Experimental & Comparative Study of Adversarial Attacks on Automatic Speech Recognition Systems

Anunya Sharma, Kiran Malik, Poonam Bansal

Indira Gandhi Delhi Technical University for Women, Kashmere Gate, New Delhi, India

Corresponding author: Anunya Sharma, Email: anunya086btcsai21@igdtuw.ac.in

Automatic Speech Recognition Systems are an integral part of modern-day artificial intelligence based machines due to advancement in smart technology. Although ASRSs have provided users with the ability to perform important tasks through voice commands, their ability to withstand adversarial attacks is uncertain. The primary objective of adversarial artificial intelligence techniques is to disrupt the functioning of machine learning models with misleading data, by exploiting their vulnerabilities in decision-making process, causing them to make misclassifications or incorrect predictions. This poses a significant threat in the realm of AI and machine learning research, particularly in areas where machine performance is crucial. Therefore, it becomes imperative to study adversarial attacks, with a particular focus on assessing and comparing their severity. This would allow distinguishment between highly detrimental attacks and those with lesser impact, enabling the development of tailored defense strategies to effectively safeguard against them. This paper encompasses an experimental and comparative study of intensity of various attack methods on wav2vec2 model from the Torchaudio hub, particularly: Fast Gradient Sign Method, Basic Iterative Method, Projected Gradient Descent, Carlini and Wagner, and Imperceptible CW attack.

Keywords: Artificial Intelligence (AI), Machine Learning (ML), Speech Recognition Systems (SRSs), Automatic Speech Recognition Systems (ASRs), Adversarial Examples, Adversarial Attacks.

in: *Advancements in Communication and Systems*. Ed. by Ashish Kumar Tripathi and Vivek Shrivastava. Computing and Intelligent Systems, SCRS, India., 2023, pp. 349–358. DOI: https://doi.org/10.56155/978-81-955020-7-3-30

1. Introduction

The field of speech recognition, which identifies individuals based on their unique voice patterns, has garnered significant interest from both academic and industry circles due to its user-friendly remote-control capabilities and cost-effectiveness. The rapid progress of ASRSs can largely be attributed to advancements in NNs, particularly DNNs. Voice-activated assistants, like Alexa, Google Home, Apple Pod, and IoT devices, have revolutionised tasks such as appointment booking, contact management, call making, email sending, and home automation. While traditional SRSs using methods like i-vector and Gaussian Mixture Model have flourished over the years, NN-based approaches have gained prominence due to their superior capabilities. However, research has revealed that NNs are vulnerable to adversarial attacks. Recent studies have demonstrated successful and efficient manipulation of voice commands to deceive audio transcription frameworks, resulting in what are known as audio adversarial examples. These vulnerabilities pose a significant threat to society, considering the growing prominence of automation in daily life.

With the increase in utilisation of neural networks, it becomes imperative to investigate their behaviour in adversarial environments. While previous research on adversarial examples has predominantly focused on images and text, encompassing tasks such as image classification, image segmentation, face detection, text classification and malware detection, research on audio, particularly in the context of automatic speech recognition, remains relatively limited. Crafting targeted adversarial examples for SRSs and ASRSs has proven to be a challenging task. Targeted attacks are represented by hidden and inaudible speech commands, but they need to be synthesised from scratch and cannot alter existing audio recordings. In this paper, the primary focus is on exploring and studying the intensities of adversarial attack methods against automatic speech recognition systems to evaluate their effectiveness across diverse applications. The implementation of these attacks in this repository can be used to evaluate the robustness of ASR models and to develop defences against such attacks. This study aims to serve as a valuable resource for practitioners and researchers, providing insights into the challenges inherent in speech recognition models and facilitating advancements for various applications.

2. Related Work

The omnipresence of machine learning systems, especially ASRSs, in daily life calls for an analysis of their vulnerabilities and defences against exploitation of the same. Piotr Zelasko et al. [1] tested robustness of two vastly different ASRSs (DeepSpeech 2 and Espresso) in absence of any counter-measure confirmed their vulnerability to every adversarial attack. Counter-measures like randomized smoothing displayed limited effectiveness while WaveGAN reduced attack success rate considerably.

Developing a universally applicable ASR system is a challenging task especially because its vulnerability is language dependent. Karla Markert et al. [2] demonstrated that, as a probable consequence of specific correlation between spoken and written form of the English language, it is more susceptible to phoneme-based attacks compared to German language. The latter is easily fooled using CW attack and requires modifications to phonetic aspects to adequately hide the attack.

Owing to recent studies on adversarial attacks, it is now possible to build specific audio adversarial examples on ASRSs. Nicholas Carlini and David Wagner [3] demonstrated cent-percent conversion of audio waveforms into target transcriptions with the help of optimization-based attacks applied end-to-end, by addition of only a slight distortion.

The ever-increasing variety of attacks methods against SRSs necessitates the need for an evaluation criterion. Jiahe Lana et al. [4] put forward the same from three aspects: i. Practicability- evaluated by transferability, universality, attack media, distance and commercial SRSs; ii. Imperceptibility- evaluated on the basis of types of adversarial audio, perturbation norm, human perception and signal-to-noise ratio; iii. Effectiveness- evaluated by Generation time, recognition accuracy, equal error rate,

false-positive rate, and false-negative rate. They also listed out the evaluation criterion for defence methods under two aspects: i. Practicability- evaluated on the basis of generality, defence media and defendable attacks; ii. Effectiveness- evaluated on the basis of defence time, detection accuracy, recognition rate, equal error rate, false-positive rate and false-negative rate.

With the advancement in machine learning, SRSs boasts of a wide range of applications including mission-critical applications. Ngoc Dung Huynh et al. [5] presented an analysis of algorithms used in speech recognition, like Hidden Markov and Neural Network, followed by a detailed description of adversarial attack methods and defence strategies, in context of strategic and critical activities carried out by machines having conversational interfaces.

Applications for automatic voice assistants with computer assistance have become more prevalent, raising concerns about their security. Code modulation and audio compression were used by Jiajie Zhang et al. [6] to propose several defence strategies against targeted audio adversarial examples in the ASRSs. Its effectiveness was tested through thorough evaluation on natural dataset.

In spite of all the progress in this field, the intrinsic properties of adversarial examples are not well studied. Wei Zong et al. [7] presented a method to visualize various decision boundary patterns which differentiate between audio adversarial examples and unaffected audios. A clear distinction between them can be observed by decision boundary-based feature extraction with dimensionality reduction. They also added that anomaly detection maybe used to find previously unidentified audio adversarial examples.

Depending on the extent of access that the adversary has to the victim learning algorithm, adversarial attacks are categorised as either white-box or black-box attacks. Saeid Samizade et al. [8] viewed defence as a problem of classification and presented a method for consistently creating datasets of adversarial example.The defence strategies now in use focus on modifying input signals and observing speech recognizer behaviour. White-box attack is a gradient-based strategy on Baidu DeepSpeech with Mozilla Common Voice while black-box attack is a gradient-free strategy on deep model-based keyword detection system with Google Speech Command dataset. For known attacks, these datasets were utilised to train a Convolutional Neural Network model containing cepstral features to accurately detect and distinguish between normal and adversarial cases, however the performance dramatically declined for unknown attacks.

As a countermeasure against white-box adversarial attacks, Sonal Joshi et al. [9] proposed three defence strategies for K2 conformer hybrid ASRS: i. denoiser pre-processor, that is unable to prevent adaptive white-box attacks; ii. adversarially fine-tuning ASR model offers more robustness; iii. adversarially fine-tuning joint model of denoiser and ASRS, which makes use of frozen parameters, gives best resistance against projected gradient descent (PGD) attack method while non-static parameters worked well against fast gradient sign method (FGSM) attack.

SRSs have the advantage of not requiring physical presence in biometric-based user identification methods, unlike fingerprint and iris. Their increasing popularity in related domains, inspired Katharina Kohls et al. [10] to explore their vulnerabilities by designing almost imperceptible psychoacoustics-based attacks, that take into account dynamic human hearing thresholds, against Kaldi ASRS, a DNN-HMM system. By including audio-agnostic universal perturbation and modelling audio distortions induced by the physical over-the-air propagation, Yi Xie et al. [11] investigated the vulnerability of DNN-based SRS to adversarial attacks and achieved a high attack success rate of 90% on a dataset of 109 English speakers. Zhouhang Li et al. [12] achieved a higher success rate of 98% for digital attack and 50% for over-the-air attack on Xvector, a DNN based SRS, with the same dataset.

Xvector based SRSs are tested against common white-box adversarial attack techniques like basic iterative method, projected gradient descent, fast gradient sign methodand Carlini-Wagner attack. For

these attacks, Sonal Joshi et al. [13] investigated four pre-processing defences- randomized smoothing, DefenseGAN, Variational Autoencoder, Parallel WaveGAN vocoder (PWG)- that do not require training with adversarial examples and concluded that SRSs are most susceptible to BIM, PGD and CW attacks. PWG and randomised smoothing were combined to produce an accuracy of 93%, as opposed to 52% in an undefended system and an improvement of >90% against BIM attacks.

Yi Xie et al. [14] designed a real-time, robust and adaptive universal adversarial attack against DNN based SRSs in white-box environment by inducing an audio-agnostic universal perturbation. They estimated the room impulse response (RIR) to model sound distortions brought on by physical over-the-air propagation to increase resilience. Magnitude of perturbations were adaptively adjusted for each individual utterance using spectral grating. On a public dataset of 109 English speakers, this technique outperformed traditional non-universal attacks with a 90% average success rate on both d-vector and Xvector SRSs and a 100x speedup on attack launching time.

Adversarial attacks consider white-box setting mostly, but for the first time, in 2021, Guangke Chen et al. [15] did a comprehensive and systematic weakness analysis of SRSs in the practical black-box environment. To support this, they proposed FAKEBOB, an adversarial attack, that's demonstrated a success rate of 99% on both open-source and commercial systems. Interestingly, it rendered four promising defence methods ineffective.

Guangke Chen et al. [16] also noticed that many real-word attack scenarios were not considered due to the use of only a few variables, such as certain combinations of source and target speakers. To comprehend transferability among 14 different SRSs, they proposed AS2T, the first attack in this field that enables the attacker to create sounds utilising target speakers and arbitrary source [16]. Various transformation functions with different parameters were applied to generate adversarial voices in over-the-air transmission.

Jesús Villalbaet al. [17] made use of representation learning based on Xvector architectures to classify attacks- with respect to the signal-to-adversarial-noise ratio, threat model or attack algorithm- in the field of speaker identification, speaker verification and speech recognition, with accuracies as high as 90%. Their models could not generalize well to attack algorithms which consequently affected attack verification. It has been considered promising that they were able to identify unknown attacks with equal error rates of roughly 19%.

Significant progress has been made in the realm of adversarial attacks in the computer vision field, while speaker recognition lags behind. According to the analysis of the issue by Arindam Jatia et al. [18], an undefended model's performance fell from 94% to 0% under the most intense attacks (Carlini 12, PGD-100). Moreover, the adversarial examples are transferrable and can lead to black-box attacks. Despite being ineffectual against 12 attacks in the experiments undertaken, PGD-based training ended up being the best option for defence. Adding white Gaussian noise to training data was also proven to be ineffective.

The ability to create real-world adversarial attacks is hampered by the adversary's limited access to system information in the actual world. Selective Gradient Estimation Attack (SGEA), a novel and successful attack on ASRSs, was used by Qian Wang et al. [19] to take down the DeepSpeech system on the LibriSpeech and the Mozilla Common Voice datasets. SGEA requires only restricted access to the neural network's output probabilities and has a high success rate. Attack success rate is increased by SGEA from 35% to 98%.

A thorough assessment of the development of SRSs, including the mainstream frameworks of SRSs, types of adversarial attacks, attack detection strategies, perturbation constraints and objects, defence training methods, refactoring against existing attacks and few commonly used datasets, has been provided by Hao Tan et al. [20].

3. Automatic Speech Recognition Systems

Automatic Speech Recognition (ASR) systems are pivotal in present day technology landscape, enabling seamless voice interactions, transcription services and accessibility solutions. ASR systems employ a combination of acoustic and language models to process audio signals and transcribe them into textual representations. ASR systems find applications in various domains, including voice assistants, transcription services, call centres, and accessibility tools, where efficient and precise conversion of spoken language to text is essential. Moreover, they have paved the way for real-time language translation, breaking down language barriers and enabling global communication and collaboration. Despite their many benefits, ASR systems encounter significant challenges. They struggle to perform optimally in noisy environments, where background disturbances or sounds can reduce their accuracy. Variations in accents, dialects and regional speech patterns can also pose difficulties, impacting their ability to understand and transcribe spoken language accurately. Additionally, understanding contextual nuances and ambiguities can be an obstacle, leading to occasional misinterpretations. Privacy concerns have arisen due to the potential for voice data to be stored and potentially misused, underscoring the need for responsible data handling practices and transparency in ASR technology deployment. such as sensitivity to noise, contextual nuances and privacy concerns, necessitating ongoing advancements to enhance their robustness and security. Developing high performing ASR models often demands substantial computational resources and extensive training datasets, which can be a barrier for smaller organisations and researchers.

4. Wav2Vec2 Model

Wav2Vec2, a deep learning model for speech recognition and representation learning, was introduced by Facebook AI Research in 2020 as an extension of the original Wav2Vec model. It leverages self-supervised learning to learn powerful speech representations from large amounts of unlabelled audio data. The model is trained using a two-step process. In the first step, a masked prediction task is performed, where the model is trained to predict the masked sections of the audio waveform given the surrounding context. This helps the model learn robust representations that capture important acoustic and linguistic information. In the second step, a contrastive loss is used to fine-tune the representations obtained from the first step. By contrasting positive pairs (segments of the same audio) with negative pairs (segments of different audio), the model learns to map similar audio segments close together in the embedding space while pushing dissimilar segments apart.

Wav2Vec2 has shown cutting-edge performance on a variety of speech-related tasks, including automatic speech recognition, keyword spotting and speaker detection. Its advancements in representation learning have contributed to significant improvements in speech-related applications and have made it a popular choice in the speech processing research community.

5. Adversarial Attacks

Adversarial attacks in the context of machine learning have garnered attention due to their potential to undermine the security and reliability of machine learning systems. Two fundamental categories of adversarial attacks are targeted and untargeted attacks. The former is characterised by their specific and goal-oriented nature, where the adversary aims to manipulate the output of a model to achieve a particular result. They require a higher level of sophistication from the attacker. Whereas, the latter are more general in their approach, focused on disrupting the performance of a model without any specific objective. This paper presents five methods of estimating additive noise, all of which are white-box attacks, as theyneed to be aware of the parameters and architecture of target model to optimize the objective function and generate adversarial examples.

5.1 Fast Gradient Sign Method (FGSM)

The FGSM attack is an adversarial technique commonly used to generate deceptive examples for neural networks and serves as a computationally efficient method for evaluating model robustness and uncovering vulnerabilities. Introduced by Goodfellow et al. in 2015, this attack exploits the model's gradient information to determine the optimal direction for perturbing the input data. By taking a step in the sign direction of the gradient and scaling it by a small value, the FGSM attack produces adversarial examples that maximise the model's prediction error while minimising the perturbation.

5.2 Basic Iterative Method (BIM)

The BIM attack is an iterative variation of the above mentioned Fast Gradient Sign Method (FGSM) attack, commonly used to create adversarial examples in deep learning models. Introduced by Kurakin et al. in 2016, BIM improves upon FGSM by iteratively applying small perturbations to the input data. BIM calculates the gradient of the loss or the cost function with respect to the input with each iteration and gradually amplifies the perturbation. This iterative approach enables BIM to generate more potent and effective adversarial examples compared to FGSM, bypassing defence mechanisms and enhancing model vulnerability assessment.

5.3 Projected Gradient Descent (PGD)

The PGD attack is an iterative optimization technique commonly used in generating adversarial examples for deep learning models. Introduced as an extension of the BIM, PGD further enhances the effectiveness of adversarial attacks by performing multiple iterations of gradient ascent while simultaneously constraining the perturbations to remain within a specified range or boundary. By iteratively updating the input data in the direction that maximises the loss function while projecting it back onto the allowed perturbation space, PGD generates strong adversarial samples which are more likely to deceive the target model with improved success rate.

5.4 Carlini and Wagner (CW) Method

The CW attack, proposed by Carlini and Wagner in 2016, attack is a powerful and widely recognized optimization-based method for crafting adversarial examples. This attack aims to find minimal perturbations that can fool a target model while adhering to certain constraints. The CW attack, in contrast to earlier attacks, formulates the creation of adversarial examples as an optimisation problem, enabling fine-grained control over the perturbations. This attack is known for its versatility and success across various types of models and defences, making it a valuable tool for assessing model vulnerability and testing the robustness of machine learning systems.

5.5 Imperceptible Carlini and Wagner (CW) Method

The Imperceptible CW attack, is an advancement of a targeted CW attack, but focuses on generating adversarial examples that are imperceptible to the human eye. Proposed as a defence-aware attack by Athalye et al. in 2018, this method introduces an additional constraint during the optimization process with the aim to minimise the perceptibility of the perturbations by considering human perception models and leveraging knowledge about the human visual system. By incorporating perceptual constraints, like spatial smoothness and colour similarity, it can generate adversarial examples that maintain high visual similarity to the original input while still successfully fooling the target model.

6. Proposed Methodology

This study delves into the implementation of adversarial attacks on an ASR model, employing five distinct adversarial attack methods. The methodology commences with the cloning two GitHub repositories, one dedicated to the adversarial attacks and the other to a convolutional deep neural

network model. Subsequently, the environment is configured to the ASR adversarial attacks directory, and requisite dependencies are installed. A pre-trained Wav2Vec2 ASR model is utilized from the torchaudio model hub.

Upon loading an audio file for attack generation, target and true transcriptions are specified. The ensuing stage encompasses both targeted (perturbing audio with a specific transcription) and untargeted instances (perturbing audio without a specific transcription), for all five attack methods: FGSM, BIM, PGD, CW and Imperceptible. This comprehensive analysis considers scenarios with and without early stopping to provide a thorough understanding of the impact of each attack method.

Finally, a systematic assessment is conductedusing the Signal-to-Noise Ratio (SNR) metric to qualify the influence of various targeted and untargeted adversarial attacks on clean audio. This comprehensive evaluation provides valuable insights into the impact of each attack method on the audio's Signal-to-Noise ratio value.

7. Metric used for Comparison of Attacks

Noise in speech recognition systems refers to any unwanted or interfering sounds present in the audio signal that can degrade the accuracy and performance of the system. The presence of noise can introduce errors and hinder the system's ability to accurately transcribe spoken words. Noise can make it challenging for the system to distinguish between speech and non-speech components, leading to reduced accuracy and increased error rates.

In the presence of background noise, the quality and clarity of a signal are measured using the Signal-to-Noise Ratio (SNR) metric. It measures the ratio of the power or amplitude of a desired signal to power or amplitude of the noise interfering with the signal. A larger value of SNR indicates a stronger signal in comparison to the noise, improving the quality of the signal and its comprehensibility. It is commonly used in various fields, including telecommunications, audio processing, and speech recognition, to assess and optimise the performance of systems in noisy environments. It acts as a useful metric for assessing the efficacy of noise reduction methods and improving overall signal fidelity under challenging acoustic circumstances. It has been used for comparing the various attack methods to check which attack adds less noise. Lower SNR is better because of the way it is measured.

8. Results

Figure 1 depicts the audio clips for targeted and untargeted attack methods, including Fast Gradient Sign Method (FGSM), Basic Iterative Method (BIM), Projected Gradient Descent (PGD), Carlini-Wagner (CW) and Imperceptible CW, in two distinct settings: one with early stopping and another without early stopping, offering a comparative view of their impact.

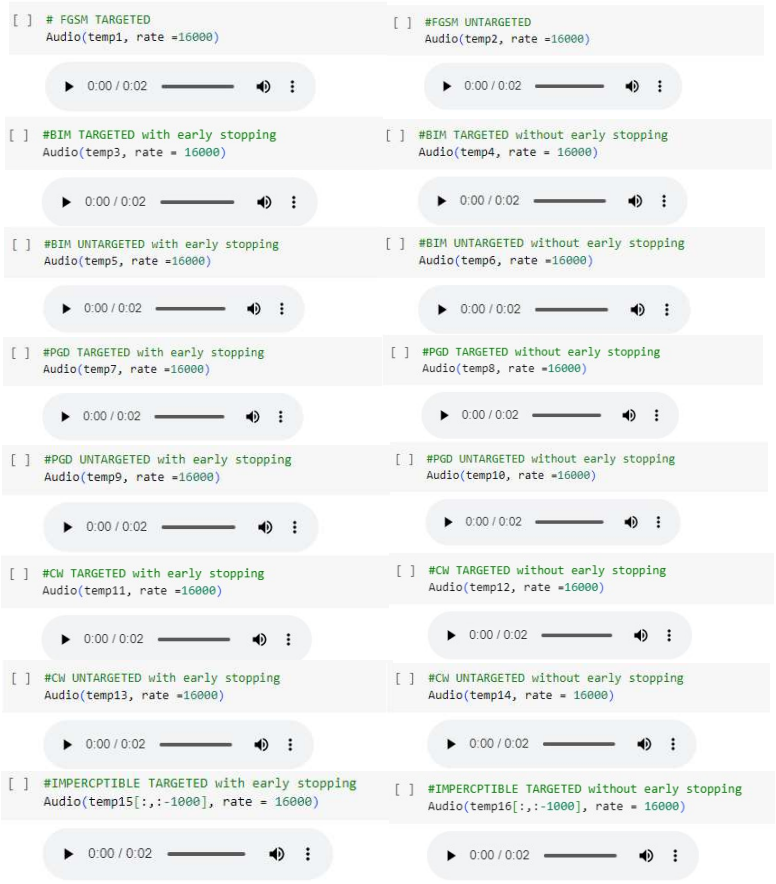

Figure 1. Audios for targeted and untargeted attacks

Table 1 and 2 present the Signal-to-Noise Ratio (SNR) values for targeted attacks and untargeted attacks respectively, detailing both scenarios with and without early stopping. Notably, SNR value for FGSM remains consistent regardless of the early stopping application. In both targeted and untargeted instances, FGSM emerges as the least detrimental attack method, boasting the highest SNR value. In the targeted scenario, BIM closely follows as the second least detrimental, without early stopping, while in the untargeted scenario, PGD claims the second position. Contrarily, in both targeted and untargeted scenarios, CW emerges as the most detrimental attack method, exhibiting lowest SNR value. Imperceptible attack falls exclusively into the targeted attacks category, demonstrating an intermediate intensity level positionedbetween FGSM and BIM in the case of with early stopping and between BIM and PGD in the case of PGDwithout early stopping, as evidenced by its SNR value.

Table 1. Results for targeted attacks.

SNR	Targeted attacks	
	With early stopping	Without early stopping
FGSM	-14.40436840057373	
BIM	-38.70874881744385	-22.978615760803223
PGD	-52.756638526916504	-52.32065677642822
CW	-57.65882968902588	-56.628432273864746
IMPERCEPTIBLE	-33.50316524505615	-33.190226554870605

Table 2. Results for untargeted attacks.

SNR	Untargeted attacks	
	With early stopping	Without early stopping
FGSM	-21.887822151184082	
BIM	-78.14002513885498	-78.14002513885498
PGD	-52.284531593322754	-52.284531593322754
CW	-94.0333890914917	-94.0333890914917

9. Conclusion & Future Work

For the purpose of this study, five different attacks methods were applied on the wav2vec2 model with the aim ofexamining differences in their intensity, using the SNR metric. Notably, the results underscore that the Fast Gradient Sign Method (FGSM) attack is the least intensive of all in both targeted and untargeted categories, having the highest SNR value. Conversely, the Carlini and Wagner (CW) attack is the most intensive of all in both targeted and untargeted attacks, having the lowest SNR value. This suggests a priority for ASR models to prioritize defence mechanisms against the Carlini and Wagner attack method. Subsequently, the order of concern should be tailored, addressing Projected Gradient Descent (PGD) method, Basic Iterative Method (BIM) or Imperceptiblemethod based on the early stopping criterion, and ultimately Fast Gradient Sign Method (FGSM).

The objective of this paper is to provide valuable assistance for future advancements in the domain of adversarial attacks on Automatic Speech RecognitionSystems(ASRSs), specifically in distinguishing and categorizing attack methods based on their levels of intensity. This research aims to serve as a valuable point of reference for upcoming investigations into the development of defensive strategies against adversarial attacks on ASR systems and to contribute significantly to the ongoing studies in the field of adversarial machine learning and speech recognition.

References

[1] Zelasko, P., Joshi, S., Shao, Y., Villalba, J., Trmal, J., Dehak, N., Khudanpur, S.: Adversarial Attacks and Defenses for Speech Recognition Systems. ArXiv abs/2103.17122 (2021)

[2] Markert, K., Mirdita, D., Bottinger, K.: Language Dependencies in Adversarial Attacks on Speech Recognition Systems. ArXiv abs/2202.00399 (2022)

[3] Carlini, N., Wagner, D.: Audio Adversarial Examples: Targeted Attacks on Speech-to-Text. IEEE Security and Privacy Workshops, 1-7 (2018)

[4] Lan, J., Zhang, R., Yan, Z., Wang, J., Chen, Y., Hou, R.: Adversarial Attacks and Defenses in Speaker Recognition Systems: A Survey. Journal of Systems Architecture 127, 102526 (2022)

[5] Huynh, N., Bouadjenek, R., Razzak, I., Lee, K., Arora, C., Hassani, A., Zaslavsky, A.: Adversarial Attacks on Speech Recognition Systems for Mission-Critical Applications: A Survey. ArXiv abs/2202.10594 (2022)

[6] Zhang, J., Zhang, B., Zhang, B.: Defending Adversarial Attacks on Cloud-Aided Automatic Speech Recognition Systems. In: Proceedings of the Seventh International Workshop on Security in Cloud Computing, pp. 23-31. Association for Computing Machinery, New York, NY, USA (2019)
[7] Zong, W., Chow, Y., Susilo, W., Kim, J., Le, N.: Detecting Audio Adversarial Examples in Automatic Speech Recognition Systems Using Decision Boundary Patterns. Journal of Imaging 8(12), 324 (2022)
[8] Samizade, S., Tan, Z., Shen, C., Guan, X.: Adversarial Example Detection by Classification for Deep Speech Recognition. ICASSP 2020- 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 3102-3106. IEEE, Barcelona, Spain (2020)
[9] Joshi, S., Kataria, S., Shao, Y., Zelasko, P., Villalba, J., Khudanpur, S., Dehak, N.: Defense Against Adversarial Attacks on Hybrid Speech Recognition Using Joint Adversarial Fine-Tuning with Denoiser. Interspeech 2022, 5035-5039 (2022)
[10] Schönherr, L., K. Kohls, K., Zeiler, S., Holz, T., Kolossa, D.: Adversarial Attacks Against Automatic Speech Recognition Systems via Psychoacoustic Hiding. ArXiv abs/1808.05665 (2018)
[11] Xie, Y., Shi, C., Li, Z., Liu, J., Chen, Y., Yuan, B.: Real-Time, Universal, And Robust Adversarial Attacks Against Speaker Recognition Systems. In: ICASSP 2020- 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1738-1742. IEEE, Barcelona, Spain (2020)
[12] Li, Z., Shi, C., Xie, Y., Liu, J., Yuan, B., Chen, Y.: Practical Adversarial Attacks Against Speaker Recognition Systems. In: Proceedings of the 21st International Workshop on Mobile Computing Systems and Applications (HotMobile'20), pp. 9-14. IEEE, Austin TX, USA (2020)
[13] Joshi, S., Villalba, J., Zelasko, P., Velazquez, 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 16, 4811-4826 (2021)
[14] Xie, Y., Li, Z., Shi, C., Liu, J., Chen, Y., Yuan, B.: Real-Time, Robust and Adaptive Universal Adversarial Attacks Against Speaker Recognition Systems. Journal of Signal Processing Systems 93, 1187-1200 (2021)
[15] Chen, G., Chenb, S., Fan, L., Du, X., Zhao, Z., Song, F., Liu, Y.: Who Is Real Bob? Adversarial Attacks on Speaker Recognition Systems. In: Proceedings of IEEE Symposium on Security and Privacy, pp. 694-711. IEEE, Piscataway, NJ, USA (2021)
[16] Chen, G., Zhao, Z., Song, F., Chen, S., Fan, L., Y. Liu, Y.: Arbitrary Source-To-Target Adversarial Attack on Speaker Recognition Systems. In: IEEE Transactions on Dependable and Secure Computing, pp. 1-17. IEEE (2022)
[17] Villalba, J., Joshi, S., Zelasko, P., Dehak, N.: Representation Learning to Classify and Detect Adversarial Attacks Against Speaker and Speech Recognition Systems. Interspeech 2021, 4304-4308 (2021)
[18] Jati, A., Hsu, C., Pal, M., Peri, R., Abdalmageed, W., Narayan, S.: Adversarial Attack and Defense Strategies for Deep Speaker Recognition Systems. Computer Speech and Language 68(5), 101199 (2021)
[19] Wang, Q., Zheng, B., Li, Q., Shen, C., Ba, Z.: Towards Query-Efficient Adversarial Attacks Against Automatic Speech Recognition Systems. IEEE Transactions on Information Forensics and Security 16, 896-908 (2021)
[20] Tan, H., Wang, L., Zhang, H., Zhang, J., Shafiq, M., Gu, Z.: Adversarial Attack and Defense Strategies of Speaker Recognition Systems: A Survey. Electronics 11(14), 2183 (2022)