Countermeasures against Adversarial Attacks on Speech Technologies

Sonal Joshi

Department of Electrical and Computer Engineering, Center for Language and Speech Processing, Johns Hopkins University, Baltimore, MD, United States of America

sjoshi12@jhu.edu

Abstract

With proliferation of speech technologies in everyday life, their security becomes of utmost importance. Recently, researchers have found stealthy and almost human-imperceptible attacks on speech technologies (speaker recognition and speech recognition) called "adversarial examples". The goal of this work is to build countermeasures against these attacks. In particular, we explore two research directions:

1. **Defense** - Robustify speech models to withstand against adversarial attacks
2. **Classify and Detect** - Categorize into attack algorithm and detect the presence of unseen attack

Index Terms: speaker recognition, speech recognition, countermeasures, adversarial attacks, security, defense, speech enhancement

1. Introduction

"*Hey Alexa*", "*Hi Siri*" and "*OK Google*", will soon be one of the first words kids learn to speak. Building on the foundations of speech technologies, the current generation of voice assistants have proliferated everyday life, be it through digital assistants, customer care centers, or banking voice authentication. The two major speech technologies that power the widespread applications are Automatic Speech Recognition (ASR) and Speaker Recognition (SR). Given a speech recording, ASR converts the speech recording to text and SR predicts the identify of the speaker. Quite recently, a new genre of attacks, termed *adversarial*–first introduced by the computer vision community [1]–has been shown to be fatal for both ASR [2, 3, 4] as well as SR systems [5, 6, 7] (See Figure 1). Adversarial examples are malicious human-imperceptible inputs that are purposely designed to fool a system. They are generated by adding a small yet carefully computed perturbation to the signal and have shown to have disastrous consequences, including but not limited to identify theft, financial losses, and digital forensic failures.

For *state-of-the-art* SR system, accuracy drops from 100% to 0-2% [8] using attack algorithms like fast gradient sign method (FGSM), basic iterative method (BIM), projected gradient descent (PGD) etc. For ASR, Carlini&Wagner show that an attacker can make the system transcribe *any* chosen phrase with 100% attack success rate [9]. Depending on the information available to the adversary, adversarial examples can be classified into White-box and Black-box attacks. White-box attack means the adversary has the knowledge of the model weights and exact parameters, while Black-box attack means that this knowledge is not available. This work aims to make speech technologies *safe and secure* by developing countermeasures against adversarial attacks. Countermeasures can be build using two major research directions, which are described in the following sections:

2. Research Direction I: Defense [1]

2.1. Research Question

How can we make existing speech models robust against adversarial attacks?

2.2. Motivation

Since adversarial examples cause performance deterioration, can we develop models that are inherently robust to them?

2.3. Key Challenges

- There are multiple adversarial attack algorithms, in both white-box and black-box settings. In addition, every attack has multiple hyperparameters including but not limited to attack strength, number of optimization iterations, etc. Building one model that successfully counteracts against all attacks is a difficult task.
- Re-training the system to defend against one particular attack algorithm is computationally very expensive.
- Using a denoising network to remove adversarial perturbation does not work for white-box attacks as the attacker can back-propogate through the denoiser model in addition to speech.
- If the adversary gets to know the defense parameters and attacks the full system including the defense (*adaptive attack*) will the defense hold?

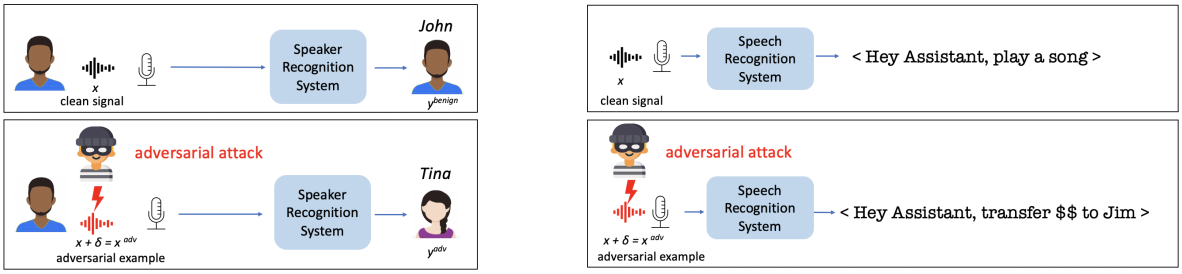
2.4. Major contributions

- We proposed pre-processing defenses that do not require any adversarial examples for training for SR Systems [7] and ASR Systems [10].
- We proposed joint adversarial fine-tuning with denoiser as a pre-processing block for removing adversarial perturbation/noise from the adversarial attacks on ASR systems [11]

2.5. Results and Discussion

With new attacks algorithms being proposed periodically, one important feature the defenses should posses is being attack-agnostic. To this end, we propose four pre-processing defenses, viz. randomized smoothing, DefenseGAN, variational auto-encoder (VAE), and Parallel-WaveGAN (PWG) vocoder to counteract against attacks on SR systems. Our best proposed method, PWG vocoder combined with randomized smoothing is able to withstand strong end-to-end adaptive white-box attacks, where both, the SR model and defense are made available to the adversary. Average accuracy improves by absolute 41% on average vs the undefended system with an notable absolute improvement $> 90\%$ for BIM attacks with $L_\infty > 0.001$

[1]This work is supported by DARPA-GARD HR001119S0026-GARD-FP-052



(a) *Speaker Identification Systems*

(b) *Speech Recognition Systems*

Figure 1: *Adversarial attacks on speech technologies*

and CW attack when the defense uses backward pass differentiable approximation. We extend this defense to state-of-the-art ASR models (DeepSpeech2 and Espresso) and it reduces the attack success rates in all evaluated scenarios and manages to recover most of the ground truth transcript (with a notable 4.4% to 15.8% WER increase for an Espresso system under an Imperceptible attack). We also proposed defenses using denoiser preprocessor, adversarial fine-tuning ASR model, and adversarial fine-tuning joint model of ASR and denoiser. We showed that denoiser pre-processor failed to defend against adaptive white-box attacks. However, adversarially fine-tuning joint model of denoiser and ASR yields a more successful defense. We show that this defense with frozen ASR parameters offers good protection, even against the strongest adaptive adversarial attacks (PGD attack with 500 iterations) yielding a mean absolute decrease of 45.08% GT WER and an increase of 68.05% adversarial target WER.) [11]

3. Countermeasure II: Detect and classify[2]

3.1. Research Question

Can we detect the presence of an adversarial attack algorithm? Can we infer which attack algorithm does the attack belong to? Can we detect if an incoming attacks belongs to a hereto unknown attack algorithm class?

3.2. Motivation

Instead of making the system inherently robust (Section 2), develop a "denial-of-service" style defense. In other words, build a gate–if the input to the system is suspected to be an adversarial attack, do not process it. Also, classifying/detecting attacks w.r.t. the attack algorithm, threat model and/or signal-to-adversarial-noise ratio might be helpful to identify the attackers, their intentions, and decide which defenses might be most effective against detected attacks.

3.3. Key Challenges

- Building one model that successfully counteracts against all attacks is a difficult task.
- Re-training the system to defend against one particular attack algorithm is computationally very expensive.

3.4. Major contributions

- We proposed a method using representation learning to generate embeddings, called "attack signatures", that give information about attack algorithm, SNR, threat model, etc. [12]
- We improved on the above system using AdvEst (Adversarial Perturbation Estimation) [11].

3.5. Results and Discussion

Our proposed method uses representation learning approach based on x-vector architecture and show that common attacks in the literature can be classified with accuracies as high as 90%. Also, *signatures* trained to classify attacks against speaker identification can be used also to classify attacks against speaker verification and speech recognition. We are able to detect unknown attacks with equal error rates of about 19%, which is promising. We also improve this approach using AdvEst [11], a method to estimate adversarial perturbation. We use adversarial perturbations as opposed to adversarial examples (consisting of the combination of clean signal and adversarial perturbation)as it eliminates nuisance information. A time-domain denoiser is used during inference to estimate the adversarial perturbations from adversarial examples. We evaluate the performance of obtained signatures on three applications: known attack classification, attack verification, and unknown attack detection. We showed that common attacks with different L_p threat models) can be classified with an accuracy of $\sim$ 96% and unknown attacks can be detected with an equal error rate (EER) of $\sim$9%, which is absolute improvement of $\sim$12% from our previous work [12].

4. Future directions

Although our preliminary work shows great results, the problem is still far from being solved. It remains an open problem of what holistic approaches would be best for making speech technologies trustworthy. As new attack algorithms keep breaking existing defenses, improving detection of unknown attacks and having a generalized defense strategy is of prime importance. We aim to keep pushing the frontiers for defenses to make existing systems robust to unforeseen and new attacks.

5. Acknowledgements

I sincerely thank my supervisor Prof. Najim Dehak for his invaluable guidance towards my dissertation. I am very grateful to my co-supervisors and collaborators Prof. Jesús Villalba, Dr. Piotr Żelasko, Prof. Laureano Moro-Velázquez for providing me an opportunity to work with them and giving useful feedback at various stages in my research. I would also like to thank all the collaborators on the DARPA-GARD project.

[2]This work supported by DARPA-RED HR00112090132

6. References

[1] I. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *International Conference on Learning Representations*, 2015.

[2] N. Carlini and D. Wagner, "Audio adversarial examples: Targeted attacks on speech-to-text," in *IEEE Security and Privacy Workshops (SPW)*, 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. 9141–9150.

[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] H. Abdullah, K. Warren, V. Bindschaedler, N. Papernot, and P. Traynor, "SoK: The Faults in our ASRs: An Overview of Attacks against Automatic Speech Recognition and Speaker Identification Systems," *IEEE Symposium on Security and Privacy*, 2021.

[6] J. Villalba, Y. Zhang, and N. Dehak, "x-vectors meet adversarial attacks: Benchmarking adversarial robustness in speaker verification," *Interspeech*, 2020.

[7] 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.

[8] 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," *arXiv*, Jan. 2021.

[9] N. Carlini and D. Wagner, "Towards Evaluating the Robustness of Neural Networks," in *IEEE Symposium on Security and Privacy*, 2017.

[10] 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.

[11] S. Joshi, S. Kataria, J. Villalba, and N. Dehak, "Advest: Adversarial perturbation estimation to classify and detect adversarial attacks against speaker identification," *Accepted at Interspeech*, 2022.

[12] J. Villalba, S. Joshi, P. Żelasko, and N. Dehak, "Representation learning to classify and detect adversarial attacks against speaker and speech recognition systems," *InterSpeech*, 2021.