.53**	5.36	7.22	11.56	74.84	91.32	6.44	25.89	63.63	95.71	97.90
+ *RS0.001*	4.66	**4.88**	10.43	13.12	79.23	94.49	**4.83**	10.48	65.68	94.73	97.17
+ *RS0.001 + DENOISER*	4.95	4.97	5.95	21.65	81.72	98.78	4.97	5.90	40.61	92.64	95.76
ADV-FINETUNE-ASR + RS0.001	4.73	**4.88**	**5.70**	20.48	82.30	99.27	4.88	5.90	40.42	93.71	96.10
ADV-FINETUNE-JOINT + RS0.001	5.22	5.07	**5.70**	**6.00**	**21.40**	**55.58**	5.27	**5.22**	12.19	78.99	92.74
ADV-FINETUNE-JOINT-ASRfrozen + RS0.001	5.64	6.09	6.24	9.65	90.05	100.00	6.24	6.39	**10.29**	**65.29**	**89.81**

where $\mathbf{x}$ is a benign signal and $\mathbf{x}'$ is the corresponding adversarial signal, $\hat{\mathbf{x}} = g(\mathbf{x}', \boldsymbol{\phi})$ is the predicted benign and $\boldsymbol{\phi}$ are the parameters of the denoiser. $\mathcal{B}$ and $\mathcal{A}$ denote the benign and adversarial domains and $P_{\mathcal{B},\mathcal{A}}$ denotes their joint distribution, which is obtained from a dataset of offline attack samples. $\mathcal{L}_{\text{sup}}$ uses M different STFT with different frame-shift and frame-lengths, which are indexed by $m = 1 \dots M$. The number of time-frequency bins in the STFT are denoted by N, while $||\cdot||_F$ refers to the Frobenius matrix norm.

3.4. Adversarial fine-tuning of joint ASR and Denoiser model

The disadvantage of using denoiser as pre-processor is that, in a fully white-box scenario, the adversary can break the system by backpropogating through the combined denoiser+ASR network and computing adaptive adversarial attacks–i.e., attacks that adapt to the defense (albeit at the expense of higher computing cost). Therefore, to make the denoiser itself robust to adaptive white-box attacks, we propose to adversarially fine-tune the pre-trained denoiser in tandem with the ASR model using on-the-fly PGD attacks, trying to minimize the ASR cross-entropy. We call this model *ADV-FINETUNE-JOINT*. Similar to standard adversarial training in (2), the optimum parameters for denoiser and ASR models are given by

$$\theta^*, \phi^* = \arg\min_{\theta,\phi} \mathrm{E}_{(\mathbf{x},\mathbf{y})\sim\mathcal{D}} \left[J(f(g(\mathbf{x}+\delta^*, \phi), \theta), \mathbf{y}) \right] \ , \quad (7)$$

where $\delta^* = \arg\max_{\delta, \|\delta\|_\infty \leq \varepsilon} J(f(g(\mathbf{x}+\delta, \phi), \theta), \mathbf{y})$.

3.5. Adversarial fine-tuning of joint ASR and Denoiser model with ASR model frozen

It is known that adversarial fine-tuning may over-fit the ASR model to work well on adversarial examples, degrading benign performance. To avoid this, we tried another variant of the joint network by freezing the ASR model and just updating the denoiser parameters. We call this model *ADV-FINETUNE-JOINT-ASRfrozen*. We expect that adversarial training on the denoiser will be more efficient than doing it on the full ASR+denoiser network, as the denoiser network is much smaller than ASR and hence will require fewer epochs to train.

4. Experimental Setup

Dataset, Baseline, and Adversarial Attacks Our experimental setup was based on LibriSpeech dataset [15]. We used a Hybrid DNN-HMM ASR model implemented on the K2 framework[1] using PyTorch [16]. A Conformer [17] network was used to compute frame-level posteriors, which were used as input to the K2 WFST decoder. The Conformer consisted of 12-layers with dimension=256, heads=4, and feed-forward dimension=2048; and used 80 log-Mel-filterbank features. The whole pipeline is end-to-end differentiable to be able to compute adversarial examples in time domain. This model was trained on the full 960 hours LibriSpeech corpus for 20 epochs, and the parameters from the last 5 epochs were averaged to get the final model. We denote this model as the undefended *baseline*. To evaluate the robustness of the model, we applied targeted FGSM and PGD attacks with different strengths, i.e., max. L_∞ norm levels: {0.0001, 0.001, 0.01, 0.1, 0.2}. For PGD, we evaluated attacks with 7 and 500 iterations (PGD-7 and PGD-500). The target phrases used to craft the attacks were taken from the LibriSpeech training set. For each utterance, we chose a target phrase with length similar to the benign transcript. We evaluated the attacks on the first 100 examples from LibriSpeech test-clean. We chose to work on this reduced set because of three reasons. First, the computational cost for PGD iterations is too high for the full set, so in the literature, it is common to experiment on a limited number of utterances [2]. Second, for the baseline, the performance of the reduced set did not statistically change w.r.t. the full set (see Table 3). Lastly, this setup is the same as in previous work [18] and in DARPA-GARD evaluations. Another important thing to point-out is that we always evaluated using white-box adaptive attacks. That means that the adversary knows the defenses and their parameters, and it is able to back-propagate through them to create the adversarial examples. Many works in the adversarial literature do not consider adaptive attacks based their defense in obfuscating the gradients of the system, as evidenced in [19].

Table 4: *Word error rate (%) for K2 ASR system under PGD-500 attacks. [Ground-truth WER (GT) and target WER (TGT). RSσ stands for randomized smoothing with a σ parameter. Arrow ↓ indicates lower is better and ↑ indicates higher is better.*

System	**Benign**	**PGD-500 Attack**			
L_∞ (max-norm)		0.01		0.1	
	GT ↓	GT ↓	TGT ↑	GT ↓	TGT ↑
Baseline (reduced LibriSpeech test-clean)	**4.53**	97.81	40.71	100.59	16.47
Baseline + *RS0.001*	4.78	101.41	16.37	101.22	17.19
Baseline + *RS0.001 + Denoiser*	4.63	94.05	55.40	92.44	71.96
ADV-FINETUNE-ASR + RS0.001	4.97	68.84	91.07	101.95	10.03
ADV-FINETUNE-JOINT + RS0.001	5.17	62.36	97.94	96.25	48.92
ADV-FINETUNE-JOINT-ASRfrozen + RS0.001	5.70	**25.40**	**100.05**	**82.84**	**93.23**

[1] https://k2-fsa.github.io/k2/index.html

Attacks Dataset: We create a dataset by generating offline PGD adversarial samples against LibriSpeech train sets using $L_2 = \{0.2, 0.5, 1.5, 1.9\}$ and $L_\infty = \{0.001, 0.01, 0.1\}$ threat models with number of iterations $\{10, 20, 50, 100, 200\}$ sampled with more bias towards high norm and high iteration attacks. Generating offline adversarial examples has the advantage that we can use computationally expensive PGD attacks with large number of iterations without letting a model run for months on online attacks. The generated attacks were used to train the denoiser as described in Section 4. Then, the pre-trained denoiser could be finetuned in tandem with ASR model to increase its robustness.

Denoiser: After experimenting with a few denoiser architectures, we choose TasNet [20], a time-domain model for source separation and speech enhancement. It is an all-convolutional 1-D Convolutional Neural Network (CNN), which consists of *encoder*, *separator*, and *decoder*. The *encoder* and *decoder* are single convolutional layers, while the *separator* stage consists of multiple CNNs called *stacks*. The *stacks* output are combined to produce a mask, which is applied to the encodings and passed to the *decoder* stage. We used 128-dim encodings obtained with kernel-size=16 and stride=8. The separator used one stack with 16 layers, with kernel dilations increasing with a factor of 2 [20]. We trained the denoiser on our dataset of offline attacks, using the corresponding benign example as a clean target.

ADV-FINETUNE-ASR: We fine-tuned the baseline ASR model on on-the-fly PGD-7 attacks. The L_∞ for these attacks were randomly sampled from a log uniform distribution [0.0001,0.02]. The learning rate was 10x lower than the one used in the training phase.

ADV-FINETUNE-JOINT: Instead of adversarially fine-tuning just the ASR model, we jointly fine-tuned the tandem denoiser+ASR. We evaluated another variant where attacks were generated on the tandem denoiser+ASR, but we only updated the denoiser weights while the ASR model is frozen. We call this defense as ***ADV-FINETUNE-JOINT-ASRfrozen***.

5. Results

We evaluate the robustness of baseline ASR and all proposed defenses against FGSM and PGD-7 attacks (Table 3) and PGD-500 attacks (Table 4). For FGSM and PGD-7, the WER w.r.t. the target phrase was always greater than 90%. In other words, the adversary is not able to make the ASR to recognize the malicious target phrase. Hence, we omitted TGT WER in Table 3 and included only WER w.r.t ground truth phrase (GT WER). We analyze the undefended baseline for full and reduced Librispeech test-clean set. We observe that the results for both sets are similar. Therefore, henceforward, all models were evaluated on the reduced set to alleviate the large cost of generating attacks (and other reasons mentioned in Section 4). Next, we evaluated the different defenses. We observe that the denoiser defense trained on offline attacks performed better than randomized smoothing for most L_∞ values and on par with the adversarially trained ASR. Both defenses jointly adv. fine-tuning denoiser and ASR performed significantly better than the offline denoise and the adv. fine-tuned ASR. *ADV-FINETUNE-JOINT* yielded the largest robustness against FGSM attacks with mean absolute decrease of 19.3% GT WER w.r.t. the baseline. *ADV-FINETUNE-JOINT-ASRfrozen* was the best for PGD-7 with a mean absolute decrease in GT WER of 22.3% w.r.t. the baseline. We can observe that for both attacks, the best defense kept the system robust up to $L_\infty \leq 0.01$.

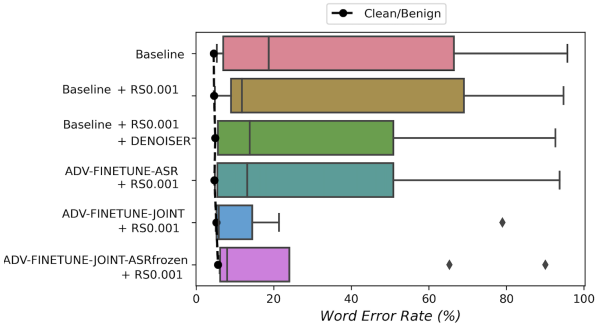

Figure 1: *Summary of systems for all attack settings using box plot. We exclude FGSM-$L_\infty = 0.2$ and PGD-$L_\infty = 0.2$) because they are perceptible attacks with SNR. The dotted line indicates the clean/benign performance of the system*

When increasing the number of PGD iterations to 500 (Table 4), the WER w.r.t. the target phrase (TGT) decreases, meaning that the attacker starts being successful in making the system to recognize a particular malicious phrase. However, to obtain a usable TGT WER of 16%, it needs to increase L_∞ to 0.1, which is a very perceptible attack. Here, the goal of the defense is increasing TGT WER while reducing GT WER. Again, the best system by far was *ADV-FINETUNE-JOINT-ASRfrozen*. The mean absolute decrease in GT WER was 45.08% GT WER and increase in TGT WER was 68.05%. Although there was a slight increase (1.17%) in the benign WER with reference to the baseline model, the gain in adversarial robustness overshadowed it. The proposed method significantly outperformed the baseline defenses in the literature, i.e., randomized smoothing and ASR adversarial training. Unfortunately, for large $L_\infty = 0.1$, the defenses could not reduce GT WER much. However, note that these are very perceptible attacks and even using Gaussian noise (non-adversarial) of that level would significantly damage the ASR system.

6. Conclusion

We evaluated the robustness of K2 Hybrid ASR model along with four defenses–pre-processing time-domain denoiser defense, adversarial fine-tuning of ASR model and two variants of joint adversarial training of pre-processing denoiser and ASR model. We evaluated these defenses against strong adaptive white-box attacks, i.e. when the adversary is aware of parameters of defense model along with those of ASR. To understand the big picture of the defenses, we convert the Table 3 to box plot shown in Figure 1. Please note that we exclude FGSM-$L_\infty = 0.2$ and PGD-$L_\infty = 0.2$) as they are perceptible. The dotted blue line indicates the clean/benign performance of the individual systems. The best defense should have WER distribution concentrated around the blue line. The results show that *ADV-FINETUNE-JOINT* is the best defense, followed closely by *ADV-FINETUNE-JOINT-ASRfrozen*. On the other hand, for PGD-500 attacks, *ADV-FINETUNE-JOINT-ASRfrozen* performs the best, consistently yielding low GT WER and high TGT WER. This is at the cost of slight degradation in benign WER (<1.2%), however this is expected for adversarial defenses. In the future, we would like to bridge this gap further.

7. References

[1] S. M. S. Talebi, A. A. Sani, S. Saroiu, and A. Wolman, "Megamind: a platform for security & privacy extensions for voice assistants," in *Proceedings of the 19th Annual International Conference on Mobile Systems, Applications, and Services*, 2021, pp. 109–121.

[2] N. Carlini and D. Wagner, "Audio adversarial examples: Targeted attacks on speech-to-text," in *SPW 2018*, 2018.

[3] Y. Qin, N. Carlini, I. Goodfellow, G. Cottrell, and C. Raffel, "Imperceptible, Robust, and targeted adversarial examples for automatic speech recognition," in *International Conference on Machine Learning (ICML)*, 2019, pp. 471–492.

[4] D. Wang, R. Wang, L. Dong, D. Yan, X. Zhang, and Y. Gong, "Adversarial examples attack and countermeasure for speech recognition system: A survey," in *International Conference on Security and Privacy in Digital Economy*, 2020, pp. 443–468.

[5] X. Yuan, Y. Chen, Y. Zhao, Y. Long, X. Liu, K. Chen, S. Zhang, H. Huang, X. Wang, and C. A. Gunter, "Commandersong: A systematic approach for practical adversarial voice recognition," in *USENIX Security Symposium*, 2018, pp. 49–64.

[6] Z. Yang, B. Li, P.-Y. Chen, and D. Song, "Characterizing audio adversarial examples using temporal dependency," in *International Conference on Learning Representations (ICLR)*, 2019.

[7] M. Esmaeilpour, P. Cardinal, and A. L. Koerich, "Class-conditional defense gan against end-to-end speech attacks," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 2565–2569.

[8] J. Cohen, E. Rosenfeld, and Z. Kolter, "Certified adversarial robustness via randomized smoothing," in *International Conference on Machine Learning*, 2019, pp. 1310–1320.

[9] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," in *International Conference on Learning Representations*, 2018.

[10] A. Jati, C.-C. Hsu, M. Pal, R. Peri, W. AbdAlmageed, and S. Narayanan, "Adversarial attack and defense strategies for deep speaker recognition systems," *Computer Speech & Language*, vol. 68, p. 101199, 2021.

[11] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and Harnessing Adversarial Examples," in *ICLR 2015*, dec 2015.

[12] F. Tramer, N. Carlini, W. Brendel, and A. Madry, "On adaptive attacks to adversarial example defenses," *Advances in Neural Information Processing Systems*, vol. 33, 2020.

[13] S. Joshi, J. Villalba, P. Żelasko, L. Moro-Velázquez, and N. Dehak, "Study of pre-processing defenses against adversarial attacks on state-of-the-art speaker recognition systems," *IEEE Transactions on Information Forensics and Security*, vol. 16, pp. 4811–4826, 2021.

[14] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "A regression approach to speech enhancement based on deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 1, pp. 7–19, 2014.

[15] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an ASR corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.

[16] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "Pytorch: An imperative style, high-performance deep learning library," *Advances in Neural Information Processing Systems*, vol. 32, pp. 8026–8037, 2019.

[17] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu *et al.*, "Conformer: Convolution-augmented transformer for speech recognition," *arXiv preprint arXiv:2005.08100*, 2020.

[18] P. Żelasko, S. Joshi, Y. Shao, J. Villalba, J. Trmal, N. Dehak, and S. Khudanpur, "Adversarial attacks and defenses for speech recognition systems," *arXiv preprint arXiv:2103.17122*, 2021.

[19] A. Athalye, N. Carlini, and D. Wagner, "Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples," in *International Conference on Machine Learning*, 2018, pp. 274–283.

[20] Y. Luo and N. Mesgarani, "Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 27, no. 8, pp. 1256–1266, 2019.