DeepAdversaryDefense: A Deep Model to Identify and Prevent Adversarial Attacks against Medical Speech Recognition

Kirtee Panwar[a], Akansha Singh[a] and Krishna Kant Singh[b]

[a] *SCSET, Bennett University, Greater Noida, India*
[b] *ASET, Amity University, Noida, India*

Abstract
Deep learning models have made significant progress in safety-critical environments such as health-care systems, machine-learning based robots, Autonomous Intelligent Vehicles (AIV)), aviation software, etc. Deep Learning models can learn from input data, the property of learning has its own drawbacks as these models can be easily affected by minor disturbances in input examples. These input examples are generally created purposely by attackers and are known as adversarial examples. A small malicious change in input can cause the model to generate incorrect output. Majority of works in literature are towards understanding and generation of adversarial attack. Most of these attacks do not effectively resist detection networks. On the other hand, adversarial example detectors have inadequate evaluation. In this paper, a secure medical speech Recognition (MSR) system is proposed that can prevent malicious attacks. Adversarial examples that pose security concerns can be detected and filtered out. With the proposed model such system-malicious inputs designed to perform an attack on safety-critical applications, even if the adversary has no access to the underlying model are prevented.

Keywords 1
Adversarial samples, Automatic Speech Recognition (ASR), Medical Speech Recognition (MSR), vulnerabilities, deep learning, security, Adversarial Loss.

1. Introduction

The use of Deep learning models has become a part of our daily life: from organizing our searches to social media feeds. Recent advances in Automatic speech recognition, machine learning and deep learning technologies have favored the advancement of speech-based conversational interfaces. This has further led to an increase in the interaction of such devices with various machine-critical applications. Machine learning-based speech recognition systems allow users to carry out essential and crucial activities for either industrial development and processes or assisted living using voice commands [1]. With the advancement in deep learning-based speech recognition systems and interfaces based for essential applications such as recognizing the transcription of medical speech in healthcare, etc., new attacks are developed also known as adversarial attacks.

Deep models can achieve acceptable accuracy levels but have been found to make mistakes more often. In literature, it can be observed that these models are vulnerable to well-designed input attacks known as adversarial examples. These inputs make the model generate incorrect output with high confidence. For example, attackers generate adversarial inputs to automatic speech recognition models through sound sensors to obtain desired target output. The model outputs un-favorable results with such malicious inputs. There are various applications of automatic speech recognition that demands security against such vulnerabilities such as Microsoft Cortana, Amazon Alexa, and Apple Siri [2]. The adversarial inputs differ very slightly from the actual inputs drawn from a certain data

IDDM-2022: 5th International Conference on Informatics & Data-Driven Medicine, November 18–20, 2022, Lyon, France
EMAIL: Kirtee.panwar@bennett.edu.in (A. 1); akanshasing@gmail.com (A. 2); krishnaiitr2011@gmail.com (A. 3)
ORCID: 0000-0001-9842-1804 (A. 1); 0000-0002-5520-8066 (A. 2); 0000-0002-6510-6768 (A. 3)

distribution that has the power to make machine misclassify examples [3] irrespective of model architecture and datasets used for training purposes thus exposing blind spots of training algorithms.

The primary cause of vulnerability of the machine learning model to adversarial attack is their linear behavior in high dimensional space [4]. This theory further leads to new methods of generating adversarial inputs for adversarial training for security enhancement purposes. In domains such as traffic control, manufacturing, advanced automotive systems, the adversarial inputs have substantial dependencies on each other which can be represented with features for non-uniform disturbances generated at the output of the machine-learning model during adversarial training [2].

With adequate analysis of inputs, it is possible to classify certain input examples as adversarial examples by identifying rules between attacker and defender based on practical scenarios [5].

One of the possible ways to reduce the vulnerability of models is to enhance scalability of the model against adversarial inputs [6]. In such cases, network-based detectors play a fundamental role in validating the security of the model. Attacks designed for distribution-based-detectors for validating the security of such detectors is critical for security-related applications [7]. Design of adversarial samples [8] that can reduce the detection rate of distribution-based detection techniques help in understanding the underlying problem with security against adversarial attacks.

Majority of works in literature are towards understanding and creating an adversarial attack. However, the attacks do not effectively resist detection networks [9]. On the other hand, adversarial example detectors have inadequate evaluation [10]. There remains a research gap in understanding the construction of adversarial examples which conflicts with the safety requirements of the ASR systems required for safety-critical applications [11].

Adversarial Attacks can be Untargeted, in which the objective of the attack is to degrade the network's performance, or targeted [12], where the aim is to make the model predict the target transcript. Adversarial attacks based on CTC(Connectionist Temporal Classification) loss function [13] or task loss of the problem, e.g., Houdini Attack [14] degrades the model's performance. Adversarial attacks include imperceptible attacks [15] in which the transcripts are hidden, Fast Gradient Sign Method (FGSM) and Project gradient Descent (PGD) attacks [4], based on generating adversarial examples that degrade the performance of the optimization process of loss function of the model.

Pre-processing defenses or adversarial training alleviates the effect of adversarial attacks. Pre-processing defenses such as randomized smoothing, WaveGAN vocoder, variational auto-encoder (VAE), etc., eliminate the disturbances caused by adversarial examples before it enters the ASR system but is ineffective against adaptive attacks [16]. On the other hand, in defenses based on adversarial training, such as FGSM adversarial training and PGD adversarial training, the system's robustness against attack is limited and requires critical tuning of parameters. The training time required for the model is another limiting factor [17].

In the proposed ASR model, the loss function incorporates Frequency domain power spectrum verification and distance between voice propagation angles as defenses to alleviate the effect of adversarial examples.

The remainder of this paper is organized as follows. Section 2 describes the proposed methodology for developing a secure ASR system based on deep learning approach. Section 3 presents experimental results and comparison with other state-of-art techniques. The paper is finally concluded in section 5 with some directions for future research.

2. Proposed Methodology

The ASR module is used in the voice control system to enable humans to interact with machines. These modules are vulnerable to adversarial voice commands that cause the system to generate undefined output. For instance, a target transcription by attacker hid within audio file below a certain threshold and imperceptible by humans but not machines can be interpreted as commands by the sensors. The medical ASR system accepts sounds from multiple speakers such as mobile phones, passengers, etc. Malicious messages may come from any of these speakers, creating negligibly perceptible changes to obtain a desired target output form ASR module. Another possible attack is the black box attack, where attacker has limited knowledge about the ASR model parameters. Such attack

is possible using keyword recognition of the ASR system that is accessible. For instance, commercial medical ASR systems use keywords, “I need urgent medical help”. Modification of such keywords by attackers can lead to desired target output.

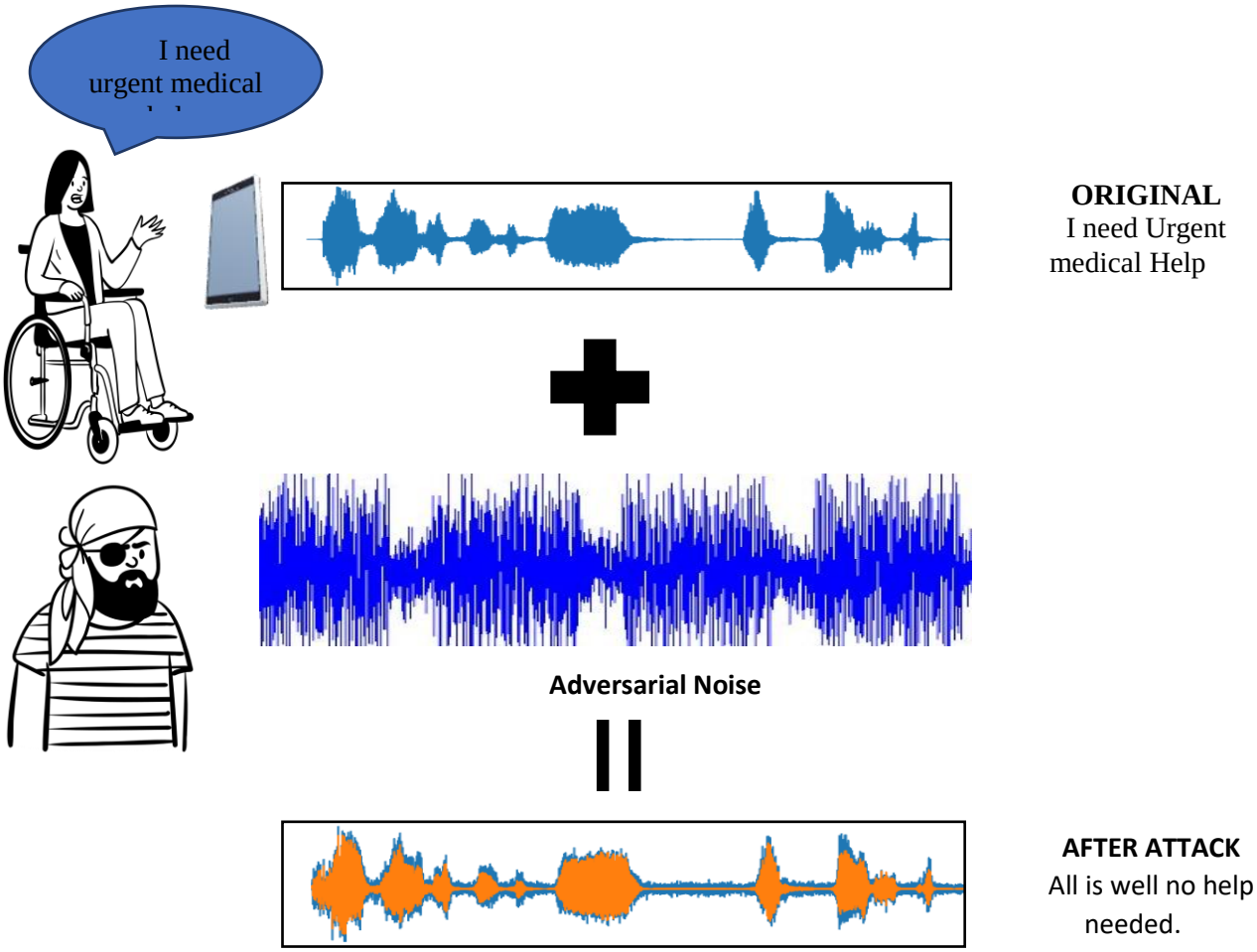

Figure 1. Adversarial attack example.

Medical ASR systems are used for critical applications and have a high demand for security against adversarial attacks that can come from various sources such as a noisy environment, loudspeakers, mobile phones, etc. For ASR systems, knowledge about adversarial perturbations is limited and it is a challenge to defend medical ASR systems against such attacks. In this paper, a deep learning model is proposed that can resist adversarial attacks. The proposed ASR system comprises of Deep neural network and transcript generation.

2.1. Deep Neural Network: DeepAdversaryDefense

With help of proposed network feature extraction and classification is performed simultaneously. The proposed deep neural network is designed extract features in human auditory system as well as to classify malicious messages and bypass original messages. Classification is performed in the feature transformation network by incorporating frequency verification and speaker verification in the loss function of the network. The loss functions incorporate:

a) Frequency domain power spectrum verification.
b) Distance between voice propagation angles.

The Proposed ASR system comprises of Feature Transformation network, Feature Decoder Network and Discriminator network. The architecture is similar to [19]. The block diagram of proposed model is given in Fig. 2.

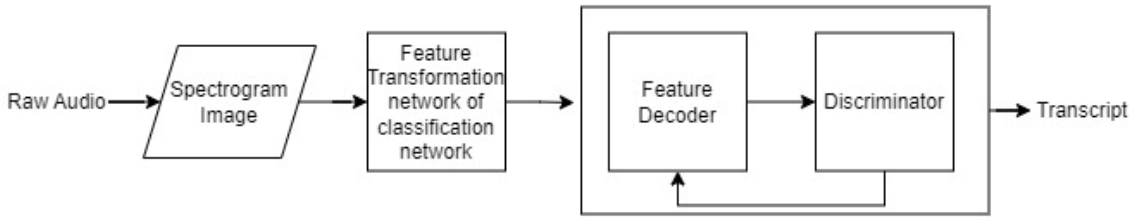

Figure 2. Block Diagram of Proposed Model

Feature transformation network Input signal is represented in the form of Mel spectrogram [20]. These features resemble a 2D image. This network consists of 1 block of convolution followed by ReLU activation function. Here the frequency domain verification and identification of driver's voice is performed. The architecture of this network is given in Table 1. In the first layer, B feature blocks of spectrogram image is obtained, these features are passed if frequency check and distance check is passed otherwise no features are passed further. The size of $B = \sqrt{N}$, where $N \times N$ is the size of spectrogram image. These features are combined with pixel shuffle layer [21] and reshaped. Here, the pixel shuffle layer has been introduced for computational efficiency.

Table 1. Architecture of feature transformation network

Layer	Kernel Size	Normalization	Activation Function
Input	B x B	Spectral, Batch	LReLU
Pixel Shuffle Layer			

Feature decoder network We use 1 convolution block with LReLu activation function for down-sampling. The decoder then converts the features to transcript. We use 1 convolution block with LReLu activation function for down-sampling sampling of features followed by 5 blocks of Resnet these features from resnet blocks are concatenated and passed to the convolution block with LReLu activation followed by convolution layer with tanh activation. Each ResNet Block consists of 2 layers of CNN Layer followed by Normalisation technique. Each of the features obtained from subsequent ResNet blocks are concatenated. The architecture of feature decoder network is given in Table 2.

Table 2. Architecture of Feature Decoder Network

Layer	Kernel Size	Normalization	Activation Function
Input	3 x 3	Spectral, Batch	LReLU
ResBlocks (5 times)	5 x 3	Spectral, Batch	LReLU
TransposeConvolution	3 x 3	Spectral, Batch	LReLU
Convolution	3 x 3	Spectral, Batch	Tanh

Discriminator Network

The discriminator network classifies the signal as malicious or original based on frequency domain power spectrum and distance between voice propagation angles. The discriminator network guides the network to create realistic transcripts for given input. The architecture of discriminator network consists of 6 layers on convolution with spectral normalization followed by self-attention [22] then convolution then self-attention layer. The output transcripts generated are meaningful due to the contextual information captured with help of self-attention layer. The final layer is sigmoid activation. The architecture of network is given in Table 3.

Table 3. Architecture of Discriminator Network

Layer	Kernel Size	Normalization	Activation Function
Input	3 x 3	Spectral	LReLU
Convolution	3 x 3	Spectral	LReLU
Convolution	5 x 3	Spectral	LReLU
Convolution	3 x 3	Spectral	LReLU
Convolution	5 x 3	Spectral	LReLU
Convolution	3 x 3	Spectral	LReLU
Self Attention Layer	-	-	-
Convolution	3 x 3	Spectral	LReLU
Self Attention Layer	-	-	-
Flatten	-	-	-
Linear	-	-	-

Loss Function:

We use Connectionist Temporal Classification (CTC) loss for feature transformation network. to produce sequential sequences for un-aligned input data. The input to the loss function is the output probability distribution y and the objective is to maximize the probability of outputting that correct transcript i.e., minimising maximum likelihood training [15]. The network then converts the input into the highest probability transcription. The loss function for spectrum detection is defined as

$$L1=O^{ML}(S,N_w)= - \sum_{(x,z)\in S} \ln(p(z/x)) \quad (1)$$

Where, S denotes set of input samples that belong to a fixed distribution D_{Xxz}, $z = (z_1, z_2, \dots z_U)$ is the target sequence whose length is smaller than or equal to the input sequence $x = (x_1, x_2, \dots x_T)$ and N_w denotes network outputs.

The feature transformation network must not pass features if Frequency domain power spectrum is not verified and distance between voice propagation angles is not within certain range. Therefore, this constrain is added to loss of the network. If realistic output is generated for malicious input, then network is penalized. The loss function L2 is defined to perform frequency domain verification which is defined as the ratio of low frequency power to total power as

$$L2 = \frac{\sum_{f=85}^{2kHz} A^2(f)}{\sum_f A^2(f)} \quad (2)$$

For original message the range is pre-defined and the range for human voice is 85 Hz to 4 kHz. If ratio R_1 is within acceptable range, then realistic transcripts are generated. A Spectrogram is like a 2D image of a signal with the time on the x-axis and frequency on the y-axis. Therefore, f represents value on the y-axis.

For identification of driver's voice, the angle of propagation of sound signal is analysed, if the angle is within a predefined range, then the features are passed on further to next layer. The loss function for identifying driver's voice is defined as

$$L3 = \alpha = \arccos\left(\Delta N.\frac{V_0}{D_0.f_s}\right) \quad (3)$$

Where ΔN length of segment, D_0 denotes distance between speakers installed.

Adversarial loss of network is min_max(L1). The total loss of the network is $advloss(L1) + L2 + L3$

3. Experimental Results

The experiments are performed using standard dataset TIMIT for training the proposed model. This dataset consists of audio aligned with each character as well as expected sentence transcription. We check the effectiveness of the proposed model to resist fast gradient-sign method (FGSM) attack [16]. Word error rate (WER(%)) is used to measure the speech to text accuracy and is calculated using Levenshtein distance [19] to determine the resistance of the proposed model against attacks using equation

$$WER = \frac{D+I+S}{N}, \qquad (4)$$

where *D* denotes deletion of words, *I* denotes insertion of words, S denotes substitution of words by the model and N denotes total number of words. Results are compared with other state of art methods and displayed in Table 4.

Table 4. WER(%) of proposed model against FGSM attack.

Method	FGSM	Average
Proposed	33.5	33.5
[23]	-	32.6
[24]	-	40.5
[25]	34.6	34.6

From Table 4. WER of the proposed model is 33.5 % for TIMIT dataset against FGSM attack. The proposed model generates valid transcription with low WER as compared to [19]. A low value of WER indicates good performance against attacks. Comparison is done on the bases of WER with [18] and values are similar, results show that WER values of proposed model is justified.

4. Conclusion and Future Scope

In this paper, we propose a secure medical speech recognition system to defeat adversarial voice command attacks in healthcare applications. We utilize the physical attributes of voices to distinguish the speaker's voice from other adversarial voices in two steps. First, multi-source signals are filtered out according to frequency domain power verification spectrum. Second, the driver's voice is determined from its propagation direction multiple microphones installed at different corners. The feature decoder network and discriminator network then use CTC loss function to transform the features and generate transcripts. Detection of adversarial examples during the ASR model's training enhances the model's scalability. The experimental results show improved WER for proposed ASR system as compared to other systems when introduced to FGSM attack.

As future work, the proposed model can be tested against various other attack models and accordingly classification network can be trained with appropriate loss functions. Some new adversarial attacks can be defined for ASR systems for adversarial training.

5. References

[1] Huynh, N.D., Bouadjenek, M.R., Razzak, I., Lee, K., Arora, C., Hassani, A. and Zaslavsky, A. ``Adversarial Attacks on Speech Recognition Systems for Mission-Critical Applications: A Survey" (2022). arXiv preprint arXiv:2202.10594.

[2] Erdemir, E., Bickford, J., Melis, L. and Aydore, S. ``Adversarial robustness with non-uniform perturbations". *Advances in Neural Information Processing Systems*, 34 (2021), pp.19147-19159.

[3] Szegedy, Christian, Zaremba, Wojciech, Sutskever, Ilya, Bruna, Joan, Erhan, Dumitru, Goodfellow, Ian J., and Fergus, Rob. ``Intriguing properties of neural networks". ICLR, abs/1312.6199, 2014b, 2013. URL http: //arxiv.org/abs/1312.6199.

[4] Goodfellow, I.J., Shlens, J. and Szegedy, C. ``Explaining and harnessing adversarial examples", arXiv preprint arXiv:1412.6572 (2014).

[5] Gilmer, J., Adams, R.P., Goodfellow, I., Andersen, D. and Dahl, G.E. ``Motivating the rules of the game for adversarial example research", arXiv preprint arXiv:1807.06732 (2018).

[6] Zhang, J. and Li, C. ``Adversarial examples: Opportunities and challenges. IEEE transactions on neural networks and learning systems", 31(7), 2019, pp.2578-2593.
[7] Yigitcan Kaya, Muhammad Bilal Zafar, Sergul Aydore, Nathalie Rauschmayr, Krishnaram Kenthapadi, "Generating Distributional Adversarial Examples to Evade Statistical Detectors", *Proceedings of the 39 th International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR* 162, 2022
[8] Wang, K., Yi, P., Zou, F. and Wu, Y. ``Generating Adversarial Samples With Constrained Wasserstein Distance", *IEEE Access*, 7, 2019, pp.136812-136821.
[9] T. Tanay and L. Griffin, ''A boundary tilting persepective on the phenomenon of adversarial examples,'' arXiv:1608.07690, 2016, URL: https://arxiv.org/abs/1608.07690
[10] Yuan, X., He, P., Zhu, Q. and Li, X. ``Adversarial examples: Attacks and defenses for deep learning", *IEEE transactions on neural networks and learning systems*, 30(9), 2019, pp.2805-2824.
[11] Serban, A., Poll, E. and Visser, J. ``Adversarial examples on object recognition: A comprehensive survey". *ACM Computing Surveys (CSUR)*, 53(3), 2020, pp.1-38.
[12] Carlini N, Mishra P, Vaidya T, Zhang Y, Sherr M, Shields C, Wagner D, Zhou W. ``Hidden voice commands". *25th USENIX security symposium (USENIX security 16)* 2016 (pp. 513-530).
[13] Amodei D, Ananthanarayanan S, Anubhai R, Bai J, Battenberg E, Case C, Casper J, Catanzaro B, Cheng Q, Chen G, Chen J. ``Deep speech 2: End-to-end speech recognition in english and mandarin". *International conference on machine learning*, 2016 Jun 11 (pp. 173-182). PMLR.
[14] Cisse M, Adi Y, Neverova N, Keshet J. ``Houdini: Fooling deep structured prediction models". arXiv preprint arXiv:1707.05373. 2017 Jul 17.
[15] Carlini N, Wagner D. ``Audio adversarial examples: Targeted attacks on speech-to-text". In 2018 IEEE security and privacy workshops (SPW) 2018 May 24 (pp. 1-7). IEEE.
[16] Schönherr L, Kohls K, Zeiler S, Holz T, Kolossa D. ``Adversarial attacks against automatic speech recognition systems via psychoacoustic hiding". arXiv preprint arXiv:1808.05665. 2018 Aug 16.
[17] Joshi S, Villalba J, Żelasko P, Moro-Velázquez L, Dehak N. ``Study of pre-processing defenses against adversarial attacks on state-of-the-art speaker recognition systems". *IEEE Transactions on Information Forensics and Security*. 2021 Sep 29;16:4811-26.
[18] Madry A, Makelov A, Schmidt L, Tsipras D, Vladu A. ``Towards deep learning models resistant to adversarial attacks". arXiv preprint arXiv:1706.06083. 2017 Jun 19.
[19] Madono, K., Tanaka, M., Onishi, M. and Ogawa, T. ``Sia-gan: Scrambling inversion attack using generative adversarial network*", IEEE Access*, 9, 2021, pp.129385-129393
[20] Graves, A., Fernández, S., Gomez, F. and Schmidhuber, J. ``Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks". *In Proceedings of the 23rd international conference on Machine learning* (pp. 369-376), 2006.
[21] Shi, W., Caballero, J., Huszár, F., Totz, J., Aitken, A.P., Bishop, R., Rueckert, D. and Wang, Z. ``Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network". In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1874-1883), 2016.
[22] Shaw, P., Uszkoreit, J. and Vaswani, A. ``Self-attention with relative position representations". *arXiv preprint arXiv:1803.02155*, 2018.
[23] Navarro, G. ``A guided tour to approximate string matching. *ACM computing surveys (CSUR)", 33*(1), 2001, pp.31-88.
[24] Żelasko, P., Joshi, S., Shao, Y., Villalba, J., Trmal, J., Dehak, N. and Khudanpur, S. ``Adversarial attacks and defenses for speech recognition systems". *arXiv preprint arXiv:2103.17122,* 2021.
[25] Yoshioka, T. and Gales, M.J. ``Environmentally robust ASR front-end for deep neural network acoustic models", *Computer Speech & Language*, *31*(1), 2015, pp.65-86.