Computer Speech & Language 68 (2021) 101199

Contents lists available at ScienceDirect

Computer Speech & Language

journal homepage: www.elsevier.com/locate/csl

Adversarial attack and defense strategies for deep speaker recognition systems

Arindam Jati[1,a,*], Chin-Cheng Hsu[1,a], Monisankha Pal[a], Raghuveer Peri[a], Wael AbdAlmageed[a,b], Shrikanth Narayanan[a,b,**]

[a] Electrical and Computer Engineering, University of Southern California (USC), Los Angeles, CA, USA
[b] USC Information Sciences Institute, Marina del Rey, CA, USA

ARTICLE INFO

Article History:
Received 19 August 2020
Revised 21 November 2020
Accepted 16 January 2021
Available online 12 February 2021

Keywords:
Adversarial attack
Deep neural network
Speaker recognition

ABSTRACT

Robust speaker recognition, including in the presence of malicious attacks, is becoming increasingly important and essential, especially due to the proliferation of smart speakers and personal agents that interact with an individual's voice commands to perform diverse and even sensitive tasks. Adversarial attack is a recently revived domain which is shown to be effective in breaking deep neural network-based classifiers, specifically, by forcing them to change their posterior distribution by only perturbing the input samples by a very small amount. Although, significant progress in this realm has been made in the computer vision domain, advances within speaker recognition is still limited. We present an expository paper that considers several adversarial attacks to a deep speaker recognition system, employs strong defense methods as countermeasures, and reports a comprehensive set of ablation studies to better understand the problem. The experiments show that the speaker recognition systems are vulnerable to adversarial attacks, and the strongest attacks can reduce the accuracy of the system from 94% to even 0%. The study also compares the performances of the employed defense methods in detail, and finds adversarial training based on Projected Gradient Descent (PGD) to be the best defense method in our setting. We hope that the experiments presented in this paper provide baselines that can be useful for the research community interested in further studying adversarial robustness of speaker recognition systems.

1. Introduction

Deep learning models have been recently found to be vulnerable to *adversarial attacks* (Szegedy et al., 2013; Biggio et al., 2013) where the attacker potentially discovers blind spots in the model, and crafts *adversarial samples* that are only slightly different from the original samples, rendering the trained model fail to correctly classify them or even to perform any other inference task on them. Over the last few years, researchers have been devoted to devising novel adversarial attack algorithms (Goodfellow et al., 2015; Madry et al., 2018; Papernot et al., 2016b; Carlini and Wagner, 2017), proposing defensive countermeasures to gain robustness (Goodfellow et al., 2015; Madry et al., 2018), and demonstrating exploratory analyses (Carlini and Wagner, 2017; Athalye et al., 2018; Carlini et al., 2019).

*Corresponding author.
**Corresponding author at: Electrical and Computer Engineering, University of Southern California (USC), Los Angeles, CA, USA.
E-mail addresses: jati@usc.edu (A. Jati), chincheh@usc.edu (C.-C. Hsu), mp_323@usc.edu (M. Pal), rperi@usc.edu (R. Peri), wamageed@isi.edu (W. AbdAlmageed), shri@sipi.usc.edu (S. Narayanan).
[1] Authors contributed equally.

https://doi.org/10.1016/j.csl.2021.101199
0885-2308/

Adversarial attack on speech processing systems With the rapid increase in the incorporation of Deep Neural Networks (DNN) within speech processing applications like Automatic Speech Recognition (ASR) (Chan et al., 2016; Audhkhasi et al., 2017), speaker recognition (Hansen and Hasan, 2015; Chung et al., 2018; Snyder et al., 2018; Jati and Georgiou, 2019), and speech emotion and human behavior studies (Narayanan and Georgiou, 2013; Huang and Narayanan, 2017; Bone et al., 2017), it is becoming essential to study the probable weaknesses of the employed models in the presence of adversarial attacks. In Carlini and Wagner (2018), the authors have shown that it is possible to achieve even 100% *success rate* in attacking deep ASR systems. In Qin et al. (2019) the authors have successfully generated imperceptible (to humans) adversarial audio samples while retaining high attack success rate. These studies highlight the vulnerability of deep ASR models against adversarial attacks.

Adversarial attack on speaker recognition systems Speaker recognition models are being widely employed in several applications including smart speakers and personal digital assistants (Wan et al., 2018; Hansen and Hasan, 2015), bio-metric systems (Nucci and Keralapura, 2012), and forensics (Becker et al., 2008). Therefore, having robust speaker recognition models that are not susceptible to adversarial perturbation is an important requirement. However, speaker recognition models have *not* been investigated extensively in the presence of adversarial attacks. Some initial work can be found in the literature (Section 3), but a detailed analysis of *white box* attacks (will be discussed in Section 2.2) with state-of-the art attack algorithms is difficult to find. Moreover, to the best of our knowledge, effective defensive countermeasures for those attacks have *not* been proposed. The present work aims to address these issues in particular.

Contributions This paper focuses on adversarial attacks and possible countermeasures for deep speaker recognition systems, with the following contributions.

- In contrast to previous works in this field (discussed in Section 3), we perform adversarial attacks directly on the time domain speech signal (as opposed to the spectrogram), which provides more insight about possible "over-the-air attacks" on speaker recognition systems (see Section 2.2.1 for detailed discussions).
- We provide an extensive analysis of the effect of multiple state-of-the-art white box adversarial attacks on a DNN-based speaker recognition model.
- We propose multiple defensive countermeasures for the deep speaker recognition system, and analyze their performance.
- We perform *transferability analysis* (Carlini et al., 2019) to investigate how adversarial speech crafted with a particular model can also be harmful to a different model.
- We present various ablation studies (*e.g.,* varying the strength of the attack, measuring signal-to-noise ratio (SNR) and perceptibility of the adversarial speech samples *etc.,*) that might be helpful to gain a comprehensive understanding of the problem.
- We share ready-to-run software implementation[2] of the present work toward supporting reproducibility and further research.

We aim to set baselines in the present exposition study, and hope it can help the community interested to continue further research in this domain.

The rest of the paper is organized as follows. In Section 2, we provide preliminaries about speaker recognition and adversarial attacks. In Section 3, we highlight related work. The adversarial attack algorithms and defense strategies are introduced in Section 4. The experimental setting and results are described in Section 5 and Section 6, respectively. Finally, conclusions and future directions are provided in Section 7.

2. Background

2.1. Speaker recognition systems

Speaker recognition systems can be developed either for identification or verification (Hansen and Hasan, 2015) of individuals from their speech. In a *closed set* speaker identification scenario (Hansen and Hasan, 2015; Jati and Georgiou, 2019), we are provided with train and test utterances from a set of unique speakers. The task is to train a model that accurately classifies a test utterance to one of the training speakers. Speaker verification (Snyder et al., 2018; Chung et al., 2018), on the other hand, is an *open set* problem. The task is to verify whether a test utterance claiming a particular speaker's identity is actually spoken by that speaker (whose enrollment utterance is available beforehand). The training data in the latter case, is generally utterances from a mutually exclusive set of speakers.

Although speaker verification differs from speaker identification during the testing phase, most of the recent state-of-the-art speaker verification systems (Snyder et al., 2018; Nagrani et al., 2017; Chung et al., 2018; Snyder et al., 2017) are trained with the objective of learning to classify the set of training speakers. In other words, these models are trained with a cross-entropy objective over the unique set of training speakers (*i.e.,* similar to a speaker identification scenario).

Formally, let $\boldsymbol{x} \in \mathbb{R}^D$ denote a time domain audio sample with speaker label y, then learning a speaker identifier model is generally done through Empirical Risk Minimization (ERM) (Madry et al., 2018):

$$argmin_{\boldsymbol{\theta}} \quad \mathbb{E}_{(\boldsymbol{x},y)\sim\mathscr{D}}[L(\boldsymbol{x}, y, \boldsymbol{\theta})] \tag{1}$$

[2] Source codes are available at https://github.com/usc-sail/gard-adversarial-speaker-id

where, $L(\cdot)$ is the cross-entropy objective, and $\boldsymbol{\theta}$ denotes the set of trainable parameters of the DNN.

An intermediate representation of the trained DNN model might be subsequently extracted as a *speaker embedding* (Snyder et al., 2018) which is expected to carry speaker-specific information. The speaker embeddings are then utilized for verification purposes. In this study, most of our detailed experiments and ablation studies will be performed on a close set speaker identification (or classification) model, and a dedicated section will be provided for speaker verification experiments (in Section 6.8).

2.2. Adversarial attack

Given an audio sample $\boldsymbol{x} \in \mathbb{R}^D$, an adversarial attack generates a perturbed signal given by

$$\tilde{\boldsymbol{x}} = \boldsymbol{x} + \boldsymbol{\eta} \quad \text{such that} \quad \|\boldsymbol{\eta}\|_p < \epsilon \tag{2}$$

with the goal of forcing the classifier to produce erroneous output for $\tilde{\boldsymbol{x}}$. In other words, if $\boldsymbol{x}$ has a true label y, then the attacker forces the classifier to produce $\tilde{y} \neq y$ for the perturbed sample $\tilde{\boldsymbol{x}}$. In this paper, we will focus on l_∞ and l_2 norms which are most widely employed in literature.

2.2.1. Attack space

Adversarial attack on speech can be performed on different signal spaces *e.g.,* the time domain raw waveform, the extracted spectrogram, or any other feature spaces (Chen et al., 2019a; Qin et al., 2019)[3]. Targeting the time domain speech signal can open up opportunities for "over-the-air" attacks where the perturbation is added to the speech signal even before it is received by the microphone of the speech processing system (speaker recognition system in our case). Our current work focuses on time domain attacks.[4]

2.2.2. Threat model

We explore *white-box* (Carlini et al., 2019) attacks in this study. This assumes that the attacker has complete knowledge of the model architecture, parameters, loss functions, and gradients. We adopt this stronger form of attack (compared to black-box attack (Carlini et al., 2019)) because it does not assume that any part of the model can be kept hidden from the attacker, and it is the most frequently employed threat model in the adversarial attack literature (Goodfellow et al., 2015; Madry et al., 2018; Papernot et al., 2016b; Carlini and Wagner, 2017).

Adversarial attacks can be *targeted* or *untargeted* (Carlini et al., 2019). An untargeted attack only forces the model to make erroneous predictions, whereas a targeted attack aims at forcing the model to predict the class that the adversary desires. We perform *untargeted attacks* in this study, and leave the targeted attack for future study.

2.2.3. Transferability

Although most of the experiments in this paper are performed for white box attacks, we study the transferability of adversarial samples in Section 6.6, which gives us a notion of performance during a black-box attack as well. The transferability test (Carlini et al., 2019; Papernot et al., 2016a) evaluates the vulnerability of a *target* model against the adversarial samples generated with a *source* model. The attacker has full knowledge about the source model, but no or limited knowledge about the target model (for example, knowledge about the fact that both source and target have convolutional layers). The goal of the attacker is to generate adversarial samples (with the source model) in such a way that they "transfer well" to the target model, *i.e.,* those samples also make the target model vulnerable.

3. Related work

This section describes key previous work on adversarial attack and defense methods proposed for speaker recognition systems.

- Li et al. (2020) showed that an i-vector (Dehak et al., 2010) based speaker verification system is susceptible to adversarial attacks, and the adversarial samples generated with the i-vector system also transfer well to a DNN-based x-vector (Snyder et al., 2018) system.[5] The attack was performed on the feature space (and not directly on the time domain speech signal), and only the Fast Gradient Sign Method (FGSM) (Goodfellow et al., 2015) (will be further discussed in Section 4) was investigated for that purpose. Moreover, no defense method was proposed.

[3] Throughout the paper, by "feature space" we refer to (mel-)spectrogram features that are generally extracted for further processing in widely-employed speaker recognition systems (Snyder et al., 2018; 2017)

[4] Note that we do *not* perform over-the-air attacks in this paper. Over-the-air attacks can be designed by following the work of Qin et al. (2019), and it is straightforward with time domain attacks as done in this work.

[5] i-vectors have been the state-of-the-art in speaker verification for a decade until DNN-based x-vectors were shown to outperform them (Snyder et al., 2017; 2018).

- Kreuk et al. (2018) demonstrated the vulnerability of an end-to-end DNN-based speaker verification system to an FGSM attack. The attack was done on feature space, and the authors discovered cross-feature transferability of the adversarial samples. No defense method was proposed in the paper.
- Chen et al. (2019a) proposed the Natural Evolution Strategy (NES) based adversarial sample generation procedure, and successfully attacked a GMM-UBM system[6] and i-vector based speaker recognition systems. They found an impressive attack success rate with their proposed method. However, the authors did not attack more recent DNN-based speaker recognition frameworks which are shown to have state-of-the-art performances. Moreover, the test set involved in their experiments only included 5 speakers (TABLE I of Chen et al., 2019a), and thus, an extensive study with a much higher number of test speakers is still needed.
- Wang et al., (2019) proposed adversarial regularization based defense methods using FGSM and Local Distributional Smoothness (LDS) (Miyato et al., 2015) techniques. The proposed method was shown to improve the performance of a speaker verification system, but only FGSM was employed as the attack algorithm, and similar to most of the above methods, the attack was performed on the feature space and not on the time domain audio.

In summary, although these studies represent important initial efforts on adversarial attacks on speaker recognition systems, many technical questions still remain to be addressed. Limitations include consideration of primarily feature space attacks (Li et al., 2020; Kreuk et al., 2018; Wang et al., 2019) (and not time domain), limited number of attack algorithms (Li et al., 2020; Kreuk et al., 2018; Chen et al., 2019a; Wang et al., 2019), limited number of speakers in the test set (Chen et al., 2019a), and no or limited number of defense methods (Li et al., 2020; Kreuk et al., 2018; Chen et al., 2019a). The present exposition study aims to address these limitations by reporting extensive experimental analysis, ablation studies, and by proposing and evaluating various defense methods.

4. Attack and defense algorithms

4.1. Attack algorithms

A group of gradient-based attack algorithms tries to maximize the loss function by finding a suitable perturbation which lies inside the l_p-ball around the sample $\boldsymbol{x}$. Formally,

$$\max_{\boldsymbol{\eta}:\|\boldsymbol{\eta}\|_p<\epsilon} \quad L(\boldsymbol{x}+\boldsymbol{\eta}, y, \boldsymbol{\theta}). \tag{3}$$

Here, the notations follow Eq. (2), *i.e.,* $\boldsymbol{\eta}$ denotes the adversarial noise being added, y denotes the true label of sample $\boldsymbol{x}$, and $\boldsymbol{\theta}$ denotes the parameters of the DNN model.

A different group of algorithms aims at decreasing the posterior of the true output class, and increasing the posterior of the most confusing wrong class. Here we present the attack algorithms we employ in our study. *Fast Gradient Sign Method (FGSM)* Goodfellow et al. (2015) proposed this computationally efficient one-step l_∞ attack to generate adversarial samples by using only the sign of the gradient function, and moving in the direction of gradient to increase the loss:

$$\tilde{\boldsymbol{x}} = \boldsymbol{x} + \epsilon \ \text{sign}(\nabla_{\boldsymbol{x}} L(\boldsymbol{x}, y, \boldsymbol{\theta})). \tag{4}$$

Projected Gradient Descent (PGD) Madry et al. (2018) proposed a more generalized version with iterative gradient based l_∞ attack:

$$\tilde{\boldsymbol{x}}_{i+1} = \Pi_{\boldsymbol{x}+\mathscr{S}}[\tilde{\boldsymbol{x}}_i + \alpha \ \text{sign}(\nabla_{\boldsymbol{x}} L(\boldsymbol{x}, y, \boldsymbol{\theta}))], \tag{5}$$

where, α is the step size of the gradient descent update, $\boldsymbol{x}+\mathscr{S}$ is the set of allowed perturbations *i.e.,* the l_∞-ball around $\boldsymbol{x}$, and $\Pi_{\boldsymbol{x}+\mathscr{S}}$ denotes the constrained projection operation in a standard PGD optimization algorithm. Throughout the text, we will denote PGD with a fixed number of T iterations with by "PGD-T".

Carlini and Wagner attack (Carlini l_2 and Carlini l_∞) Carlini and Wagner (2017) defined the general methodology of their attack by

$$\begin{aligned} &\text{minimize} && \|\boldsymbol{\eta}\|_p + c \cdot g(\tilde{\boldsymbol{x}}) \\ &\text{such that} && \tilde{\boldsymbol{x}} \in [0,1]^D. \end{aligned} \tag{6}$$

Here, $g(\cdot)$ defines the objective function given by

$$g(\tilde{\boldsymbol{x}}) = \left[Z(\tilde{\boldsymbol{x}})_t - \max_{j \neq t} \left(Z(\tilde{\boldsymbol{x}})_j \right) + \delta \right]_+ \tag{7}$$

where, $Z(\cdot)$ is the output vector containing posterior probabilities for all the classes, t denotes the output node corresponding to the true class y, δ is the confidence margin parameter, and $[\cdot]_+$ denotes the $\max(\cdot, 0)$ function. Intuitively, the attack tries to maximize the posterior probability of a class that is *not* the true class of $\boldsymbol{x}$, but has the highest posterior among all the wrong classes.

[6] GMM-UBM stands for Gaussian Mixture Model-Universal Background Model, a classical model in speaker recognition (Reynolds et al., 2000).

The norm can be either l_2 or l_∞. For Carlini l_∞ attack, the minimization of $\| \boldsymbol{\eta} \|_\infty$ is not straightforward due to non-differentiability, and an iterative procedure is employed in Carlini and Wagner (2017).[7]

4.2. Defense algorithms

Adversarial training

The intuition here is to train the model on adversarial samples generated by a certain adversarial attack. The adversarial samples are generated online using the training data and the current model parameters. Madry et al. (2018) introduced the generalized notion of adversarial training by a *mini-max optimization* given by:

$$argmin_{\boldsymbol{\theta}} \quad \mathbb{E}_{(\boldsymbol{x},y)\sim\mathscr{D}}\Big[\max_{\boldsymbol{\eta}:\|\boldsymbol{\eta}\|_p<\epsilon} \quad L(\boldsymbol{x}+\boldsymbol{\eta},y,\boldsymbol{\theta})\Big] \tag{8}$$

The *inner maximization* task is addressed by the attack algorithm utilized during adversarial training, and the *outer minimization* is the standard ERM (Eq. (1)) employed to train the model parameterized with $\boldsymbol{\theta}$. We separately apply both one-step FGSM (Eq. (4)), and T-step PGD (Eq. (5)) algorithms to solve the inner maximization problem. Throughout the remaining text, we refer to these as "FGSM adversarial training" and "PGD-T adversarial training" respectively.

Notably, the overall training is done on clean as well as adversarial samples. The overall loss function is given by:

$$L_{\text{AT}}(\boldsymbol{x},\tilde{\boldsymbol{x}},y,\boldsymbol{\theta}) = (1-w_{\text{AT}})\cdot L(\boldsymbol{x},y,\boldsymbol{\theta}) + w_{\text{AT}}\cdot L(\tilde{\boldsymbol{x}},y,\boldsymbol{\theta}), \tag{9}$$

where $w_{\text{AT}} \in [0,1]$ is the weight of the adversarial training.

Adversarial Lipschitz Regularization (ALR) This approach of gaining robustness is based on learning a function that is not much sensitive to a small change in the input. In other words, if we can learn a relatively smooth function, then the posterior distribution should not vary abruptly if the input perturbation is within the maximum allowed limit. We propose a training strategy equipped with the recently proposed adversarial Lipschitz regularization technique (Terjék, 2020). Similar to the regularization based on local distribution smoothness in Virtual Adversarial Training (VAT) (Miyato et al., 2015), ALR imposes a regularization term defined using Lipschitz smoothness:

$$\|f\|_L = \sup_{\boldsymbol{x},\tilde{\boldsymbol{x}}\ \in X, d_X(\boldsymbol{x},\tilde{\boldsymbol{x}})>0} \frac{d_Y\Big(f(\boldsymbol{x}),f(\tilde{\boldsymbol{x}})\Big)}{d_X(\boldsymbol{x},\tilde{\boldsymbol{x}})}, \tag{10}$$

where $f(\cdot)$ the function of interest (implemented by the neural network) that maps the input metric space (X,d_X) to output metric space (Y,d_Y). In our case of speaker classification, we chose $f(\cdot)$ as the final log-posterior output of the network, *i.e.*, $f(\boldsymbol{x}) = \log p(y|\boldsymbol{x},\boldsymbol{\theta})$, l_1 norm as d_Y, and l_2 norm as d_X. The adversarial perturbation $\boldsymbol{\eta} = \epsilon\ \boldsymbol{\eta}_k$ in $\tilde{\boldsymbol{x}} = \boldsymbol{x}+\boldsymbol{\eta}$ is approximated by the power iterations:

$$\boldsymbol{\eta}_{i+1} = \frac{\nabla_{\boldsymbol{\eta}_i} d_Y(f(\boldsymbol{x}),f(\boldsymbol{x}+\xi\boldsymbol{\eta}_i))}{\|\nabla_{\boldsymbol{\eta}_i} d_Y(f(\boldsymbol{x}),f(\boldsymbol{x}+\xi\boldsymbol{\eta}_i))\|_2}, \tag{11}$$

where, $\boldsymbol{\eta}_0$ is randomly initialized, and ξ is another hyperparameter (see Section 5.5). The regularization term added to training is

$$L_{ALR} = \left[\frac{d_Y\Big(f(\boldsymbol{x}),f(\tilde{\boldsymbol{x}})\Big)}{d_X(\boldsymbol{x},\tilde{\boldsymbol{x}})} - K\right]_+, \tag{12}$$

where K is the desired Lipschitz constant we wish to impose.

5. Experimental setting

We implement the core of most of our attack and defense algorithms (except ALR) through the Adversarial Robustness Toolbox (Nicolae et al., 2018). For ALR, we follow the original implementation of Terjék (2020). The rest of the experimental details are described below.

5.1. Dataset

We employ Librispeech (Panayotov et al., 2015) (the "train-clean-100" subset) dataset for all the experiments. It contains 100 hours of clean speech from 251 unique speakers (125 females). We utilize all the speakers for our experiment. For every speaker, we employ 90% of the utterances for training the classifier, and the remaining 10% utterances for testing. The train-test split is deterministic, and it is kept fixed throughout all the experiments.

[7] We suggest the readers to refer to (Carlini and Wagner, 2017) for detailed information about the iterative workaround for l_∞ attack, and also for choosing the values for the weight parameter, c.

5.2. Model architectures

We implemented our classifier, $f : \mathscr{X} \rightarrow \mathscr{Y}$, by combining a Convolutional Neural Network (CNN) with a digital signal processing (DSP) front-end. The DSP front-end is non-trainable but differentiable, and it extracts log Mel-spectrogram which can be viewed as a temporal signal of F channels, where F is the number of Mel frequency bins. The back-end is either of the two DNN models described below. As both modules are differentiable, the adversarial attack schemes introduced in Section 4 can be applied to create time-domain perturbation directly.

5.2.1. 1D CNN

Our classifier comprises three components: a DSP front-end, a speaker embeddings extractor, and a linear classifier. The extractor consists of 8 stacks of convolutional layers and transforms the spectrogram into a single vector of speaker embedding $\boldsymbol{v} \in \mathbb{R}^{32}$. All of the CNN layers are coupled with a batch normalization and ReLU non-linearity. Max-pooling is employed 3 times to down-sample the feature map. The 32D speaker embedding is obtained by max pooling over time of the output of the final CNN layer. The linear classifier maps the embedding into the class logits. The model has around 219 thousand trainable parameters in total. The complete network architecture is shown in Table 5. We analyze this model throughout the paper.

5.2.2. TDNN

The Time Delay Neural Network (TDNN) (Snyder et al., 2017; 2018) is one of the current state-of-the-art models for speaker recognition. We adopt the model architecture proposed in Snyder et al. (2018) for the experiments related to transferability analysis (Section 6.6). The model consists of time-dilated convolutional layers along with a statistics pooling module, and it has $\sim$4.4 million trainable parameters, and hence, is much larger than the 1D CNN model. We use this model only in transferability experiments where we create adversarial perturbation from TDNN to attack our model, and vice versa.

5.3. Training parameters

We employ the Adam optimizer (Kingma and Ba, 2014) with a learning rate of 0.001, $\beta_1 = 0.5$, and $\beta_2 = 0.999$. We use a mini-batch size of 128. We train all the models until the training loss saturates.

5.4. Attack parameters

Our main results (Section 6.3) are obtained from the experiment with attack strength $\epsilon = 0.002$ for l_∞ attacks, and confidence margin $\delta = 0$ for a Carlini l_2 attack. We empirically chose $\epsilon = 0.002$ because it results in an average SNR of around 30 dB and $\sim$3.5 PESQ score for the FGSM/PGD adversarial samples. This will be explained in more detail in Section 6.1 and Section 6.2. Furthermore, we vary the strength of different attacks, and the results are shown in Section 6.4. The PGD attack is performed for 100 iterations with a step size $\alpha = \epsilon/5$.

5.5. Defense parameters

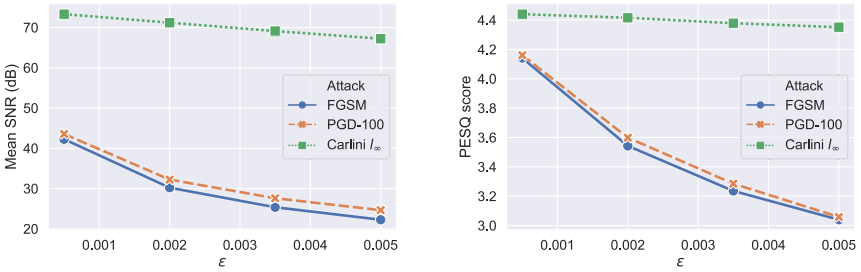
In the ALR method, we set the number of power iterations $K = 1$, and the hyperparameter $\xi = 10$, as recommended in Terjék (2020). The FGSM- and PGD-based adversarial training algorithms are run with $\epsilon = 0.002$. Hence, the main results (Section 6.3) employ the same ϵ value in both the attack and the adversarial training based defense. The ablation study in Section 6.4 is particularly designed to investigate the effect of using different values of ϵ during the attack. Specifically, the study varies ϵ above and below the vicinity of $\epsilon = 0.002$ (set during defense training), and analyzes the effectiveness of the defense method. The PGD adversarial training uses 10 iterations (*i.e.,* PGD-10 as introduced in Section 4.2)[8], although we evaluated it against PGD attacks with higher number of iterations (Sections 6.3 and 6.5). During adversarial training, we create minibatches containing equal number of clean and adversarial samples, *i.e.,* in Eq. (9), we set $w_{\text{AT}} = 0.5$.

6. Results and discussion

6.1. *Attack strength* vs. *SNR*

To have a substantial understanding about the strength of different attack algorithms, we computed the mean Signal-to-Noise Ratio (SNR) of all the test adversarial samples for every level of attack strength. For l_∞ attacks, ϵ varies between $\{0.0005, 0.002, 0.0035, 0.005\}$, and for Carlini l_2 attack the confidence margin δ varies between $\{0, 0.001, 0.01, 0.1\}$. The curves for l_∞ attacks are shown in Fig. 1a. There are two important observations. First, the average level of SNR is $\sim$30 dB higher for Carlini l_∞ than FGSM and PGD. Second, the SNR level tends to decrease faster with increase in ϵ for PGD and FGSM as compared to Carlini l_∞. The reason might be attributed to the optimization algorithms that various attack methods use for generating the adversarial samples. For example, the Carlini method enforces a minimum perturbation required to change the output prediction,

[8] PGD adversarial training is slow, and we could not afford to run it for more than 10 iterations.



(a) Attack strength *vs.* SNR (b) Attack strength *vs.* perceptibility

Fig. 1. Mean SNR and PESQ score of the test adversarial samples generated by different l_∞ attack algorithms at different strengths.

while PGD enforces perturbation projected inside the l_∞-ball of size ϵ around $\boldsymbol{x}$. We have also computed mean SNR for Carlini l_2 attack for different values of the confidence margin δ. The SNR level tends to stay around 75 dB, and does not vary much with increasing δ. A visualization of the adversarial spectrograms is shown in Appendix A for a more detailed analysis of the attack algorithms.

6.2. *Attack strength* vs. *perceptibility*

We measure the perceptibility of the generated adversarial samples by employing Perceptual Evaluation of Speech Quality (PESQ) (Recommendation, 2001; Rix et al., 2001). While subjective measure with multiple human annotators can be more accurate, it is time-consuming and costly. The objective PESQ measure has been the ITU-T standard for measuring telephonic transmission quality. It gives a mean opinion score by comparing the degraded speech signal with the original speech recording. The PESQ score is between -0.5 to 4.5, and a higher value indicates better quality. Fig. 1b depicts the average PESQ scores of all the test adversarial samples generated via different attack methods at various strengths. We can see the gradual degradation of audio quality with the increase of attack strength. Similar to the findings in the SNR analysis (Section 6.1), Carlini l_∞ attack produces more perceptible adversarial audio samples than PGD-100 and FGSM. The degradation is also slower for Carlini attack. It is noteworthy that at $\epsilon = 0.0005$ the attack algorithms are able to achieve high audio quality (PESQ score > 4.0), but force the classifier to produce erroneous outputs (Section 6.4). We have also computed the average PESQ score for the Carlini l_2 attack. The PESQ score is ~ 4.4, and does not vary much with change in the confidence margin δ.

6.3. *Main results*

Table 1 presents the test performance of standard training (no defense) and all the employed defense methods for three l_∞ attacks, and one l_2 attack. All the performances are averaged over 10 random runs. Please note that "benign accuracy" denotes the accuracy on the clean samples, and "adversarial accuracy" denotes the accuracy on the adversarial samples generated with a particular attack algorithm. We will use this terminology throughout the remainder of the text. The l_∞ attacks are with $\epsilon = 0.002$, and all the adversarial training methods are run with the same ϵ value. As we can see, the accuracy of the standard training method drops from 94% by a significant margin for all the attacks. This shows the vulnerability of the model, and further underscores the need for strong countermeasures. A comparison between the three l_∞ attacks shows the FGSM is the weakest one (25% adversarial accuracy for standard training), and PGD-100 (PGD with 100 iterations) is the strongest one (0% adversarial accuracy for standard training).

Comparing different defense methods, we can see that FGSM-based adversarial training is the weakest defense strategy. The ALR method is better than FGSM adversarial training, although it fails to defend against a PGD-100 attack. The PGD-10 adversarial

Table 1
Different attacks on a speaker recognition system, and performance of different defense methods. "Benign accuracy" denotes accuracy on clean samples, and "Adversarial accuracy" denotes accuracy on adversarial samples generated with different attack algorithms. "AT" stands for Adversarial Training (the defense method described in Section 4.2). Accuracy is on a scale of [0,1].

Defense method	Benign accuracy	Adversarial accuracy with different attacks			
		FGSM	Carlini l_∞	PGD-100	Carlini l_2
No defense	0.94	0.25	0.02	0	0
FGSM AT	0.82	0.20	0.09	0	0
ALR	**0.96**	0.44	0.10	0	0
PGD-10 AT	0.92	**0.73**	**0.58**	**0.43**	**0.09**

training is found to perform the best in our experiments. It is interesting to see that PGD adversarial training with 10 iterations is able to defend against a PGD attack with 100 iterations. The proposed PGD-10 adversarial training gives absolute improvements of 48%, 56% and 43% over the undefended performance against FGSM, Carlini l_∞ and PGD-100 attacks, respectively.

As observed in the previous literature, the performance gain achieved by PGD-10 adversarial training against different adversarial attacks generally results in a drop in benign accuracy. Similarly, in our experiment, the accuracy on the clean test samples drops for both FGSM- and PGD-based adversarial training methods, with the FGSM variant getting higher drop in performance. The ALR method, on the other hand, achieves a 2% absolute improvement in benign accuracy compared to the model with standard training, possibly because of lesser overfitting due to the presence of the penalty term shown in Eq. (12).

The last column of Table 1 shows the performance of different defense methods against Carlini l_2 attack. Standard training, FGSM adversarial training, and ALR algorithms are unable to defend against this attack. Defense with PGD-10 adversarial training also performs poorly for this l_2 attack. The reason might be attributed to the adversarial training methodology which is based on l_∞ perturbation, and thus, probably fails to defend against a strong l_2 attack.

A related ablation study is provided in Appendix B which shows the similarity between the misclassified predictions made by the model under different attacks. This could possibly reveal some inherent similarities between different attacks.

6.4. Ablation study 1: Varying attack strength

Fig. 2 shows how the performances of different defense methods vary when we vary the strength of the adversarial attacks. Note that the adversarial training-based defense methods still employ the same $\epsilon = 0.002$ during training, but ϵ of the attack algorithm varies.

We can observe that the general trend of the curves is downward with the increase of the strength of any attack. The only exception is the unprotected model under FGSM attack. The performance surprisingly increases in the beginning and then saturates.

Comparing different defense methods, we can see the PGD-10 adversarial training continues to outperform all the other defense methods for all attack types, and for all strength levels. The proposed ALR training is found to be the next best defense technique.

Another interesting observation is that the accuracy curves for both the Carlini methods are more flat in nature compared to FGSM and PGD attacks. The reason might be attributed to the relatively less drop in SNR values of the test adversarial samples generated by Carlni method as the attack strength increases, as explained in Section 6.1.

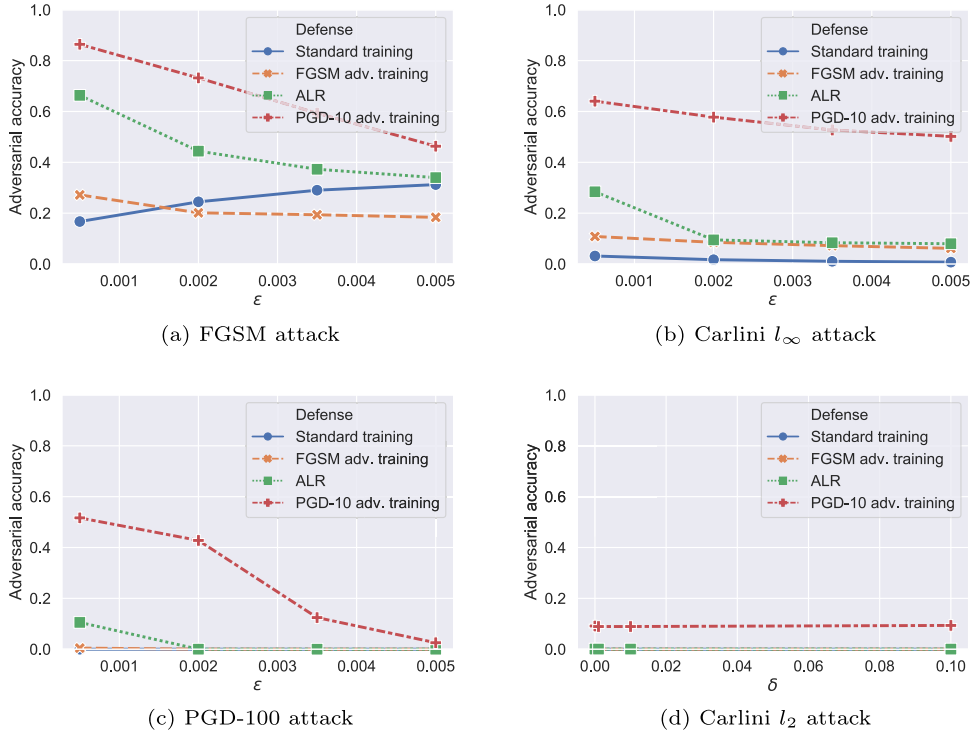


Fig. 2. Ablation study 1: Varying the strength (ϵ for three l_∞ attacks, δ for the Carlini l_2 attack) in different attack algorithms, and performance of different defense methods.

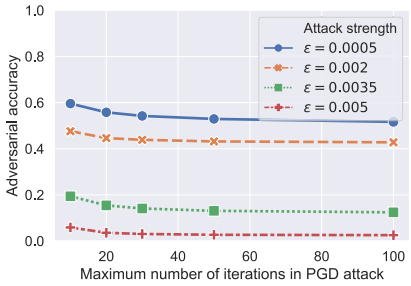


Fig. 3. Performance of PGD-10 adversarial training against PGD attack at different strengths, and with different number of iterations.

6.5. Ablation study 2: Analyzing the best defense method

Here we analyze the best defense method, *i.e.,* PGD-10 adversarial training, in further detail. Specifically, we investigate its behavior when we attack it with PGD attack with different number of iterations and at different strengths. Fig. 3 shows the variation of the adversarial accuracy for PGD-10 defense. Each line denotes attack at a particular strength *i.e.,* a particular ϵ value. The horizontal axis denotes the iteration number, varying in the range $\{10, 20, 30, 50, 100\}$. A closer inspection reveals that after the first drop in performance for PGD-10 attack to PGD-20 attack, the accuracy value tends to decrease very slowly. For example, at $\epsilon = 0.002$, adversarial accuracy against PGD-10 attack is $\sim 48\%$. Then a 3% absolute drop is observed when we perform a PGD-20 attack, and the adversarial accuracy becomes $\sim 45\%$. The accuracy tends to drop very slowly afterwards, and we see a $\sim 43\%$ accuracy against a PGD-100 attack. We conjecture that 20 iterations are sufficient for the adversary to find the best perturbation within the l_∞-ball with radius ϵ around $\boldsymbol{x}$. As a result, the effects of PGD with larger number of iterations are marginal.

6.6. Ablation study 3: Transferability analysis

We perform the transferability analysis between the smaller 1D CNN model, and the larger TDNN model defined in Section 5.2. The goal of this experiment is to investigate whether the adversarial samples generated with a simple convolutional model (such as the 1D CNN model) is able to make a different model (the TDNN model) vulnerable. We choose the off-the-shelf TDNN model for this purpose because of its widespread usage in speaker recognition experiments. Table 2 shows the performance of different models for benign and adversarial samples. We can see that adversarial samples crafted from the "source" model tend to be harmful to the "target" model as well, as evident from the significant drop in the performances. Adversarial samples generated with the larger model (TDNN) tend to be more effective in attacking the smaller model. Further studies are needed to fully understand the observed pattern of transferability.

6.7. Ablation study 4: Effect of noise augmentation

Noise augmentation is a standard technique employed during training a speaker recognition model (Snyder et al., 2018; Jati et al., 2019). Here, we experiment with augmenting the dataset with white Gaussian noise (scaled with a factor equal to the ϵ used in the attack) during training the *undefended* model. The model is trained with both clean and noisy samples. The

Table 2
Transferability of adversarial samples between different models. The adversarial samples are generated with the "source" model, but seem to be effective against the "target" model as well. $\epsilon = 0.002$ is employed for this experiment. Accuracy is on a scale of [0,1].

Benign accuracy		
	1D CNN	TDNN
	0.94	0.95
Adversarial accuracy for FGSM attack		
Source / Target	1D CNN	TDNN
1D CNN	0.24	0.37
TDNN	0.14	0.03
Adversarial accuracy for PGD-100 attack		
Source / Target	1D CNN	TDNN
1D CNN	0	0.28
TDNN	0.06	0

Table 3
Effect of training data augmentation with white Gaussian noise. $\epsilon = 0.002$ is employed for this experiment. Accuracy is on a scale of [0,1].

		Adversarial accuracy with different attacks	
Training data augmentation	Benign accuracy	*FGSM*	*PGD-100*
No	0.94	0.25	0
Yes	0.95	0.17	0

Table 4
Recall rates of our model in speaker verification settings. For this experiment, we employ FGSM with 10 random initializations, which is found to be stronger than the vanilla FGSM employed in the experiments of Section 6.3.

Recall Type	Attack	Benign	$\epsilon = 0.0005$	$\epsilon = 0.002$	$\epsilon = 0.0035$	$\epsilon = 0.005$
		without adversarial training				
positive	FGSM (10 init)	0.8684	0.5955	0.4873	0.4404	0.4076
	Carlini l_∞	"	0.5846	0.3906	0.3736	0.3644
	PGD-10	"	0.3911	0.0095	0.0032	0.0029
negative	FGSM (10 init)	0.8530	0.5298	0.4518	0.4310	0.4162
	Carlini l_∞	"	0.5314	0.3864	0.3791	0.3779
	PGD-10	"	0.3817	0.0335	0.0210	0.0157
		with PGD-10 adversarial training				
positive	FGSM (10 init)	0.8728	0.8248	0.7593	0.7077	0.6599
	Carlini l_∞	"	0.8231	0.8066	0.7890	0.7689
	PGD-10	"	0.8071	0.7179	0.6424	0.6102
negative	FGSM (10 init)	0.8495	0.8188	0.7702	0.7274	0.6870
	Carlini l_∞	"	0.8071	0.7974	0.7875	0.7759
	PGD-10	"	0.8133	0.7305	0.6571	0.6236

experimental observations are tabulated in Table 3. As expected, the benign accuracy improves. However, we could not find any improvement in the performance of the model for defending against FGSM and PGD-100 adversarial attacks. The reason might be attributed to the ability of attack algorithms to generate more novel noise samples (compared to simple white Gaussian noise) that force the model posteriors to change.

6.8. Speaker verification

Finally, we experiment our speaker recognition model on an open-set speaker verification (SV) setup. For this purpose, we compute the cosine similarity between the speaker embeddings of the enrollment and a query utterance, denoted by $\boldsymbol{v}_e$ and $\boldsymbol{v}_q$, respectively. We determine the decision boundary b on the dev-clean set (40 speakers unseen during training) of Librispeech, and apply the boundary for speaker verification on the test-clean set (another 40 unseen speakers). To create adversarial perturbation, we formulate verification as a binary classification problem where a positive case ($Y = 1$) denotes the decision that query and enrollment are the same speaker, and negative ($Y = 0$) otherwise. The model prediction is given by

$$P(Y|\boldsymbol{v}_q) = \frac{1}{1 + e^{-(d-b)}}, \quad \text{where } d = \frac{\boldsymbol{v}_q^\top \boldsymbol{v}_e}{\| \boldsymbol{v}_q \| \, \| \boldsymbol{v}_e \|}. \tag{13}$$

For positive pairs, in which the speaker in the enrollment and the query is the same, we enumerate all possible combinations (resulting in a total of 91,961 pairs). For negative pairs, we randomly drew 100 queries per enrollment from utterances of other 39 speakers (resulting in a total of 262,000 pairs). Following (Chen et al., 2019b), we report the positive and negative recall rates separately in Table 4. The random seed was kept fixed, so the benign recall rate remained the same across positive (or negative) experiments. The goal of attacking a positive pair is to fool the system into believing that the query is not the enrolled speaker, and the goal of attacking a negative pair is to make it believe the imposter is the enrolled speaker. Our experiments show that our speaker recognition model with adversarial training is more robust to adversarial attacks even in speaker verification settings compared to the unprotected baseline model. Moreover, even though the adversarial training includes only one type of attack (PGD-10 in our case), the robustness can be generalized to other types of attacks. In conclusion, the robustness induced by adversarial training can be transferred to a related task.

7. Conclusion and future directions

The paper presented an extensive exploratory analysis of adversarial attacks on a closed set speaker recognition system. We reported results obtained from experiments with multiple state-of-the-art attack algorithms with varying attack strengths. We

also investigated state-of-the-art defense methods, and adopted them for employing as countermeasures for the speaker recognition model. We performed several ablation studies to understand the SNR characteristics and perceptibility of the adversarial speech, analyze the transferability of the adversarial attacks, and the effectiveness of white noise augmentation during training. The main observations are the following:

- Speaker recognition system such as the one employed in the current study is vulnerable to white box adversarial attacks. The performance of the undefended model dropped from 94% to 0% with the strongest attacks (PGD-100, Carlini l_2) in our experiment even at 40 dB SNR and PESQ score > 4.
- Adversarial samples crafted with the Carlini and Wagner method are found to have the best perceptual quality in terms of the PESQ score.
- The adversarial samples generated with a particular source model are found to transfer well to a different target model, and hence, are also harmful for the target model. This is particularly alarming because it can open up chances for *black box* attacks.
- Augmenting training data with white Gaussian noise is *not* found to be effective.
- Experimenting with several defense methods showed that PGD-based adversarial training is the best defense strategy in out setting.
- Although PGD adversarial training is the best defense method, it is *not* found to be effective against l_2 attack in our experiments, probably because of employing l_∞ norm during training.
- Robustness induced by adversarially training a speaker recognition model translates to speaker verification.

We hope the source codes published along with this paper can be helpful to the research community interested in pursuing further work in this domain. Several important future directions can be taken from here.

- Metric learning such as triplet training (Schroff et al., 2015) are shown to learn compact and robust embeddings against adversarial attacks for images (Mao et al., 2019; Zhong and Deng, 2019). Metric learning is also found to be useful for learning robust speaker embeddings in Jati et al. (2019). A natural extension can be to verify the adversarial robustness of speaker embeddings learning via metric learning.
- Adaptive attacks (Tramer et al., 2020) are particularly designed to break any specific defense algorithm. The strategies introduced in Tramer et al. (2020) can be a starting point to perform model-specific adversarial attacks on existing defense methods proposed for speaker recognition systems.
- Studying targeted attacks might be another good direction from here, especially, since this could be a potential threat for biometric systems that rely on speaker recognition modules.
- Finally, further research can be done on crafting imperceptible (to human judgement or by retaining high PESQ score) adversarial audio samples with high attack success rate such as in Qin et al. (2019), and also formulating effective detection (Speakman et al., 2018) and defense algorithms as countermeasures.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

The work is supported by DoD.

Appendix A. Visualizing spectrograms

Fig. 4 shows the mel-spectrograms of a randomly chosen utterance for different attacks at varying ϵ values. Here, for exploratory analysis, we increase ϵ beyond the range specified in our experiments described in the main text. We can see that for both FGSM and PGD, the noise is visible in the mel-spectrogram for $\epsilon = 0.002$. The signal becomes extremely noisy for $\epsilon = 0.2$ (SNR drops below -10 dB, and PESQ score < 1.5). On the other hand, for Carlini l_∞ attack, the noise is almost invisible at $\epsilon = 0.002$ and $\epsilon = 0.02$, as also evident from the high SNR values and PESQ scores. The noise becomes somewhat visible at $\epsilon = 0.2$ where the SNR drops to 10 dB, and the PESQ score becomes ~ 3.4.

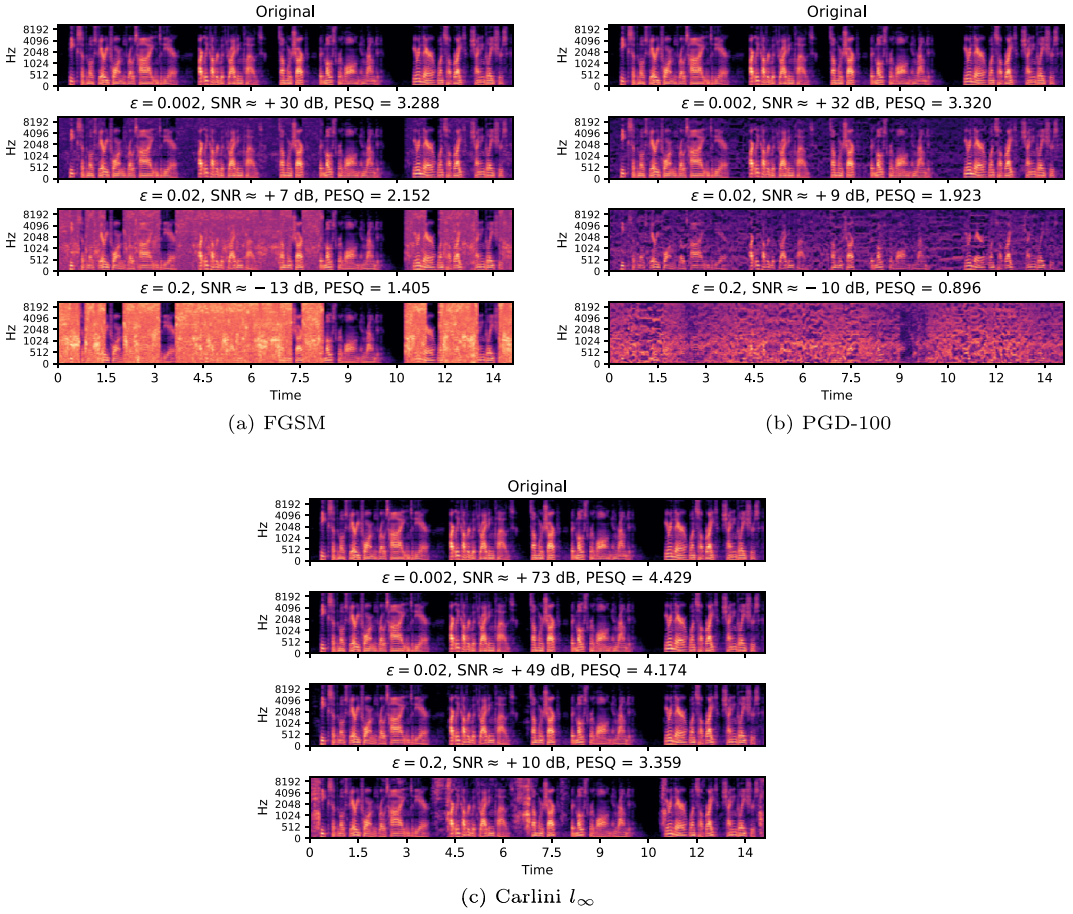


Fig. 4. Mel-spectrograms of an original utterance and its perturbed versions under different l_∞ attacks at varying strengths.

Appendix B. Similarity in misclassification for different attacks

We investigate whether different attack algorithms force the model to misclassify a particular input utterance as the same (wrong) speaker. This could possibly reveal similarity between different attack algorithms. Fig. 5 shows the fraction of similarity (*i.e.*, average number of matches) between the wrong predictions made by the model for different attacks. As evident, wrong predictions for Carlini l_∞ and Carlini l_2 attacks are very similar ($> 90\%$ similarity for all the ϵ values), possibly because the inherent strategy of the Carlini attack remains the same in the two variants. The similarity between FGSM and the two Carlini attacks is also noticeable. More interestingly, the similarity scores tend to decrease when ϵ increases. We hypothesize that a low ϵ constrains the attack algorithm with a smaller space for perturbation, and hence, the model generally tends to wrongly predict the

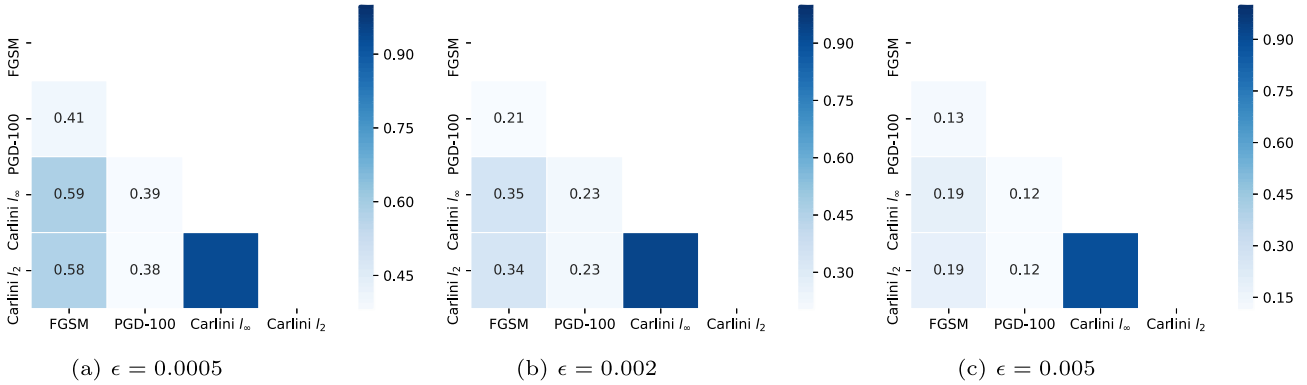


Fig. 5. Similarity (on a scale of [0,1]) between wrong predictions made by the model for different attacks.

closest class (one that causes the most confusion). On the other hand, a high ϵ opens up a lot more allowed space for the perturbation, and hence, the similarity between the wrong predictions tends to decrease.

Appendix C. Network architecture

The employed 1D CNN model is shown in Table 5.

Table 5
The network architecture of the CNN model we used in this paper. It takes waveform as input and predicts class logits. In all of the CNNs, the number of padded points (two-sided) is $\frac{k}{2}$.

audio waveform $\boldsymbol{x} \in [-1, 1]^T$
DSP: $[-1, 1]^T \rightarrow \mathscr{R}^{32 \times T'}$
Conv1D(32, 64, k=3) BatchNorm, ReLU; MaxPool1D(2)
Conv1D(64, 128, k=3); BatchNorm, ReLU
Conv1D(128, 128, k=3); BatchNorm, ReLU
Conv1D(128, 128, k=3); BatchNorm, ReLU; MaxPool1D(2)
Conv1D(128, 128, k=3); BatchNorm, ReLU
Conv1D(128, 64, k=3); BatchNorm, ReLU, MaxPool1D(2)
Conv1D(64, 32, k=3), BatchNorm, ReLU
MaxPool1D over time (the output is treated as speaker embedding)
FullyConnected(32, 251)

References

Athalye, A., Carlini, N., Wagner, D.A., 2018. Obfuscated gradients give a false sense of security: circumventing defenses to adversarial examples. In: Proceedings of the ICML, pp. 274–283.

Audhkhasi, K., Ramabhadran, B., Saon, G., Picheny, M., Nahamoo, D., 2017. Direct acoustics-to-word models for english conversational speech recognition. In: Proceedings of the Interspeech 2017, pp. 959–963.

Becker, T., Jessen, M., Grigoras, C., 2008. Forensic speaker verification using formant features and gaussian mixture models. Ninth Annual Conference of the International Speech Communication Association.

Biggio, B., Corona, I., Maiorca, D., Nelson, B., Šrndić, N., Laskov, P., Giacinto, G., Roli, F., 2013. Evasion attacks against machine learning at test time. In: Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases. Springer, pp. 387–402.

Bone, D., Lee, C.-C., Chaspari, T., Gibson, J., Narayanan, S., 2017. Signal processing and machine learning for mental health research and clinical applications. IEEE Signal Process. Mag. 34 (5), 189–196.

Carlini, N., Athalye, A., Papernot, N., Brendel, W., Rauber, J., Tsipras, D., Goodfellow, I., Madry, A., Kurakin, A., 2019. On evaluating adversarial robustness. arXiv:1902.06705.

Carlini, N., Wagner, D., 2017. Towards evaluating the robustness of neural networks. In: Proceedings of the IEEE symposium on security and privacy (sp). IEEE, pp. 39–57.

Carlini, N., Wagner, D., 2018. Audio adversarial examples: targeted attacks on speech-to-text. In: Proceedings of the IEEE Security and Privacy Workshops (SPW). IEEE, pp. 1–7.

Chan, W., Jaitly, N., Le, Q., Vinyals, O., 2016. Listen, attend and spell: a neural network for large vocabulary conversational speech recognition. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 4960–4964.

Chen, G., Chen, S., Fan, L., Du, X., Zhao, Z., Song, F., Liu, Y., 2019a. Who is real bob? adversarial attacks on speaker recognition systems. arXiv:1911.01840.

Chen, G., Chen, S., Fan, L., Du, X., Zhao, Z., Song, F., Liu, Y., 2019. Who is real bob? Adversarial attacks on speaker recognition systems. CoRR.abs/1911.01840

Chung, J.S., Nagrani, A., Zisserman, A., 2018. Voxceleb2: deep speaker recognition. In: Proceedings of the Interspeech, pp. 1086–1090.

Dehak, N., Kenny, P.J., Dehak, R., Dumouchel, P., Ouellet, P., 2010. Front-end factor analysis for speaker verification. IEEE Trans. Audio Speech Lang. Process. 19 (4), 788–798.

Goodfellow, I., Shlens, J., Szegedy, C., 2015. Explaining and harnessing adversarial examples. In: Proceedings of the International Conference on Learning Representations.

Hansen, J.H.L., Hasan, T., 2015. Speaker recognition by machines and humans: a tutorial review. IEEE Signal Process Mag. 32 (6), 74–99.

Huang, C.-W., Narayanan, S., 2017. Deep convolutional recurrent neural network with attention mechanism for robust speech emotion recognition. In: Proceedings of the IEEE International Conference on Multimedia and Expo (ICME). IEEE, pp. 583–588.

Jati, A., Georgiou, P., 2019. Neural predictive coding using convolutional neural networks toward unsupervised learning of speaker characteristics. IEEE/ACM Trans. Audio Speech Lang. Process. 27 (10), 1577–1589.

Jati, A., Peri, R., Pal, M., Park, T.J., Kumar, N., Travadi, R., Georgiou, P.G., Narayanan, S., 2019. Multi-task discriminative training of hybrid DNN-TVM model for speaker verification with noisy and far-field speech. In: Proceedings of the Interspeech, pp. 2463–2467.

Kingma, D. P., Ba, J., 2014. Adam: a method for stochastic optimization. arXiv:1412.6980.

Kreuk, F., Adi, Y., Cisse, M., Keshet, J., 2018. Fooling end-to-end speaker verification with adversarial examples. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 1962–1966.

Li, X., Zhong, J., Wu, X., Yu, J., Liu, X., Meng, H., 2020. Adversarial attacks on GMM i-vector based speaker verification systems. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 6579–6583.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., Vladu, A., 2018. Towards deep learning models resistant to adversarial attacks. In: Proceedings of the International Conference on Learning Representations.

Mao, C., Zhong, Z., Yang, J., Vondrick, C., Ray, B., 2019. Metric learning for adversarial robustness. Advances in Neural Information Processing Systems, pp. 480–491.

Miyato, T., Maeda, S.-i., Koyama, M., Nakae, K., Ishii, S., 2015. Distributional smoothing with virtual adversarial training. arXiv:1507.00677.

Nagrani, A., Chung, J.S., Zisserman, A., 2017. Voxceleb: a large-scale speaker identification dataset. In: Proceedings of the Interspeech, pp. 2616–2620. https://doi.org/10.21437/Interspeech.2017-950.

Narayanan, S., Georgiou, P.G., 2013. Behavioral signal processing: deriving human behavioral informatics from speech and language. Proc. IEEE 101 (5), 1203–1233.

Nicolae, M.-I., Sinn, M., Tran, M.N., Buesser, B., Rawat, A., Wistuba, M., Zantedeschi, V., Baracaldo, N., Chen, B., Ludwig, H., Molloy, I., Edwards, B., 2018. Adversarial robustness toolbox v1.2.0. CoRR.1807.01069

Nucci, A., Keralapura, R., 2012. Hierarchical real-time speaker recognition for biometric voIP verification and targeting. US Patent 8,160,877.

Panayotov, V., Chen, G., Povey, D., Khudanpur, S., 2015. Librispeech: an ASR corpus based on public domain audio books. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 5206–5210.

Papernot, N., McDaniel, P., Goodfellow, I., 2016a. Transferability in machine learning: from phenomena to black-box attacks using adversarial samples. arXiv:1605.07277.

Papernot, N., McDaniel, P., Jha, S., Fredrikson, M., Celik, Z.B., Swami, A., 2016. The limitations of deep learning in adversarial settings. In: Proceedings of the IEEE European Symposium on Security and Privacy (EuroS&P). IEEE, pp. 372–387.

Qin, Y., Carlini, N., Cottrell, G., Goodfellow, I., Raffel, C., 2019. Imperceptible, robust, and targeted adversarial examples for automatic speech recognition. In: Proceedings of the International Conference on Machine Learning, pp. 5231–5240.

Recommendation, I.-T., 2001. Perceptual evaluation of speech quality (PESQ): an objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs. Rec. ITU-T P. 862.

Reynolds, D.A., Quatieri, T.F., Dunn, R.B., 2000. Speaker verification using adapted gaussian mixture models. Digit. Signal Process. 10 (1–3), 19–41.

Rix, A.W., Beerends, J.G., Hollier, M.P., Hekstra, A.P., 2001. Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 01CH37221), 2. IEEE, pp. 749–752.

Schroff, F., Kalenichenko, D., Philbin, J., 2015. Facenet: A unified embedding for face recognition and clustering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 815–823.

Snyder, D., Garcia-Romero, D., Povey, D., Khudanpur, S., 2017. Deep neural network embeddings for text-independent speaker verification. In: Proceedings of the Interspeech, pp. 999–1003.

Snyder, D., Garcia-Romero, D., Sell, G., Povey, D., Khudanpur, S., 2018. X-vectors: robust dnn embeddings for speaker recognition. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 5329–5333.

Speakman, S., Sridharan, S., Remy, S., Weldemariam, K., McFowland, E., 2018. Subset scanning over neural network activations. arXiv:1810.08676.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., Fergus, R., 2013. Intriguing properties of neural networks. arXiv:1312.6199.

Terjék, D., 2020. Adversarial Lipschitz regularization. In: Proceedings of the 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia. https://openreview.net/forum?id=Bke_DertPB. OpenReview.net.

Tramer, F., Carlini, N., Brendel, W., Madry, A., 2020. On adaptive attacks to adversarial example defenses. arXiv:2002.08347.

Wan, L., Wang, Q., Papir, A., Moreno, I.L., 2018. Generalized end-to-end loss for speaker verification. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 4879–4883.

Wang, Q., Guo, P., Sun, S., Xie, L., Hansen, J.H.L., 2019. Adversarial regularization for end-to-end robust speaker verification. In: Proceedings of the Interspeech, pp. 4010–4014.

Zhong, Y., Deng, W., 2019. Adversarial learning with margin-based triplet embedding regularization. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 6549–6558.