Vulnerability of Neural Network based Speaker Recognition systems

Thesis submitted in partial fulfillment
of the requirements for the degree of

Master Of Science
in
Computer Science and Engineering by Research

by

Ritu Srivastava
2018701002
ritu.srivastava@research.iiit.ac.in

International Institute of Information Technology Hyderabad
500 032, India

June 2024

International Institute of Information Technology Hyderabad
Hyderabad, India

CERTIFICATE

This is to certify that work presented in this thesis, titled ***Vulnerability of Neural Network based Speaker Recognition systems*** by *Ritu Srivastava* has been carried out under my supervision and is not submitted elsewhere for a degree.

_____ _____

Date Advisor: Dr. Vineet Gandhi

To my family and friends

Acknowledgments

I would first like to express gratitude to my advisor, Prof. Vineet Gandhi, for the wonderful experience and support throughout my research program. I am grateful for the calm and friendly manner in which he engaged with me during our discussions. He encouraged the initial explorations, provided freedom and flexibility to pursue new areas, and helped formulate ideas while helping me become an independent thinker. I thank him not only for being a research guide but also for his role in shaping my career.

I would also like to thank Sarath Sivaprasad, Saiteja Kosgi, Neha SahipJohn for their valuable discussions in understanding and approaching any idea. I also thank my colleagues at CVIT for the vibrant work environment they created, which motivated me to pursue excellence in my career.

I want to extend my gratitude to all professors and IIIT academics who introduced me to various evolving fields in technology. I also thank the overall CVIT ecosystem for the unparalleled access to resources. I thank the staff and administration for promptly resolving any issues and thank all my friends for supporting me and making an enriching academic journey.

I am deeply grateful to my family for their unwavering support and for standing by me throughout my academic life. I can never be thankful enough to them for enabling me to pursue my career and, specifically, for making this thesis happen. I dedicate this thesis to my beloved family.

Above all, I owe it all to Almighty God for granting me wisdom, health, and strength to undertake research and enabling me to complete it.

I must express gratitude to the pioneering researchers whose immense contributions laid the groundwork upon which I built. I also acknowledge science enthusiasts and educators who played a vital role in shaping my outlook and cultivating an environment conducive to broad scientific exploration. I thank my advisor and fellow researchers for my learning throughout the process.

Abstract

Speaker recognition (SR) involves automatic identification of individual speakers based on their voices, often representing acoustic traits as fixed-dimensional vectors through speaker embedding. A standard speaker recognition system (SRS) consists of three key phases: training, enrollment, and recognition. In each stage, acoustic features are extracted from raw speech signals using an acoustic feature extraction module, resulting in the acquisition of essential acoustic characteristics. Commonly used acoustic features include speech spectrogram, filter bank, and Mel-frequency cepstral coefficients.

During the training stage, a background model is trained to establish a mapping from training voices to embeddings. The traditional background model employs a Gaussian Mixture Model (GMM) to generate identity-vector (ivector) embeddings. In contrast, more recent and promising background models leverage deep neural networks (DNNs) to generate deep embeddings, like xvector. In the enrollment stage, a voice spoken by an individual undergoing enrollment is mapped to an enrollment embedding using the previously trained background model. In the recognition stage, the process begins by retrieving the testing embedding of a given voice from the background model. Subsequently, the scoring module is engaged to measure the similarity between the enrollment and testing embeddings. The scoring module evaluates the similarity between the speaker and recorded embedding. Following the assessment, the scoring and decision module makes a decision based on the similarity score. A decision threshold is established, which serves as a criterion to determine whether the claimed identity of the speaker is accepted or rejected.

The concept of voiceprint is rapidly gaining prominence as one of the emerging biometrics, primarily owing to its seamless integration with natural and human-centered Voice User Interface (VUI). The fast progress of Speaker Recognition Systems (SRSs) is intricately linked to the evolution of Neural Networks (NNs), with a particular emphasis on Deep Neural Networks (DNNs). With strides made in deep learning, Speaker Recognition (SR) has also benefitted and found extensive applications across hardware and software platforms.

However, it has been shown that NNs are vulnerable to adversarial attacks, highlighting a challenge that needs to be addressed. Thus, even though users have the convenience of authentication with Speaker Recognition services, it has become evident that these solutions are vulnerable to adversarial attacks. This vulnerability highlights that Speaker Recognition (SR) is encountering security threats, raising significant concerns about user privacy.

Adversarial attack was initially implemented with images, where an image classification model was successfully deceived using adversarial examples. Drawing inspiration from the progress made in ad-

versarial attacks within the image domain, there is a growing interest in extending these techniques to the audio field. With emerging trends, convolutional neural networks have demonstrated instability to artificially crafted perturbations that remain undetectable to the human eye. Virtually every type of model, ranging from CNN to graphical neural network (GNN), has shown vulnerability to adversarial examples, particularly in the domain of image classification.

Deep learning models typically get audio input by converting the audio into a spectrogram for further processing. A spectrogram serves as a condensed representation of an audio input. Given its image-like nature, the audio spectrogram is frequently used as input data for deep learning models, especially Convolutional Neural Networks (CNNs) adapted for audio tasks. CNN-based architectures were initially designed for image processing.

This thesis contributes to the assessment of Convolutional Neural Networks (CNNs) for their resilience against adversarial attacks, a domain that is yet to be extensively investigated concerning end-to-end trained CNNs for speaker recognition. This examination is essential for sustaining the integrity and security of speaker recognition systems. Our study fills this gap by exploring the variations of iterative Fast Gradient Sign Method (FGSM) to carry out adversarial attacks. We note that using a vanilla iterative FGSM technique can alter the identity of each speaker sample to any other speaker within the LibriSpeech dataset.

Additionally, we introduce adversarial attacks specific to Mel spectrogram features by (a) constraining the number of manipulated pixels, (b) confining alterations to certain frequency bands, (c) limiting changes to particular time segments, and (d) employing a substitute model to generate the adversarial sample. Through comprehensive qualitative and quantitative analyses, we illustrate the vulnerability and counterintuitive behavior of existing CNN-based speaker recognition systems, wherein the predicted speaker identities can be inverted without discernible alterations in the audio. The samples are available at “https://advdemo.github.io/speech/”

Contents

List of Figures

List of Tables

Symbols

$\odot$	Element-wise product
J	Jacobian cost function
∇	Gradient
$A^\intercal$	Transpose of matrix A

List of Related Publications

- **Adversarial Robustness of Mel Based Speaker Recognition Systems** , Ritu Srivastava, Saiteja Kosgi, Sarath Sivaprasad, Neha Sahipjohn and Vineet Gandhi , **In APSIPA** : Asia Pacific Signal and Information Processing Association, 2023, Taipei, Taiwan

Chapter 1

Introduction

Motivation and Background

While deep learning excels in accomplishing various Computer Vision tasks with impressive accuracies, [1] initially identified a notable vulnerability in deep neural networks for image classification. They demonstrated that, despite achieving high accuracies, contemporary deep networks exhibit unexpected susceptibility to adversarial attacks characterized by slight perturbations to images that remain nearly imperceptible to the human visual system. These attacks have the potential to completely change the predictions made by a neural network classifier for input images. What's even more concerning is that the compromised models show high confidence in incorrect predictions. Additionally, the same perturbed image can fool multiple network classifiers.

These perturbed adversarial samples indicate that traditional neural network architectures cannot comprehend concepts or high-level abstractions. Additionally, adversarial examples possess the potential to be dangerous. This fact has led to growing concern over the increasing number of artificial intelligence integrations into various systems crucial for our safety and convenience. These include banking, surveillance systems, ATMs, laptop face recognition, and self-driving cars. Recently, safety considerations regarding AI have been centered around ethical implications.

The use of deceptive data to trick models is an escalating threat in the AI and machine learning research community. Attacks on AI models can either unintentionally influence the classifier or breach security. In the former, the attack disrupts the model during the prediction process, while in the latter, malicious data is provided, leading to its classification as legitimate.

Already operational systems might be susceptible to attacks. For instance, researchers demonstrated the vulnerability of a self-driving car in production by strategically placing small stickers on the ground, causing the vehicle to veer into the opposite lane. Similarly, subtle alterations to an image sample have been shown to deceive medical analysis systems, leading them to misclassify a benign mole as malignant. The growing use of AI is expected to be accompanied by increased adversarial attacks. Even the most proficient forensic classifiers, such as AI systems designed to differentiate between authentic and synthetic content, are vulnerable to adversarial attacks. Even though not entirely new, this poses a concerning challenge for organizations aiming to commercialize fake media detectors, especially given

the rapid proliferation of deepfake content on the internet. The malicious use of fake media has the potential to influence opinions during an election, falsely implicate an individual in a crime, or defraud a major energy producer. Adversarial examples provide a starting point for addressing AI safety concerns. When delving into the study of AI safety, our focus often turns to some of the most challenging issues in the field. These include ensuring that highly intelligent reinforcement learning agents, surpassing human capabilities, exhibit behavior as intended by their designers.

The security and safety risks posed by adversarial samples also present prohibitions in applying neural networks to numerous critical scenarios. Hence, it becomes crucial to design neural networks that are not only accurate but also robust. Consequently, a comprehensive assessment of robustness is essential to evaluate the network's resilience efficiently. The space of adversaries and metrics makes the current state-of-the-art challenging to understand. Additionally, most recent adversarial attacks are white-box, limiting their applicability for assessing hybrids, non-standard neural networks, and other classifiers in general.

It turns out that evaluating robustness proves to be a formidable challenge due to the multitude of possibilities, definitions, exceptions, and trade-offs involved. Moreover, these adversarial samples highlight deficiencies in the reasoning of existing machine learning algorithms. Enhancements in robustness should also lead to learning systems capable of more adept reasoning over data and achieve a heightened level of abstraction. Therefore, robustness assessment is instrumental in identifying and rectifying failures in both reasoning and high-level abstractions.

Why is Vulnerability Assessment Necessary? The frequency and severity of cybersecurity breaches are on the rise. The primary targets for these attacks are business sectors such as financial services, SaaS/webmail, and social media, accounting for more than 50% of all reported incidents.

In the contemporary landscape, organizations function within a highly interconnected environment. They engage with software applications, execute programs, and handle data in the cloud. Additionally, they utilize networked devices and collaborate with numerous vendors and third parties, significantly enlarging the potential attack surface for cybercriminals. These malicious actors probe systems for misconfigurations, outdated or unpatched software, weak credentials, and other security gaps that render systems susceptible to attacks. Hence, it is logical to proactively identify and rectify these vulnerabilities before they can be exploited by malicious entities.

Addressing every identified vulnerability through patching is often impractical. It is crucial to prioritize vulnerabilities that most impact your organization, which could be in critical systems or those that could result in significant financial harm. Swiftly patching or implementing fixes for these prioritized vulnerabilities is essential.

According to a 2021 survey by Gartner, security concerns emerged as a primary obstacle to adopting AI, sharing the top position with the complexity of integrating AI solutions into existing infrastructure. Thus, the challenges associated with securing AI and machine learning systems are considerable.

1.1 Speech Processing

Various mission-critical applications include rescue missions, military operations, search and recovery, and emergency management actions. A Mission-Critical Application refers to a system in which a hazard has the potential to degrade or impede the successful completion of a planned operation [2]. Hence, any interruption in a mission-critical application has direct negative financial or life-threatening ramifications. Recent strides in Artificial Intelligence (AI) have refreshed the notion of direct communication with computer systems. The evolution of Conversational User Interfaces (CUIs) for mission-critical applications represents a significant advancement, featuring two-way interaction between humans and machines. This facilitates end-users in resolving issues through voice support. Development of these conversational systems [3] uses deep learning techniques in mission-critical applications such as recognizing medical speech transcriptions in healthcare. Recently, there has been a shift towards crafting end-to-end neural networks and harnessing extensive datasets to enhance the precision of speech conversational systems. In these end-to-end models, the components of the conventional system, constituting the acoustic, pronunciation, and language models, are collectively optimized as a unified system.

Nevertheless, while conversational systems relying on speech recognition have yielded numerous advantages for mission-critical applications, the most significant threat, particularly for deep learning-based systems, is posed by adversarial attacks. Adversarial attacks on speaker recognition systems have increased in importance due to ethical considerations, security concerns, and practical implementations in real-world scenarios. [4] showcase effective attack algorithms that specifically target conventional speech recognition models relying on Hidden Markov Models (HMMs) and Gaussian Mixture Models (GMMs) that work on Mel Frequency Cepstral Coefficient (MFCC) representation of audio. In the case of Hidden Voice Commands [4], for instance, the attacker employs inverse feature extraction to produce obscured audio capable of being played over-the-air to launch attacks on Automatic Speech Recognition (ASR) systems.

Individuals use language to express their emotions and sentiments effectively. Through speech, people can communicate and convey meaning in their preferred language. Speech processing is a dedicated field that involves examining and applying methods for analyzing and manipulating speech signals. This field covers various tasks, including speech synthesis or text-to-speech, automatic speech recognition (ASR), and speaker recognition (SR) [5]

Signal processing is a discipline that comprises of the study of quantities showing variations in space or time commonly denoted as signal. Sound signals are particularly defined as fluctuations in air pressure. Hence, a speech signal is recognized as a kind of sound signal, namely pressure variations, created by humans to assist in spoken communication. A signal that consistently repeats at a fixed interval, referred to as a period, is categorized as periodic in the domain of signal processing. The frequency of the signal is represented by the reciprocal of this period. Speech signals are frequently converted into digital form for the processing of speech. The sampling rate, determined by the number of samples gathered per second, dictates the level of detail in this digitization process [5]. Speech features refer to numerical representations of speech signals employed for analysis, recognition, and

synthesis. In general, speech signals can be categorized into two main types: time-domain features and frequency-domain features.

Time-domain features are directly derived from the amplitude of the speech signal across time. These features are simple to calculate and are frequently utilized in real-time speech-processing applications. Common time-domain features include:

- Energy is a numerical gauge of the amplitude characteristics of a speech signal across time. The calculation involves squaring each sample in the signal and summing them within a defined time window. This process captures the overall strength and dynamics of the signal, providing insight into temporal variations in intensity.

- Zero-crossing rate represents the number of times the speech signal crosses the horizontal axis (zero amplitude) within a given time frame.

- Pitch refers to the fundamental frequency of a periodic signal, which is the frequency of the repetition of the basic waveform of the signal. Pitch in audio signals can be estimated using autocorrelation and cepstral analysis.

- Linear Predictive Coding (LPC) is a robust method that utilizes an autoregressive model to characterize the speech signal as a linear combination of previous samples.

Frequency domain features are extracted from the signal's representation in the frequency domain, commonly known as its spectrum. A spectrum captures the energy distribution of signal across different frequencies. Spectrograms are visual representations in two dimensions that capture variations in a signal's spectrum over time [5]. In a typical image, neighboring pixels of an object usually belong to the same object, a characteristic not shared by spectrograms where overtones from the same sound source can lie far apart. Also, various sound sources or objects may exhibit overlapping frequency content, resulting in a scenario where a single pixel in the spectrogram can be associated with multiple sounds [6]. Common frequency-domain features include:

- Mel-spectrogram represents the short-term power spectrum of a sound signal. It is computed by converting the power spectrum of a speech signal into a mel-scale, a nonlinear scale that approximates the human perception of sound. The mel-scale partitions the frequency range into a series of mel-frequency bands, offering finer resolution in the lower frequencies and coarser resolution in the higher frequencies. The Mel-spectrogram records not only the spectral content but also the temporal dynamics of the signal. This is advantageous in situations where capturing both the spectral patterns and frequency content of the signal holds significance for the analysis or classification task.

- MFCC coefficients encapsulate the power spectrum of a sound within a brief interval by employing a linear cosine transformation of a logarithmically-scaled power spectrum on a non-linear

mel frequency scale. Comprising a set of coefficients that collectively constitute a Mel-frequency cepstrum, MFCCs offer an efficient means of analyzing audio with just 12 parameters related to frequency amplitude. This compact representation strikes a balance, providing an adequate number of frequency channels for audio analysis.

Frequency-based representations like the Mel spectrogram and MFCC are extensively employed in speech processing since they are comparatively more robust to noise than temporal variations of the sound. [7].

Extensive research has been conducted in speaker recognition, focusing on addressing two key objectives: speaker identification and speaker verification.

- Speaker identification(SI) seeks to differentiate the identity of an unknown individual from a set of known speakers. This entails detailed examination of the speaker's voice characteristics such as pitch, tone, accent, and other relevant features such as vocal cord size, to establish their identity.
- Speaker Verification (SV) is a procedure that entails validating the identity of a speaker based on their speech. It distinguishes itself from speaker identification as it specifically confirms whether a speaker is indeed who they claim to be by comparing their voice with an existing speaker template. Deep learning-based speaker verification relies on low-dimensional vector representations from speech signals, capturing distinctive speaker characteristics such as pitch and speaking style.

Early Speaker Recognition (SR) research primarily focused on understanding human capabilities. While recognizing voices over the phone is common for humans, implementing autonomous SR tasks presents considerable challenges. Over the years, autonomous SR systems have experienced both successes and setbacks. Significant progress over made the last five decades has contributed to overcoming many significant challenges associated with SR systems. Expanding upon Noll's pitch detection, research in [8] employed cepstral analysis for modeling speech. The Cepstral speech model has become an important tool for Speaker Recognition systems.

Over the years, various types of feature modeling methods have been used, including the Hidden Markov Model, Vector Quantization, and template matching models. The Gaussian Mixture Model (GMM) stands out as a noteworthy advancement in recognition systems. Arguably, one of the most notable advancements in speech-related research was the introduction of the Universal Background Model (UBM) [9]. This approach not only involves modeling an individual's voice and testing the likelihood of that person being the authenticated user but also proposes using a set of individuals who are not authenticated user. Support Vector Machines (SVM) are employed for data classification. In the verification task, SVM is utilized to classify data into categories of authenticated users or impostors. The SVM classifier minimizes false reject and false accept error rates by using an optimized non-linear decision boundary, as opposed to a simple threshold.

SRS implementations include the ivector-PLDA [10] method, widely adopted in academic research and industry. This method attains state-of-the-art performance across various speaker recognition tasks

[11] [12]. Another approach involves GMM-UBM based methods, where a Gaussian mixture model (GMM) [9] is trained as the Universal Background Model (UBM). In recent times, deep neural networks (DNN) have gained prominence in both speech processing and speaker recognition, as exemplified by approaches like xvector-PLDA [13]. Methods based on Deep Neural Networks (DNN) typically depend on much larger amounts of training data, leading to a significant increase in computational complexity when compared to ivector and GMM-based methods. As a result, DNN-based methods may not be well-suited for offline enrollment on client-side devices [14].

The domain of speech processing has experienced a transformative shift with the emergence of deep learning.

1.2 Convolutional Neural Network

CNNs are modified versions of fully connected neural networks extensively utilized for processing data with a grid-like topology. Examples of such data include time-series data (1D grid) with samples at regular intervals or images (2D grid) with pixels, both exhibiting a grid-like structure. 1D CNN is suitable for tasks like time series prediction and signal identification. It is used on time-series data, such as audio signals, ECG signals, or sensor signals. In contrast, a 2D CNN finds applications in tasks such as image classification, object detection, image segmentation, and face recognition. Moving to 3D CNNs, they are employed in tasks like human action recognition and object recognition/detection.

The speech spectrogram preserves more information compared to hand-crafted features, encompassing speaker characteristics like vocal tract length variations and distinct speaking styles. Spectrogram exhibits high correlations in both time and frequency. Given these characteristics, the spectrogram serves as a fitting input for a CNN processing pipeline that necessitates maintaining locality along both the frequency and time axes. Modeling local correlations with CNNs is advantageous for speech signals. Additionally, CNNs can adeptly extract structural features from the spectrogram, thereby reducing model complexity through weight sharing. Given that spectrograms are two-dimensional visual representations, one can harness CNN architectures used extensively for visual data processing (images and videos) by carrying out convolutions in two dimensions.

CNNs can replace previously popular methods like HMMs and GMM-UBM in various applications. Additionally, CNNs demonstrate the capability to capture features that maintain robustness even in the face of variations in speech signals caused by diverse speakers, accents, and background noise [5]. CNNs have demonstrated their versatility as effective tools for various speech-processing tasks. They have been successfully applied to speech recognition [15], including hybrid NN-HMM models and multi-class classification. Additionally, CNNs have been suggested for speaker recognition in emotional speech, with a constrained CNN model outlined in [16]. Both 1D and 2D CNNs have emerged as the core building blocks for numerous speech processing models, including acoustic models [17] in Automatic Speech Recognition (ASR) systems. Similarly, VGGVox [18] achieved state-of-the-art results in speaker recognition by using a CNN with VGG architecture to learn speaker embeddings from Mel

spectrograms. Furthermore, CNNs have found extensive use in developing state-of-the-art architectures for speech enhancement and text-to-speech applications.

1.3 Adversarial Attacks on Deep Neural Network

An adversarial attack seeks to trick a deep learning model into producing a specific output with minimal changes to the input data or by extracting valuable information from the model through diverse tactics. The attacker's level of access to the target deep learning models, including their weights and training dataset, may vary.

Why do Adversarial Examples Exist? The initial and original hypothesis attempting to explain adversarial examples originated from [1]. In this work, the argument was presented that these examples arise from the existence of low-probability "pockets" in the manifold, indicating excessive non-linearity and poor regularization of the network. Subsequently, an alternative theory emerged, spearheaded by [19], contending that adversarial examples actually result from excessive linearity in modern machine learning, particularly in deep learning systems. The third and widely embraced hypothesis explaining adversarial examples is the tilted boundary [20]. In essence, the authors present that since the model never achieves a perfect fit to the data, adversarial pockets of inputs will always exist between the classifier's boundary and the actual sub-manifold of sampled data.

Traditionally, attacks on machine learning have been generally classified based on two criteria: the extent of information available to the attacker and the timing of the attack. Regarding information, the attacker might possess complete knowledge of a model, including parameters, features, and training data. Conversely, the attacker may lack any insight into the internal mechanisms of the model and only have access to its predictions. Concerning timing, an attacker can opt to target the learning algorithm during the model training phase or target a pre-trained model at the point when it makes a decision.

Numerous decision-time attacks hinge on inference for generating adversarial examples. A malicious actor can submit input data accepted by a trained model and observe the results, namely the model's classification. This iterative process allows the attacker to continually adjust the input features, gaining new insights and inferring the decisions made by the model.

Generally, the more extensively an adversary understands your model, features, and training data, the simpler it becomes to create subtly altered adversarial examples that cross decision boundaries. This can lead to misclassification by the model, exposing decision boundaries to the attacker and facilitating the development of subsequent attacks. Intellectual property theft is another distinct yet plausible motivation for an attack, specifically stealing the model itself.

The origins of adversarial attacks can be attributed to various source papers. However, focusing specifically on the deep learning domain, [1] was the first to showcase the vulnerability of Convolutional Neural Networks (CNN) to adversarial examples. The initial advancement in the realm of adversarial attacks related to computer vision and image-related tasks is the white-box attack. Among the

foundational attacks introduced, the Fast Gradient Sign Method (FGSM) [19] stands out as one of the earliest.

Essentially, the machine learning approach employed in modern AI systems is vulnerable to attacks targeting the public APIs offered by the model and the platforms where they operate. Attackers can compromise the confidentiality and privacy safeguards of both data and models by exploiting the model's public interfaces and providing data inputs within the acceptable range. The following section describes the classification of attacks from various perspectives.

Attack Classification

1.3.1 Stages of learning to perform attacks

In the Machine Learning literature, learning paradigms include supervised learning, where models are trained using labeled data during training and optimization aims to minimize a specific loss function; unsupervised learning, which involves training models with unlabeled data; semi-supervised learning, wherein a small subset of examples is labeled, while the majority remains unlabeled; reinforcement learning involves an agent interacting with an environment, learning an optimal policy to maximize its reward; federated learning is a collaborative approach where a group of clients jointly trains machine learning model by communicating with a server, which aggregates model updates. Ensemble Learning, on the other hand, is a machine learning strategy that aims to enhance predictive performance by combining predictions from multiple models [21].

The literature on adversarial machine learning primarily examines attacks against AI systems, which may occur either during the training stage or the deployment stage of machine learning (ML). In the ML training stage, the attacker could exert control over aspects such as training data, labels, model parameters, or the code of machine learning algorithms. In contrast, during the machine learning deployment or decision time, the ML model is already trained, and adversaries may launch attacks to compromise integrity, altering the model's predictions. Alternatively, they might target the trained model to infer sensitive information about the training data or the machine learning model.

1.3.1.1 Training-time attacks

Attacks during the ML training stage can manifest as poisoning Attacks or byzantine attacks.

In a data poisoning attack [22], an adversary gains control over a subset of the training data, , achieved by either inserting or modifying training samples. Conversely, in a model poisoning attack [23], the adversary controls the model and its parameters. Consequently, these attacks may involve data injection, data modification, and logic corruption.

In data injection, the adversary lacks access to the training data or learning algorithm but can corrupt the target model by inserting adversarial samples into the training dataset. On the other hand, in data modification, the adversary lacks access to the learning algorithm but possesses full access to the training

data, allowing them to poison the training data directly by modifying it before it is used for training the target model. In cases of logic corruption, the adversary possesses the capability to tamper with the learning algorithm.

Attacks can also occur on federated learning where malicious actors operate one or more participating edge devices. Such attacks are called byzantine attacks.

1.3.1.2 Testing-time attacks

Potential attacks at testing, deployment, or decision time include evasion, oracle, and privacy attacks.

Evasion attacks involve altering testing samples to create adversarial examples [24]. These adversarial examples closely resemble the original samples (according to specific distance metrics) but alter the model predictions to align with the attacker's choices.

In an Oracle Attack, the adversary tries to gain insights into the model's internals or the training data by providing a series of carefully crafted inputs. This type of attack encompasses three categories – Extraction Attacks, Inversion Attacks, and Membership Inference Attacks.

In Extraction Attacks, the adversary extracts the structure or parameters of the model by observing the model's predictions, typically including probabilities assigned to each class. In Inversion Attacks, the inferred characteristics might enable the adversary to reconstruct the data employed for training the model, potentially revealing personal information that infringes upon an individual's privacy. In Membership Inference Attacks, the adversary employs returns from queries to the target model to discern whether certain data points belong to the same distribution as the training dataset. This is accomplished by making use of differences in the model's confidence for the samples that were or were not encountered during the training process.

Privacy attacks are typically initiated by attackers with query access to a machine learning model. These attacks can be further categorized into data and model privacy attacks.

1.3.2 Attacker Goals and Objectives

The attacker's objectives are categorized along three dimensions, aligning with the primary types of security violations scrutinized when assessing the security of a system, namely availability breakdown, integrity violations, and privacy compromise.

In an Availability Attack, the attacker seeks to disrupt the model's performance during testing or deployment. This attack aims to disrupt the availability of systems, applications, or data to users. This is often achieved by overwhelming the target with a high volume of traffic, rendering it unable to respond to legitimate requests.

An Integrity Attack aims to compromise the integrity of an ML model's output, leading to incorrect predictions. This alters the model's accuracy and validity. These attacks can be carried out by malware that deletes or modifies data or by unauthorized users manipulating information for various purposes.

Attackers may be interested in acquiring information about the training data, leading to Data Privacy attacks, or gaining insights into the machine learning model itself, resulting in Model Privacy attacks.

An adversary's attempt to furnish an altered input to a classification system can lead to an inaccurate output classification. The objective of the adversary is inferred from the incorrectness of the model. Depending on the impact on the integrity of the classifier output, adversarial goals can be categorized into Confidence Reduction or Misclassification. In Confidence Reduction, a genuine image of a 'stop' sign may be predicted with lower confidence, indicating a reduced probability of belonging to the correct class. In misclassification, a genuine image of a 'stop' sign would be predicted as a class different from the actual 'stop' sign class.

1.3.3 Attacker Capabilities

An adversary could leverage six types of capabilities below to attain their objectives:

- Training Data Control: The attacker may seize control of a portion of the training data by inserting or modifying training samples. This capability is utilized in data poisoning attacks.
- Model Control: The attacker might take control of the model parameters by either creating a Trojan trigger and embedding it in the model or sending malicious local model updates in federated learning.
- Testing Data Control: This capability allows attackers to introduce perturbations to testing samples during model deployment to create adversarial examples.
- Label Limit: This capability pertains to restricting adversarial control over the labels of training samples in supervised learning. In clean-label poisoning attacks, it is assumed that the attacker does not have control over the labels of the poisoned samples - a realistic poisoning scenario. In contrast, regular poisoning attacks assume label control over the poisoned samples.
- Source Code Control: The attacker could alter the source code of the machine learning algorithm, including components like the random number generator or any third-party libraries, which are frequently open source.
- Query Access: In scenarios where the machine learning model is managed by a cloud provider(e.g., Machine Learning as a Service - MLaaS), the attacker could submit queries to the model and obtain predictions, including labels or model confidences. This capability is exploited in black-box evasion attacks, energy-latency attacks, and privacy attacks.

1.3.4 Data Modality impacted by attacks

Adversarial attacks on machine learning have been identified across various data modalities employed in numerous application domains. Historically, most attacks focused on a single modality. However, a recent trend in machine learning involves utilizing multimodal data. The prevalent data modalities in the literature on adversarial machine learning include:

Image: Adversarial examples within the image data modality [19] possess the advantage of a continuous domain, enabling the direct application of gradient-based methods for optimization.

Text: Natural Language Processing (NLP) stands as a widely adopted modality, with proposed attacks across various classes, including evasion, poisoning, and privacy [25]. Additionally, audio systems and text generated from audio signals have also been targeted [26].

Cybersecurity: Industrial Control Systems (ICS) and Supervisory Control and Data Acquisition (SCADA) systems constitute integral components of modern Critical Infrastructure (CI), including power grids, power plants (nuclear, fossil fuel, renewable energy), water treatment plants, oil refineries, and more. ICS becomes an appealing target for adversaries due to the potential to cause significant disruptions to critical infrastructure [27].

Tabular Data: Various attacks targeting machine learning models processing tabular data within finance, business, and healthcare applications have been demonstrated.

1.3.5 Attacker Knowledge

Another aspect of attack classification pertains to the level of knowledge the attacker possesses about the machine learning system. The primary attack types based on this dimension are white-box, black-box, and gray-box attacks [21].

White-box attacks: These attacks presuppose that the attacker has complete knowledge of the machine learning system, including details such as the training data, model architecture, and model hyperparameters. Despite operating under very strong assumptions, examining white-box attacks is essential to assess a system's vulnerability against worst-case adversaries and evaluate potential mitigations.

Black-box attacks: Black-box attacks assume minimal knowledge about the machine learning system. In these scenarios, adversaries may have query access to the model but possess no other information about the model's training. These attacks are particularly practical, as they operate under the assumption that the attacker lacks knowledge about the AI system and relies on system interfaces readily available for normal use.

Gray-box attacks: Gray-box attacks span across a spectrum of adversarial knowledge situated between black-box and white-box attacks. [28] introduced a framework for categorizing gray-box attacks. In a gray-box scenario, an attacker may be aware of the model architecture but not its parameters or the attacker may know both the model and its parameters but not the training data. Typical assumptions for gray-box attacks also include the attacker's access to data distributed identically to the training data and knowledge of the feature representation. The latter assumption is particularly pertinent in applications where feature extraction is performed before training a machine learning model, such as cybersecurity, finance, and healthcare.

1.3.5.1 White Box Adversarial Attack Methods

The attacker has knowledge of the attack model's intricacies, including data preprocessing, model structure, and model parameters. White-box attacks represent the most challenging scenario, as attackers have full access to both the model architecture and its weights.

L-BFGS Attack: The L-BFGS method is employed to minimize a differentiable scalar function across unconstrained real-vector values. It utilizes an estimate of the inverse Hessian matrix to guide its exploration through the variable space. In contrast to the BFGS method, which stores a dense n×n approximation of the inverse Hessian (where n is the number of variables in the problem), L-BFGS only retains a few vectors that implicitly represent the approximation. This linear memory requirement makes the L-BFGS method well-suited for optimization problems characterized by a large number of variables.

Fast Gradient Sign Method: The FGSM attack operates by using the gradients of the loss function with respect to the input sample to generate a new sample that maximizes the loss. In its fundamental form, FGSM involves introducing noise (non-random noise), the direction of which aligns with the gradient of the cost function concerning the data.

DeepFool The DeepFool attack is an untargeted white-box attack. [29] investigated the problem of minimally distorted adversarial examples. Instead of computing the gradient of a loss function, DeepFool searches the shortest distance from the original input to the nearest decision boundary. This is achieved through an iterative linear approximation of the decision boundary or hyperplane, along with the orthogonal projection of the input onto the approximated decision boundary.

Carlini and Wagner : Carlini and Wagner's (C&W) attack [30] is designed to identify the minimally perturbed disturbance. Carlini and Wagner transformed the box-constrained optimization problem into an unconstrained optimization problem, making it amenable to standard optimization algorithms instead of relying on the L-BFGS algorithm. Carlini and Wagner explored three distinct methods to eliminate the box constraint

$$x' \in [0, 1]^n$$

where x' represents the adversarial example. These methods include projected gradient descent, clipped gradient descent, and change of variables.

1.3.5.2 Black Box Adversarial Attack Methods

In a black-box attack, the attacker lacks information about the network and the training set. They can only query the probability/confidence score for each class (score-based black box attack) or the label (decision-based black box attack). Additionally, because black-box attack models do not rely on gradient calculation, unlike most white-box attacks, defense strategies based on masking gradients or non-differentiability are ineffective against these attacks.

Black-box attacks can be categorized into three groups: transfer-based attacks, score-based attacks, and decision-based attacks.

Transfer-based Attacks: This attack leverages the transferability of adversarial examples and can be extended to no-box settings, where the attacker lacks access even to the model's output.

[19] emphasized the transferability of adversarial examples, demonstrating that the same adversarial image can be misclassified by various classifiers with different architectures or trained on different datasets. This characteristic proves beneficial in black-box settings. In instances where an attacker lacks knowledge of the model architecture, they can create a substitute model to mimic the target model. Subsequently, they can employ white-box attack methods, such as the L-BFGS attack, on a newly created substitute model to generate an adversarial example.

Score-based Attacks: These attacks work by iteratively querying the model for its predictions and using those predictions to guide the search for an adversarial example that satisfies a specific goal (e.g., misclassified as a specific target class).

- ZOO Attack: Chen et al. [35] introduced a ZOO attack, which directly estimates the gradients of the target model using confidence/probability scores to produce an adversarial example. This method does not require the training of a substitute model. Since the gradient is inadmissible, and one can only use the function evaluation, this makes it a zeroth-order optimization method.

- Square Attack: A black-box attack, despite requiring numerous queries, generally exhibits lower performance compared to a white-box attack. Andriushchenko et al. [26] introduced the square attack (SA), which enhances query efficiency and success rate by utilizing random search and a task-specific sampling distribution.

Decision-based Attack : These attacks rely solely on the label of machine learning output for adversarial attacks, rendering it easily applicable in real-world machine learning scenarios.

- Boundary Attack: Avoiding dependence on both training data and the transferability assumption, the boundary attack employs a simple rejection sampling algorithm. It generates minimal perturbation adversarial samples by utilizing a constrained independent and identically distributed Gaussian distribution as a proposed distribution and incorporating dynamic step-size adjustment inspired by Trust Region methods.

Adversarial Attacks can lead to serious consequences, especially in fields like finance, healthcare, and security, by making incorrect predictions.

1.4 Thesis Contributions

The major contribution of this thesis is evaluating adversarial attacks on speaker recognition systems. We demonstrate four different methods of adversarial attacks: Top-k pixels, frequency band, time band, and using substitute model. Different experimental attack setups for source-to-target speakers include male to female, female to male, or any speaker to any other speaker. We generate adversarial examples using FGSM, a white box method; however, we perform the attack in a black box setting.

Only the backbone model is publicly available. VGGVox architecture is used as a backbone. The substitute model needs to be finetuned using own dataset, which is Librispeech in this thesis. The substitute model is not publicly available. We use CNN architectures to be applied directly to raw spectrograms and trained in an end-to-end manner. There are other methods that use deep neural networks (DNN) as feature extractors combined with classifiers but are not trained end-to-end.

We perform comprehensive experiments to evaluate the robustness of speaker recognition systems to adversarial attacks. Convolutional neural networks are designed for images. Sound images in the form of mel spectrogram are different from visual images in the way that sound objects are "transparent" so that multiple objects can have energy at the same frequency, where a given pixel in a visual image almost always corresponds to only one object. We use the mel spectrogram of audio as an image in our experiments and explore how CNN performs when applied to perturbed mel spectrograms, questioning the role of high accuracy systems while perturbation to the audio remains imperceptible.

1.5 Thesis Outline

Rest of the thesis is organized as follows:

- Chapter 2 discusses the related work and contemporary approaches for speaker systems and adversarial attacks on various systems.
- Chapter 3 presents the dataset and extensive experiments performed to examine the impact of adversarial attacks on speaker recognition systems.
- Chapter 4 summarizes our work and presents the concluding thoughts with future direction on defense against adversarial attacks and limitations of AI with trustworthiness.

Chapter 2

Related Work

Early research in SR systems devoted significant resources to the study of crafting the suitable feature set for speaker recognition [31] [32]. MFCC-like features were primarily used in speaker recognition systems in the pre-deep learning era.

2.1 Factor Analysis and I-vectors

Classifiers using i-vectors [33], a low-dimensional speaker and channel-dependent space learned using factor analysis, achieved state-of-the-art results in Speaker Recognition.

The need for a substantial volume of speech data limited the extensive adoption of speaker verification technology in everyday applications [33]. Many research studies, particularly those centered on Joint Factor Analysis (JFA) and Support Vector Machine-based speaker verification, concentrated on reducing the required amount of speech data while achieving satisfactory performance. As indicated by these investigations, although performance experiences a significant decline in very short utterances for both methodologies, JFA [34] proves to be a preferable option in such circumstances compared to Support Vector Machines [32]

The i-vector approach presented in [35] offers the benefit of utilizing Cosine Similarity Scoring (CSS) kernel for direct verification, streamlining the scoring process and rendering it quicker and less intricate compared to other speaker verification methods, such as JFA or Support Vector Machines (SVM) supervector techniques.

Due to the intricate nature of speech signals, they are highly susceptible to alterations in channel conditions during the acquisition and transmission phases, leading to distortions in the speech signal. Consequently, tracking variability poses a significant challenge for robust speaker verification approaches. Various techniques have been employed in speaker verification to address channel variability, and one of these approaches is Joint Factor Analysis.

As introduced in [36], a front-end factor analysis method known as i-vector extraction originates in JFA (Joint Factor Analysis). I-vectors offer a simplified and highly efficient approach to speaker recognition. It is defined in total variability space [37] and is based on the cosine similarity between low-dimensional vectors. A classifier that has been trained on i-vectors can be applied to a range of

speaker recognition applications. The scores produced by the classifier serve as the basis for making determinations, such as confirming or denying a claimed identity.

2.2 DNN and X-vectors

Subsequently, methods based on x-vector features [38] learned from deep networks outperformed earlier models [39]. X-vectors are extracted from a deep neural network trained for speaker classification [38].

X-vector architecture is designed to extract speaker embeddings from acoustic features, allowing for effective speaker recognition. It captures both the local and global context in speech data, making it suitable for a wide range of speaker recognition tasks. X-vectors are fixed-length representations of variable-length speech segments. They are known to capture speaker characteristics, even when the speakers have not been seen during the DNN training. It is a type of embedding extracted from a Deep Neural Network (DNN) trained to predict the speaker's identity in an input speech recording. The x-vector architecture is based on a time-delay neural network (TDNN). TDNNs are a type of DNN that is well-suited for speech-processing tasks, as they can model the temporal dependencies in speech signals.

The x-vector represents a high-level speaker embedding. Once the x-vectors had been extracted, they were classified with different classification methods like cosine similarity, Linear Discriminant Analysis(LDA) + cosine similarity, or Probabilistic Linear Discriminant Analysis (PLDA) [38].

2.3 CNN and Speaker Recognition

[40] proposed modeling the SR system as an end-to-end convolutional network over 2D representation of Mel frequency intensities. End-to-end CNN models gave promising results and offered a methodically simple solution to SR. Since [40], using convolutions on Mel features has become popular in many speech and audio-based tasks like audio tagging [41], acoustic scene classification [42], speech emotion classification [43], sound event detection [44], and style capture in speech generation [45].

The current forefront of speaker recognition tasks incorporates the utilization of joint factor analysis (JFA) based techniques [46] and i-vectors [47]. Nevertheless, these methodologies depend on a low dimensional representation computed from the audio input like Mel Frequency Cepstrum Coefficients (MFCCs). The degradation in performance of MFCCs in the presence of real-world noise is mentioned in [48] [49]. Furthermore, focusing only on short frames' overall spectral envelope, MFCCs may lack speaker-specific discriminative features like pitch information. This has prompted a transition from handcrafted features to deep learning techniques, particularly Convolutional Neural Networks (CNNs), which are suited for higher-dimensional inputs [50] [51] and find applications in speaker identification [52].

A CNN architecture has the ability to select the pertinent features necessary for the speaker recognition task. This flexibility reduces audio data pre-processing, thereby preventing the loss of valuable

information during this phase. The initial step involves the conversion of all audio into single-channel streams. Subsequently, spectrograms are computed in a sliding window manner using a Hamming window. Mean and variance normalization is applied to each frequency bin of the spectrum. No additional speech-specific pre-processing steps, such as eliminating silence, voice activity detection, or removing unvoiced speech, are employed. These short-time magnitude spectrograms are subsequently utilized as input for the CNN.

Considering that speaker identification under a closed set context can be framed as a multi-class classification problem, an architectural foundation rooted in the VGG-M [53] convolutional neural network (CNN) is used in [40]. This network is known for its good classification performance on image data and has been adapted to accommodate spectrogram inputs. The fully connected fc6 layer is substituted with a pair of layers: one is a fully connected layer designed for operations in the frequency domain, and the other is an average pooling layer tailored to accommodate integral value components. The configuration of the latter depends on the length of the input speech segment. As a result of this modification, the network achieves temporal position invariance but not frequency. Additionally, the output dimensions alignment with the original fully connected layer are kept intact. Moreover, this adjustment reduces the number of parameters, which in turn aids in preventing overfitting.

2.4 Neural Networks and Adversarial Perturbations

Since CNNs are known to be susceptible to adversarial attacks in image classification literature [54], it is natural to ask: How robust are speech and audio systems built with convolutions on spectrograms to adversarial attacks? Especially those like SR that are used in user authentication systems.

The introduction of artificial changes in natural images has been demonstrated to be effective in causing misclassification by deep neural networks(DNNs). Thus, algorithms have been proposed for creating such samples known as ”adversarial images” [19] [55]. The creation of adversarial images involves the addition of a minute, well-tuned additive perturbation that is invisible to human vision, to a correctly classified natural image. Despite its subtle nature, this tiny modification can lead the classifier to predict the altered image as an entirely different class.

Many prior adversarial attack methods have not explored highly constrained scenarios. In those prior works, the number of altered pixels is substantial, potentially leading to perceptibility by human vision [55], as depicted in Figure 2.1 [54]. Moreover, delving into generating adversarial images under highly constrained conditions can provide fresh insights into the geometrical attributes and overall behavior of DNN models in high-dimensional spaces [56].

The intricate nature of high-dimensional spaces has led to relatively less exploration of the geometrical characteristics of DNN boundaries. Nevertheless, assessing the resilience of DNNs concerning adversarial perturbations can offer insights into this complex problem [56]. For instance, both random and natural images are found to be susceptible to adversarial perturbations. If we presume these images

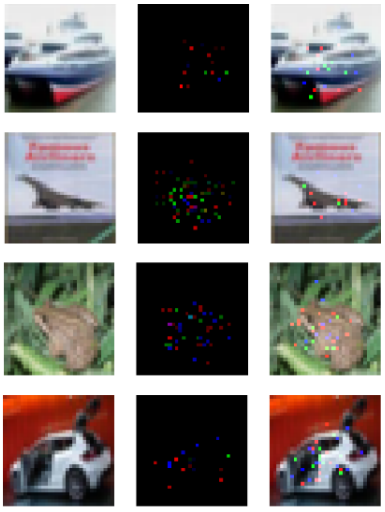


Figure 2.1 Adversarial images with altered pixels perceptible to human [55]

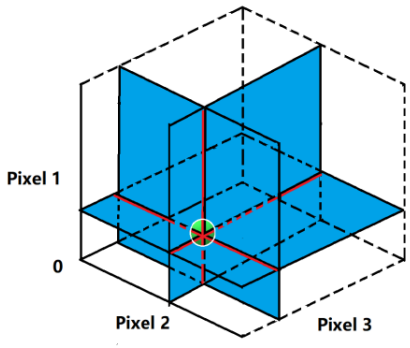


Figure 2.2 Illustration of one and two-pixel perturbation attack in 3-dimensional input space [54]

are uniformly distributed, it implies that a substantial portion of data points in the input space tends to cluster near the boundaries [56].

The one-pixel alteration can be interpreted as adjusting the data point in a direction that aligns with one of the n dimensions. Likewise, three or five-pixel modification shifts the data points within three or five-dimensional regions. In essence, a few-pixel attack involves perturbing the input space along lower-dimensional sections. In fact, a one-pixel perturbation enables the modification of an image in any of the n potential directions with a customizable magnitude. This is depicted in Figure 2.2

Different speaker recognition methods have been studied for their adversarial robustness. A detailed survey of existing research on the robustness of speaker recognition systems is presented in [57]. Like other systems built upon neural networks, speech and speaker recognition systems can be susceptible to attacks involving perturbed inputs. Nonetheless, the end-to-end structure of speech and speaker systems, along with the distinctive nature of their inputs, make attacks against them significantly different from those that occur in the domain of image-based systems [57].

The progress in neural network technology has played a crucial role in making Voice Processing Systems(VPSes) a viable solution. While various architectures, such as Hidden Markov models, have been utilized in the past, systems built upon neural networks are now gaining prominence. Although neural networks have led to considerable enhancements in transcription and identification accuracy, ample literature in the domain of adversarial machine learning has revealed their vulnerability to a diverse range of attacks.

Specifically, the research community has dedicated considerable effort to illustrate that image classification and, only recently, Voice Processing Systems (VPSes) built upon neural networks have shown that they can be exploited through slight alterations to their inputs.

The potential to extend the existing research on attacks from the image domain to the audio domain presents several key differences. Speech recognition pipelines differ considerably from those utilized in image recognition due to data preprocessing before sending as input to a machine learning model. While image classifiers typically work directly with the raw pixel data encoding the image, audio classifiers frequently depend on feature extraction components not necessarily learned from the training data. Instead, these features are derived through signal processing algorithms that human experts manually encode [58].

Given the sequential nature of audio data, machine learning models designed for audio-related tasks are trained using architectures that incorporate statefulness. This enables them to capture contextual dependencies within an audio sample. While recurrent neural architectures facilitate the precise analysis of audio signals, the temporal dimension adds complexity to conducting successful attacks against Voice Processing Systems (VPSes). To create a perturbation, an adversary must account for all time steps simultaneously, a process known as "unrolling" in machine learning terminology. Although unrolling permits backpropagation through time, it complicates the optimization process, potentially leading to issues such as exploding or vanishing gradients [59]. Since gradient-based algorithms depend on this optimization for convergence, discovering adversarial samples becomes more challenging in such scenarios.

The sole optimization attack that has showcased transferability, albeit in a constrained manner, is the Commander Song [60]. The authors successfully transferred samples created for the Kaldi ASR [61] to iFlytek [62], but their attempts to transfer these samples to DeepSpeech [63] proved unsuccessful.

Furthermore, a rising number of vendors are shifting towards Neural Network (NN) based systems, departing from the Hidden Markov Models (HMMs) that the Kaldi ASR internally utilizes. Consequently, the consideration of transferability in the context of optimization attacks becomes more pertinent, particularly in the context of Neural Network based systems.

[64] studies the vulnerability of x-vector based SR system by developing adversarial samples that remain effective while being played over-the-air. This is achieved by modeling the air channel using the estimated room impulse response, then convolving it with the input, followed by FGSM. [65] also shows that x-vector based SR system can change speakers across genders without perceptible distortion in the input.

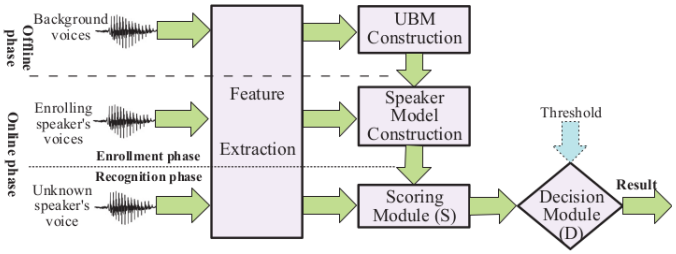


Figure 2.3 Overview of a Speaker Recognition System [67]

An over-the-air adversarial attack is carried out in real-world scenarios, as detailed in [64]. This attack involves playing adversarial examples through a loudspeaker with the intent of undermining the functionality of speaker recognition devices. Leveraging the estimated room impulse response (RIR) encapsulates acoustic channel state information to create practical audio adversarial examples in the physical environment. The suggested adversarial attack presents several advantages compared to traditional replay and synthesis attacks: (1) The proposed attack eliminates the need for the adversary to gather audio clips from the victim, rendering it more practical and easier to execute. (2) The attack allows the generated speech to be recognized as any speaker within the enrolled set, as the adversary desires, without requiring the adversary to acquire voice data from each individual speaker explicitly.

The research conducted in [65] focuses on generating adversarial perturbations in audio that are imperceptible to human senses instead of introducing slight noise to the clean speech samples. To achieve this, they employ the concept of frequency masking, which pertains to the event wherein a faint but audible sound (the ”maskee”) becomes inaudible when a louder audible sound (the ”masker”) is present [66]. This helps in adapting adversarial perturbations to be inaudible, provided that the perturbation remains below the masking threshold of the original speech.

[67] studies the robustness of SR systems to adversarial attacks to understand their security weakness in the practical black-box setting. They propose an adversarial attack and explore its efficacy on i-vector based systems, GMM-UBM systems, and x-vector based systems.

Their investigation delves into the adversarial attack on the three facets of Speaker Recognition Systems (SRSs): (1) Exploring attacks in a more realistic black-box setting. (2) Analyzing the effectiveness of attacks on the open-set identification task [68], which inherently encompasses both close-set identification and speaker verification. (3) Assessing the effectiveness of attacks when deployed over-the-air in real-world physical environments.

Attacks conducted in a black-box setting are often more practical but also more demanding compared to existing white-box attacks [69] [70]. It is worth highlighting that the scoring and decision-making mechanisms of Speaker Recognition Systems (SRSs) exhibit variations across different recognition tasks [71] as depicted in Figure 2.3

While the effectiveness of adversarial attacks on image recognition systems has been extended to speech recognition systems, both in white-box scenarios [26] [72] and black-box settings [73] [74], there has been relatively limited research conducted on Speaker Recognition Systems (SRSs). The speech signal of an utterance fundamentally comprises two main components: the speaker's characteristics and the textual content. In speech recognition, the primary goal is to reduce speaker-dependent variations to accurately decipher the underlying text or command. On the other hand, in speaker recognition, phonetic variations are regarded as extraneous noise, and the focus is on identifying the source of the speech signal. Consequently, adversarial attacks customized for speech recognition systems may be less effective when applied to Speaker Recognition Systems (SRSs).

The prior art to be seen as closest to our work is [65], which proposed an imperceptible white-box psychoacoustics-based method to attack an x-vector model. In the context of neural networks and deep neural networks (DNNs), psychoacoustic hiding generally pertains to methods employed for embedding information within an audio signal in a way that remains imperceptible to human listeners. Embedding information may entail altering specific features of the audio signal or mel spectrogram in a manner that leverages psychoacoustic principles, ensuring that the modifications are perceptually indistinguishable by the human ear. We study the adversarial robustness of speaker recognition systems, which are end-to-end CNN based models that take raw spectrograms as input with a softmax layer to predict the speaker label. Our work focuses on making a comprehensive analysis by restricting the attacks in time and frequency bands, limiting the number of editable pixels, and crafting adversarial samples by making minimal assumptions about the model.

Chapter 3

Adversarial Robustness of Mel Based Speaker Recognition Systems

Machine Learning (ML) algorithms are recognized for their ability to discern the inherent patterns in data, enabling them to make decisions independently without explicit instructions. In the field of research and literature, numerous captivating studies have been conducted to comprehend and replicate human sensory reactions, such as those related to speech and vision [75] [76] [77] [78]. Deep architectures have a benefit when solving complex learning problems compared to shallow architectures. Heaping multiple linear and non-linear processing units in a layered approach imparts the capacity to capture intricate representations across various levels of abstraction. Research has demonstrated that deep convolutional neural networks (CNNs), when provided with an ample amount of training data, can learn invariant representations and potentially attain performance levels similar to human capabilities. Beyond their application in supervised learning, deep CNNs offer the opportunity to extract valuable features from extensive sets of unlabeled data. Recent findings highlight the feasibility of transferring various feature levels, spanning both low-level and high-level features, to a general recognition task by harnessing the principles of Transfer Learning (TL) [79] [80].

3.1 Introduction

Convolutional neural networks have matured over the last decade and continue to show performance gains on numerous tasks [81]. However, one major concern with CNN models is that they are susceptible to small input perturbations. Previous works have shown that CNN models can be tricked to misclassify highly confident predictions by adding a small additive targeted noise to the input image [54]. Consequently, adversarial robustness has been an area of interest for deep learning research in recent years, and the robustness to adversarial attacks has become an important factor in evaluating systems. The study of adversarial robustness is crucial in applications where the cost of misclassification is high, such as biometric authentication. With access to the trained model (white-box attacks) or otherwise by multiple trials, an attacker can compute the gradients or its numerical approximation to break the authentication systems. To this extent, distortion-less adversarial attacks are malicious to authentication systems as they make it difficult to debug and trace the attack.

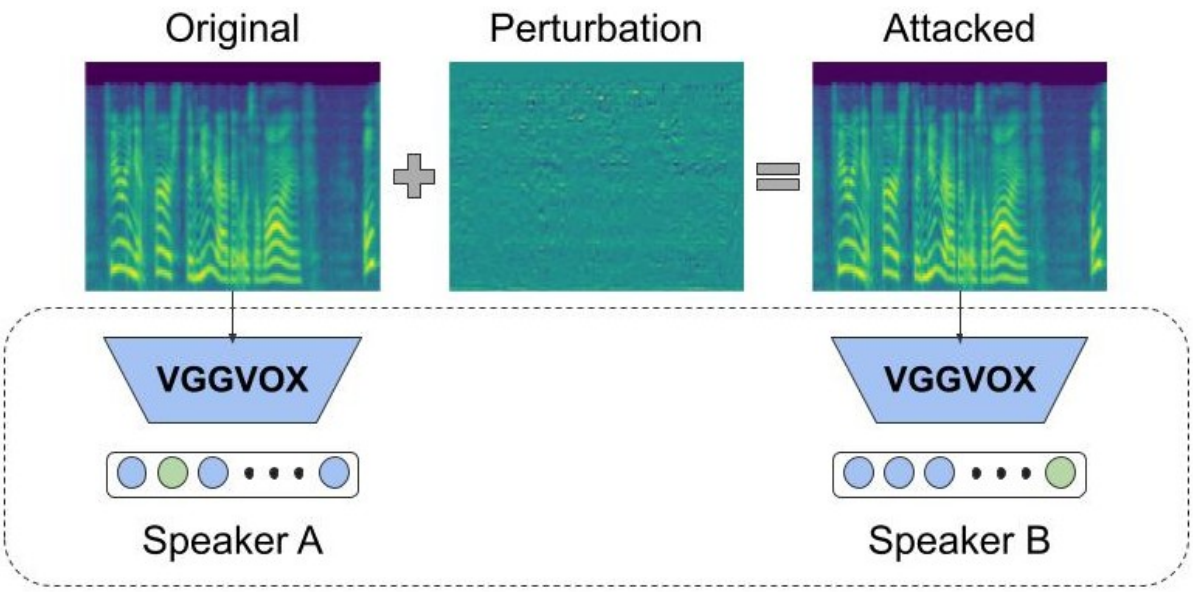


Figure 3.1 Adversarial attack on Speaker Recognition Systems

Speech signals are a popular modality for biometric authentication. Various features and modeling techniques have been proposed over the years to enhance the performance of Speaker Recognition (SR) systems. Earlier approaches relied on classifiers learned on top of common feature representations like i-vector [82] and x-vector [83]. The seminal work by [84] shows that an end-to-end CNN, when applied to raw spectrograms, gives promising results. Using Mel spectrograms as the input also circumvents the need to pre-process the audio data. Further, this approach of learning CNNs on Mel spectrograms has found success in numerous downstream tasks like automatic speech recognition, speaker recognition, and style capture in speech generation.

Audio data, whether presented as spectrograms or other high-dimensional features, encompasses a wealth of information and intricate patterns. Adversarial attacks within this high-dimensional space have the potential to manipulate these intricate patterns and exploit the model's susceptibility to minor changes in input.

An interesting characteristic of adversarial examples is the well-documented transferability from a substitute model to a target model [85]. This feature has been leveraged in the development of transferable black-box attacks. Although I-FGSM enhances the attack strength of FGSM, it comes with a trade-off of reduced transferability. This has given rise to a common perception that the white-box effectiveness of an attack conflicts with its transferability [85]. In our approach, we employ white-box FGSM only to generate adversarial samples while utilizing black-box setting for transferability.

Speaker recognition, among other applications, requires systems to be secure from malicious attacks and have been tested for their robustness to such attacks. Traditional testing methods focused on spoofing attacks like replaying recorded speech [86], mimicking human voice [87] and use of synthesized

speech [88]. More recent forms of attacks attempt to fool models by supplying deceptive or adversarial input *i.e.*, to generate acoustic utterances that sound as one speaker (say speaker A) to the human ear but are classified by the system as another speaker (see Figure 3.1 for an illustration). Adversarial attacks on speaker identification systems have been studied for i-vector and x-vector systems, relying on Mel frequency cepstrum or MFCC features [64, 65, 89]. The adversarial robustness of current state-of-the-art, end-to-end CNN model on raw Mel spectrograms remains largely unexplored.

Our work focuses on this aspect and examines the adversarial resilience of end-to-end trained CNN models designed for speaker recognition. We conduct our experiments with VGGVox architecture [84] using variant of the Fast Gradient Sign Method (FGSM) [90] on LibriSpeech dataset [91]. Our results indicate that by making minimal alterations to the input pixels of the Mel spectrogram, vanilla iterative FGSM attack has the capability to alter the prediction of each speaker to every other speaker label in the selected dataset. We note that corresponding distortion to the audio is in proportion to the number of gradient updates through FGSM.

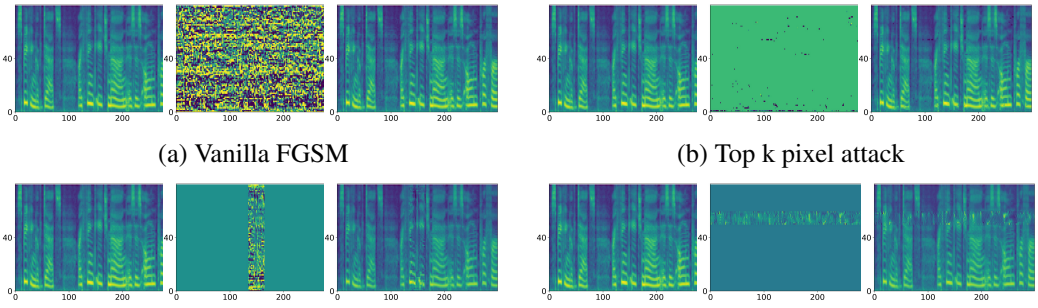
We also evaluate the efficacy of adversarial attacks constrained to specific regions of the spectrogram. Predicted speaker labels can be flipped by limiting changes to a small time band, a small frequency band, or a few pixels (as illustrated in Figure 3.2). We also use a substitute model to compute the adversarial sample to show that the attack can be performed with minimal assumptions on the target model.

FGSM-like attacks become infeasible in the absence of access to the entire model. This mitigates the risk of adversarial attacks in many in-house developed real-world systems. However, since most models are trained using transfer learning, these backbone models are often available in public. We show that adversarial samples can be crafted using a different model with the same backbone. Overall, our work makes the following contributions:

- We explore the robustness of VGGVox architecture to FGSM attack and show that it can change the prediction of a sample to any other class in the given dataset.
- We propose three constrained FGSM attacks and show that predictions can be altered with minimal changes to spectrogram that are mostly imperceptible to the human ear.
- We evaluate the performance of models when the adversarial samples are crafted without access to the actual model but using a different model with the same backbone.
- We present qualitative and quantitative results to demonstrate the efficacy of proposed methods with ablation studies analyzing the role of the number of iterations of FGSM, number of pixels, and different time/frequency bands.

3.2 Approaches for Adversarial Attacks

Adversarial samples are inputs explicitly designed to confuse or mislead a neural network, resulting in the misclassification of the sample. FGSM uses the gradients of the neural network to create



(a) Vanilla FGSM (b) Top k pixel attack

(c) Selected time band (d) Selected frequency band

Figure 3.2 Figure shows the effect of the four attacks used on the spectrogram. For each experiment the three spectrograms in the figure shows: the original spectrogram, the computed perturbation and the attacked spectrogram respectively

the adversarial example. This method is known to achieve the attack through imperceptible additive perturbations.

Given an input spectrogram $\mathbf{x}$ belonging to y, we wish to generate $\tilde{\mathbf{x}} \sim \mathbf{x} + \eta$ with perturbation η that model will misclassify into y^*. FGSM generates perturbations using gradients of the loss J at $\mathbf{x}$ using y^* to generate $\tilde{\mathbf{x}}$ as,

$$\tilde{\mathbf{x}} = \mathbf{x} + \epsilon \operatorname{sign}\left(\nabla_{\mathbf{x}} J(\mathbf{x}, y^*; \theta)\right) \tag{3.1}$$

where θ is the network parameter and ϵ is a hyper-parameter controlling magnitude of the perturbation. We apply FGSM iteratively in our experiments with initialization $\tilde{\mathbf{x}}_0 = \mathbf{x}$ as,

$$\tilde{\mathbf{x}}_t = \tilde{\mathbf{x}}_{t-1} + \epsilon \operatorname{sign}\left(\nabla_{\tilde{\mathbf{x}}_{t-1}} J(\tilde{\mathbf{x}}_{t-1}, y^*; \theta)\right) \tag{3.2}$$

and terminating the loop when the predicted class label is flipped to the desired $y^* = \operatorname{argmax}_y f(\tilde{\mathbf{x}}^*; \theta)$. We propose variants of iterative FGSM , all of which can be considered as different masks applied on the gradient before applying on $\tilde{\mathbf{x}}$. The masked variations is,

$$\tilde{\mathbf{x}}_t = \tilde{\mathbf{x}}_{t-1} + \epsilon \left(\mathbf{M}_{t-1} \odot \operatorname{sign}\left(\nabla_{\tilde{\mathbf{x}}_{t-1}} J(\tilde{\mathbf{x}}_{t-1}, y^*; \theta)\right)\right) \tag{3.3}$$

where $\mathbf{M}_{t-1}$ is the binary mask and $\odot$ represents element-wise product. We evaluate three different masks in this work.

Top-k pixels: In this variant, we restrict perturbations to a fraction k of all pixels in the spectrogram $\tilde{\mathbf{x}}$ with the largest gradient magnitude in the first iteration. Mask $[\mathbf{M}]_{ij} = 1$ is computed at the first iteration consisting of pixels with magnitude of the corresponding entry in gradient $[\text{abs}(\nabla_{\tilde{\mathbf{x}}} J(\tilde{\mathbf{x}}, y^*))]_{ij} > \tau$ and $[\ \cdot\ ]_{ij}$ is indexing operation into (i, j) element. Threshold τ is chosen such that $k * N$ pixels are allowed to be altered. N is the total number of pixels. In subsequent iterations, only pixels corresponding to $[\mathbf{M}]$ are updated. The resulting mask satisfies $\mathbf{1}^\top \mathbf{M} \mathbf{1} = k * N$ where $\mathbf{1}$ is a vector of ones.

Frequency band: In this case, we allow the attack only on a chosen frequency band of the spectrogram. Psychoacoustic hiding is useful in frequency band masking for adversarial attacks. The frequency band is a horizontal strip in the spectrogram, as shown in Figure 3.2(c). The mask here is a fixed matrix across all iterations set to $[\mathbf{M}_t]_{ij} = 1$ only for $i \in [\tilde{\nu} - \nu/2, \tilde{\nu} + \nu/2]$ for a chosen frequency $\tilde{\nu}$ and window length of ν around it. We report experiments for various frequencies $\tilde{\nu}$.

Time period: Similar to the frequency band restriction, we consider limiting attack to a specific time span of the speech signal. This corresponds to constraining the attack to a vertical strip of the spectrogram and the mask is fixed across all iterations to $[\mathbf{M}_t]_{ij} = 1$ only for $j \in [\tilde{\rho} - \rho/2, \tilde{\rho} + \rho/2]$ for a chosen time $\tilde{\rho}$ with window ρ.

Using a substitute model: We evaluate the adversarial robustness of SR models in two different scenarios. We train a model(M_1) on N speakers and craft an adversarial sample using this. The target model (M_2) has n speakers, such that $n << N$. M_2 is fine-tuned with the backbone of M_1. In the first setting, the two sets of speakers used to train source model M_1 and target model M_2 are mutually exclusive and exhaustive. In the second setting, one of the speakers is common between the training sets of M_1 and M_2.

3.3 Experiments

3.3.1 Dataset

The LibriSpeech corpus comprises audiobooks from the LibriVox project, aligning and segmenting read speech with corresponding text while automatically excluding segments with noisy transcripts. The majority of audiobooks originate from Project Gutenberg. Dev and test data are split into 'clean' and 'other.' Additionally, the corpus includes n-gram language models and corresponding texts selected from Project Gutenberg. LibriSpeech has been employed in diverse research projects, spanning speech recognition, speaker identification, and language modeling. In our research, we leverage the dataset to assess the vulnerability of speaker recognition systems to adversarial attacks. We experiment with a CNN speaker recognition model as in [84] trained on LibriSpeech [91] corpus. While the complete LibriSpeech corpus consists of about 1000 hours of speech from 2400 speakers, we use only a clean subset of 360 hours comprising 927 speakers. We sample the waveforms at 22KHZ and convert them into Mel spectrograms with 80 Mel channels by short-time Fourier transform using a Hamming window of length 1024 and hop length of 256. We train the speaker discriminator with a simple training strategy using no augmentation.

3.3.2 Attack Procedure

We perform our FGSM attacks using the aforementioned trained model. For all our experiments, we compare the audio generated from the attacked Mel spectrogram with the audio generated from the original spectrogram using MelGan [92]. Also, we limit the maximum number of perturbation iterations

allowed to 5000. In other words, we label the adversarial attack as unsuccessful if the network prediction is not converted to the target label within 5000 updates.

Vanilla iterative attack vs. Top-k pixel attack: In the first experiment, we apply the vanilla iterative FGSM method described in Equation 3.2 on one random sample of each 927 speakers. Each sample is perturbed to change the prediction to that of every other speaker in the dataset. We report the average number of iterations taken to change the label for each speaker. In the second experiment, we evaluate the efficacy of the top-k pixels attack strategy for two distinct values, $k = 1.25\%$ and $k = 2.5\%$. We measure the Perceptual Evaluation of Speech Quality (PESQ), Signal to Noise Ratio (SNR), and Mean Square Error (MSE) between the audio from the attacked Mel spectrogram and the audio from the original spectrogram for quantitative evaluation.

Frequency band and time period attack: We choose 50 speakers from the dataset for restricted frequency and time band adversarial attack experiments. We use stratified sampling to include the same number of male and female speakers for this subset. We attack a sample from each speaker from this subset with all other speakers as target label. We choose a mask of width 1 KHz and evaluate the attack on ten such non-overlapping positions. Frequency values $\tilde{\nu}$ are chosen such that the 10 non-overlapping windows span the entire width of the spectrogram ranging from 0-8 KHz. In the fourth experiment, we evaluate the efficacy of time segment based attacks. We evaluate time segment attack by choosing a mask of width ρ as small as 2 frames to a maximum of 30 frames. We position the mask at the spectrogram's beginning, middle, and end.

Using a substitute model: We perform two experiments to evaluate the adversarial robustness of models when the adversarial sample is crafted using a substitute model. In the first experiment, we show the transferability of attacks using samples crafted from a model trained on a different set of labels. This experiment evaluates the adversarial robustness when a model with the same backbone is publicly available for crafting adversarial samples. We use 800 speakers to train the source model (M_1) and 121 different speakers to train the target model (M_2). M_2 is initialized with the backbone of M_1, and the backbone weights are frozen during the training. There are no common speaker identities between the trained set of M_1 and M_2, and we attack using the M_1 model to change the label to a speaker from the opposite gender and evaluate it on M_2. We flip between the two gender labels provided in the LibriSpeech dataset.

In this attack, we assume we cannot access the model M_2. For a sample $s \in testdata(M_2)$, we make a prediction using M_1. We then apply an iterative FGSM attack in M_1 to flip the identity so that the gender label flips. We call the attacked sample $\hat{s}$, following $(gender(M_1(s)) \neq gender(M_1(\hat{s})))$. We evaluate $\hat{s}$ on M_2 and call the attack successful if the prediction changes, in other words, $M_2(s) \neq M_2(\hat{s})$.

In the second experiment, we use the source model trained on 800 speakers and fine-tune the model on a new set of 127 speakers with one speaker C_s common to the substitute and source models. We perform targeted FGSM attack using the source model, such that the identity of every sample in the target model is changed to C_s.

3.3.3 Audio Quality Evaluation

We perform a human survey to qualitatively measure the distortion in the audio generated by the attacked Mel spectrogram. For each of our four experiments, we ask fifty users to evaluate twenty pairs of audio each. Each pair constitutes one audio generated from the original Mel spectrogram and the other generated from the attacked Mel spectrogram. We ask the participants to rate the distortion in the attacked audio compared to the former on a Likert scale of three, with zero corresponding to no distortion at all, one for minor distortions, and two for significant distortions. We report the participants' Mean Opinion Score (MOS) for all four experiments.

3.4 Results and Discussion

Vanilla iterative attack vs. Top-k pixel attack: On training the CNN [84] model on the selected 927 speakers of LibriSpeech [91] corpus, we get an accuracy of 98.7% on the test set with an average confidence of 93% for prediction. Despite the network having such high accuracy and high confidence in the prediction, Table 3.1 shows that vanilla iterative FGSM can change samples from any 927 speakers to every other speaker in the dataset with minimal changes. The distortions in the audio caused by the attack are imperceptible in most cases, with a mean MOS of 0.08 in distinguishing the speech generated from original and attacked spectrograms. We also show the results for the fifty samples that incurred the most iterations in Table 3.1. We observe higher distortion in these samples, with mean MOS jumping to 0.79. This correlation between the number of iterations and distortion can be observed from both qualitative and quantitative results Table 3.1.

Table 3.1 further shows that allowing attack on only 2.5% pixels can still give an attack success rate of 89.2% in flipping every speaker to any other speaker in the dataset. It gives no significant added distortion to the audio as well. With a lower k, as low as 1.25% (about 300 pixels), we observe that 69% of the attacks are successful. As the MOS values indicate, both these attacks are almost imperceptible to humans.

To strengthen the claim of adversarial vulnerability of these models, we train the model with data augmentation (SpecAugment [93]) for randomly selected 200 speakers. We observe no significant improvement in robustness in simple iterative FGSM. We observe that the attack success rate remains the same post-augmentation, with the average number of iterations being 123.19 (without augmentation,

Table 3.1 Qualitative and quantitative evaluation. ASR refers to attack success rate and ITER. is the number of iterations.

	ASR	ITER.	MOS	Quantitative Results		
				PESQ	SNR	MSE
V-FGSM	100%	61	0.08	3.71	57.72	10.21
V-FGSM for 50 largest t	100%	182	0.79	2.56	59.17	17.34
Top-$k, k = 1.25$	63.9%	1401	0.17	3.20	56.47	10.91
Top-$k, k = 2.5$	89.2%	1048	0.15	3.22	58.41	12.65

the average number of iterations for the selected two hundred speakers is 113.11). The corresponding distortions are imperceptible.

Time period and frequency band attack: When the number of pixels modified is restricted to a time band as small as 2 frames (equivalent to 23ms), the average MOS score across different placements of the time band is 1.05. Table 3.4 shows that the distortion is more perceptible when the time band appears in the middle of the audio clip compared to the beginning or end. In our study, the MOS when the mask is placed in the center is at least 50% more than the next best placement. It is interesting to note that the model prediction changes with a distortion limited to just 2ms output of audio that spans three seconds. The distortion becomes less (and almost imperceptible) when the time window is increased to 30 frames (or 0.34s) with a mean MOS of 0.6.

Table 3.3 shows that the distortion caused by restricting attacks to a specific frequency band is almost imperceptible irrespective of where we place the band except on the two bands on the extreme end of the spectrum. We hypothesize that since the lower frequencies(0-1 KHz) have higher activity with higher amplitudes, they are more perceptible with even minor distortions leading to significant audible differences. Our experiments show that with a sufficiently large band of 1 KHz, we can successfully attack the system in the range of 2-7 KHz without incurring much distortion in the audio.

Using a substitute model: On training the CNN [84] model such that target model (M2) has 121 speakers fine-tuned with the backbone of model M1 pre-trained on the selected 800 speakers of LibriSpeech [91] corpus, we get an accuracy of 100%. Despite the network having such high accuracy, Table 3.2 shows that iterative FGSM can change samples from any male speaker to every other female speaker in the dataset and vice versa with minimal change.

Transferability of attacks: We show the transferability of attacks using samples crafted from publicly available pre-trained models. We create a substitute model by training speaker detection on a portion of speakers. We fine-tune the model on a new set of 121 speakers with no speaker common to the substitute model label set. Here, we observe an attack success rate of 84.2% when the attack sample is crafted using a substitute pre-trained model, and the attack is performed on the fine-tuned target model. The attack success rate is 91.6% when the source speaker is male and the target speaker is female. Similarly, the attack success rate is 85.2% when the source speaker is a female from the set of 800 speakers, and the target speaker is a male from the set of 121 speakers.

Table 3.2 Qualitative and quantitative evaluation for Substitute model. ASR refers to attack success rate and ITER. is the number of iterations. The rows show Iterative FGSM for substitute model for male to female and female to male respectively

	ASR	ITER.	MOS	Quantitative Results	
				PESQ	MSE
(F to M)	85.2%	112	0.14	3.12	0.0023
(M to F)	91.6%	110	0.26	2.74	0.0037

Table 3.3 MOS evaluation of different frequency bands.

	ν_7	ν_6	ν_5	ν_4	ν_3	ν_2	ν_1	ν_0
MOS	0.82	0.07	0.10	0.28	0.17	0.14	0.21	0.60
PESQ	3.61	3.84	3.48	3.36	3.17	3.2	2.8	2.6

Table 3.4 MOS evaluation of different time bands.

	Start	Middle	End
$\rho = 2$	0.92	1.38	0.84
$\rho = 30$	0.44	0.9	0.46

We also fine-tune the model on a new set of 200 speakers with one speaker C_s common to the substitute model label set. We observe an attack success rate of 60% (with attack target C_s) when the attack sample is crafted using a substitute pre-trained model, and the attack is performed on the fine-tuned target model. This shows that without access to the fine-tuned model, we can successfully attack a model with assumptions about the backbone.

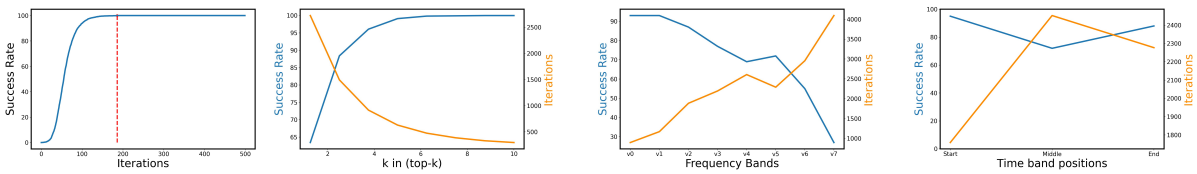


Figure 3.3 Iterations vs. attack success rate trade-offs. See Section 3.4 on Results and Discussion for description, to be viewed in color.

Attack Cost vs. Success Rate In this section, we discuss the trade-off between the Attack Success Rate (ASR) and the cost incurred for an attack measured in terms of the iterations required for the attack to succeed. In each run, we fix the maximum number of iterations required. We report the success rate over a range of iterations. Figure 3.3(a) shows the attack success and the iterations required for a vanilla iterative FGSM. With just 186 iterations, we can achieve a 100% success rate with an average of 61 iterations across samples. For the Top-k pixel attack, we can observe from Figure 3.3 (b) that with an increase in k, the attack success rate increases, which suggests it is easy to attack if we can modify more pixels. It can also be observed that attack success rates and iterations required are inversely correlated, suggesting that if it is easy to attack a sample, it would require fewer iterations, and the attack success rate would be higher. Conversely, if it is difficult to attack, the attack success rate would be less and require a large number of iterations.

We study the success-rate vs. iteration trade-off in frequency band and time band restricted attacks. Figure 3.3(c) shows that the attack success rates vary across frequency bands. When limiting the number of iterations to 5000, the attack success rate largely decreases with frequency while the number of

iterations required increases. Figure 3.3(d) shows the impact of the adversarial attack on time bands positioned at different positions (Start, Middle, and End) for a time band of 30 frames. The attack success rate remains above 75% across all three positions, while the number of iterations required is lower for the first position.

Chapter 4

Conclusion

The exploration of adversarial robustness performed for state-of-the-art speaker recognition system built from convolutional operations involved investigating the vulnerability to adversarial attacks. These attacks create specially crafted inputs that fool the system into producing incorrect outputs. We note that the vanilla version of FGSM attack can effectively change the speaker label of every speaker to any other sample without noticeable alterations to the audio. Further, the distortion in audio is proportional to the number of gradient updates in FGSM but is not correlated to the distance between the embeddings learned by the speaker discriminator. We also evaluate the efficacy of adversarial attacks when the attack is restricted to a small set of pixels or a small band of time or frequency within the spectrogram, and empirical results show that it is possible to trick the classifier without significant perturbations to the input. We would like to study further if these observations hold to the numerous downstream tasks, which are very likely, and explore avenues for improving these high-accuracy systems to be more robust to such attacks.

4.1 Challenges and Future Work

Adversarial threats present a risk to the integrity of deep learning models. Nevertheless, there is a scarcity of research on methods for detecting adversarial attacks, particularly within the multi-modal domain. Present methodologies predominantly concentrate on image classification models, emphasizing a necessity for further investigations into diverse modal types, such as Speech and Natural Language Processing (NLP).

There remains potential for improving the transferability of adversarial examples, particularly by developing more efficient transferable attacks and gaining a deeper insight into the essential causes of transferability. Moreover, finding a balance between perturbation visibility and attack success is important for devising effective adversarial attack methods. Conversely, it is crucial to formulate novel strategies for safeguarding against transferable adversarial attacks and develop defenses against them accordingly. Introducing techniques for detecting and mitigating attacks and exploring the use of adversarial examples for model improvement is a relevant direction for future work. Diverse defense

mechanisms can be aggregated into an ensemble solution to account for the lack of diversity observed in a single defense mechanism. Focus is also needed on reducing attack costs.

4.2 Trustworthiness

A machine learning system that exhibits accuracy but is easily vulnerable to adversarial attacks is unlikely to inspire trust. Similarly, a machine learning system that generates detrimentally biased or unfair outcomes, even if robust, is unlikely to be trusted. In situations where both fairness and privacy are crucial, weighing the trade-off between privacy and fairness becomes essential. Simultaneously maximizing the performance of the AI system concerning these attributes is not feasible. Trade-offs exist between adversarial robustness and explainability as well.

Bibliography

[1] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," *arXiv preprint arXiv:1312.6199*, 2013.

[2] K. Fowler, "Mission-critical and safety-critical development," *IEEE Instrumentation and Measurement Magazine*, vol. 7, no. 4, pp. 52–59, 2004.

[3] J. Gao, M. Galley, and L. Li, "Neural approaches to conversational ai," in *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, ser. SIGIR '18, 2018, p. 1371–1374.

[4] N. Carlini, P. Mishra, T. Vaidya, Y. Zhang, M. Sherr, C. Shields, D. Wagner, and W. Zhou, "Hidden voice commands," in *25th USENIX Security Symposium (USENIX Security 16)*. USENIX Association, 2016.

[5] A. Mehrish, N. Majumder, R. Bharadwaj, R. Mihalcea, and S. Poria, "A review of deep learning techniques for speech processing," *Information Fusion*, vol. 99, p. 101869, 2023.

[6] L. Wyse, "Audio spectrogram representations for processing with convolutional neural networks," *arXiv preprint arXiv:1706.09559*, 2017.

[7] Z. K. Abdul and A. K. Al-Talabani, "Mel frequency cepstral coefficient and its applications: A review," *IEEE Access*, vol. 10, pp. 122 136–122 158, 2022.

[8] A. Oppenheim and R. Schafer, "Homomorphic analysis of speech," *IEEE Transactions on Audio and Electroacoustics*, vol. 16, pp. 221–226, 1968.

[9] D. A. Reynolds, "Automatic speaker recognition using gaussian mixture speaker models," in *The Linkoln Laboratory, MIT*, 1995.

[10] M. K. Nandwana, L. Ferrer, M. McLaren, D. Castán, and A. D. Lawson, "Analysis of critical metadata factors for the calibration of speaker recognition systems," in *Interspeech*, 2019.

[11] G. Heigold, I. Moreno, S. Bengio, and N. M. Shazeer, "End-to-end text-dependent speaker verification," *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5115–5119, 2015.

[12] S. S. Tirumala and S. R. Shahamiri, "A review on deep learning approaches in speaker identification," in *International Conference on Signal Processing Systems*, 2016.

[13] D. Snyder, D. Garcia-Romero, G. Sell, A. McCree, D. Povey, and S. Khudanpur, "Speaker recognition for multi-speaker conversations using x-vectors," *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5796–5800, 2019.

[14] G. Chen, S. Chenb, L. Fan, X. Du, Z. Zhao, F. Song, and Y. Liu, "Who is real bob? adversarial attacks on speaker recognition systems," in *2021 IEEE Symposium on Security and Privacy (SP)*, 2021, pp. 694–711.

[15] O. Abdel-Hamid, A. rahman Mohamed, H. Jiang, L. Deng, G. Penn, and D. Yu, "Convolutional neural networks for speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, pp. 1533–1545, 2014.

[16] N. Simic, S. Suzic, T. V. Nosek, M. Vujovic, Z. H. Perić, M. S. Savic, and V. Delić, "Speaker recognition using constrained convolutional neural networks in emotional speech," *Entropy*, vol. 24, 2022.

[17] S. Kriman, S. Beliaev, B. Ginsburg, J. Huang, O. Kuchaiev, V. Lavrukhin, R. Leary, J. Li, and Y. Zhang, "Quartznet: Deep automatic speech recognition with 1d time-channel separable convolutions," *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6124–6128, 2019.

[18] J. S. Chung, A. Nagrani, and A. Zisserman, "Voxceleb2: Deep speaker recognition," in *Interspeech*, 2018.

[19] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *3rd International Conference on Learning Representations, ICLR San Diego, CA, USA, Conference Track Proceedings*, Y. Bengio and Y. LeCun, Eds., 2015.

[20] T. Tanay and L. D. Griffin, "A boundary tilting persepective on the phenomenon of adversarial examples," *ArXiv*, 2016.

[21] A. Oprea and A. Vassilev, "Adversarial machine learning: A taxonomy and terminology of attacks and mitigations," *National Institute of Standards and Technology*, 2023.

[22] B. Biggio, B. Nelson, and P. Laskov, "Poisoning attacks against support vector machines," in *Proceedings of the 29th International Coference on International Conference on Machine Learning*, 2012.

[23] Y. Liu, S. Ma, Y. Aafer, W.-C. Lee, J. Zhai, W. Wang, and X. Zhang, "Trojaning attack on neural networks," in *Network and Distributed System Security Symposium*, 2018.

[24] B. Biggio, I. Corona, D. Maiorca, B. Nelson, N. Šrndić, P. Laskov, G. Giacinto, and F. Roli, *Evasion Attacks against Machine Learning at Test Time*. Springer Berlin Heidelberg, 2013.

[25] S. Zanella-Béguelin, L. Wutschitz, S. Tople, V. Rühle, A. Paverd, O. Ohrimenko, B. Köpf, and M. Brockschmidt, "Analyzing information leakage of updates to natural language models," in *Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security*, 2020.

[26] N. Carlini and D. Wagner, "Audio adversarial examples: Targeted attacks on speech-to-text," in *2018 IEEE Security and Privacy Workshops (SPW)*, 2018, pp. 1–7.

[27] T. Conway, R. M. Lee, and M. J. Assante, "Analysis of the cyber attack on the ukrainian power grid. defense use case," in *Technical report*. SANS ICS, 2016.

[28] O. Suciu, R. Marginean, Y. Kaya, H. Daumé, and T. Dumitras, "When does machine learning fail? generalized transferability for evasion and poisoning attacks," *ArXiv*, 2018.

[29] S.-M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard, "Deepfool: A simple and accurate method to fool deep neural networks," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[30] N. Carlini and D. A. Wagner, "Towards evaluating the robustness of neural networks," *CoRR*, 2016.

[31] K. Murty and B. Yegnanarayana, "Combining evidence from residual phase and mfcc features for speaker recognition," *IEEE Signal Processing Letters*, vol. 13, 2006.

[32] M. McLaren, R. Vogt, B. Baker, and S. Sridharan, "Experiments in svm-based speaker verification using short utterances," in *The Speaker and Language Recognition Workshop*, 2010.

[33] A. Kanagasundaram, R. Vogt, D. Dean, S. Sridharan, and M. Mason, "i-vector based speaker recognition on short utterances," in *Interspeech*, 2011.

[34] R. Vogt, B. Baker, and S. Sridharan, "Factor analysis subspace estimation for speaker verification with short utterances," in *Interspeech*, 2008.

[35] N. Dehak, R. Dehak, J. Glass, D. Reynolds, and P. Kenny, "Cosine Similarity Scoring without Score Normalization Techniques," in *Proc. The Speaker and Language Recognition Workshop (Odyssey 2010)*, 2010, p. paper 15.

[36] S. Shum, N. Dehak, R. Dehak, and J. R. Glass, "Unsupervised speaker adaptation based on the cosine similarity for text-independent speaker verification." in *Odyssey*, 2010, p. 16.

[37] N. Dehak, "Discriminative and generative approaches for long- and short-term speaker characteristics modeling: Application to speaker verification," Ph.D. dissertation, Ecole de Technologie Superieure Montreal , Canada, 2009.

[38] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5329–5333.

[39] D. Snyder, D. Garcia-Romero, G. Sell, A. McCree, D. Povey, and S. Khudanpur, "Speaker recognition for multi-speaker conversations using x-vectors," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 5796–5800.

[40] A. Nagrani, J. S. Chung, and A. Zisserman, "Voxceleb: A large-scale speaker identification dataset," in *Interspeech*, 2017.

[41] J. Pons, O. Nieto, M. Prockup, E. M. Schmidt, A. F. Ehmann, and X. Serra, "End-to-end learning for music audio tagging at scale," *ArXiv*, vol. abs/1711.02520, 2017.

[42] Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, and M. D. Plumbley, "Panns: Large-scale pretrained audio neural networks for audio pattern recognition," *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, vol. 28, p. 2880–2894, 2020.

[43] J. Zhao, X. Mao, and L. Chen, "Speech emotion recognition using deep 1d and 2d cnn lstm networks," *Biomedical Signal Processing and Control*, vol. 47, pp. 312–323, 2019.

[44] T. Chen and U. Gupta, "Attention-based convolutional neural network for audio event classification with feature transfer learning," in *Neural Processing Letters*, 2018.

[45] Y. Wang, D. Stanton, Y. Zhang, R.-S. Ryan, E. Battenberg, J. Shor, Y. Xiao, Y. Jia, F. Ren, and R. A. Saurous, "Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis," in *International conference on machine learning*, 2018, pp. 5180–5189.

[46] P. Kenny, "Joint factor analysis of speaker and session variability: Theory and algorithms," in *Computer Science, draft version*, 2006. [Online]. Available: https://api.semanticscholar.org/CorpusID:9214179

[47] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, pp. 788–798, 2011.

[48] U. H. Yapanel, X. Zhang, and J. H. L. Hansen, "High performance digit recognition in real car environments," in *Interspeech*, 2002.

[49] J. H. L. Hansen, R. Sarikaya, U. H. Yapanel, and B. L. Pellom, "Robust speech recognition in noise: an evaluation using the SPINE corpus," in *EUROSPEECH 2001 Scandinavia, 7th European Conference on Speech Communication and Technology, 2nd INTERSPEECH Event, Aalborg, Denmark, September*, 2001, pp. 905–908.

[50] T. N. Sainath, R. J. Weiss, A. W. Senior, K. W. Wilson, and O. Vinyals, "Learning the speech front-end with raw waveform cldnns," in *Interspeech*, 2015.

[51] S. Hershey, S. Chaudhuri, D. P. W. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, M. Slaney, R. J. Weiss, and K. W. Wilson, "CNN architectures for large-scale audio classification," *CoRR*, vol. abs/1609.09430, 2016.

[52] Y. Lukic, C. Vogt, O. Dürr, and T. Stadelmann, "Speaker identification and clustering using convolutional neural networks," in *2016 IEEE 26th International Workshop on Machine Learning for Signal Processing (MLSP)*, 2016.

[53] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman, "Return of the devil in the details: Delving deep into convolutional nets," *Proceedings of the British Machine Vision Conference*, 2014.

[54] J. Su, D. V. Vargas, and K. Sakurai, "One pixel attack for fooling deep neural networks," *IEEE Transactions on Evolutionary Computation*, vol. 23, pp. 828–841, 2019.

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

[56] A. Fawzi, S.-M. Moosavi-Dezfooli, and P. Frossard, "The robustness of deep networks: A geometrical perspective," *IEEE Signal Processing Magazine*, vol. 34, pp. 50–62, 2017.

[57] H. Abdullah, K. Warren, V. Bindschaedler, N. Papernot, and P. Traynor, "Sok: The faults in our asrs: An overview of attacks against automatic speech recognition and speaker identification systems," *2021 IEEE Symposium on Security and Privacy (SP)*, pp. 730–747, 2020.

[58] S. Sigurdsson, K. B. Petersen, and T. Lehn-Schiøler, "Mel frequency cepstral coefficients: An evaluation of robustness of mp3 encoded music," in *International Society for Music Information Retrieval Conference*, 2006.

[59] R. Pascanu, T. Mikolov, and Y. Bengio, "On the difficulty of training recurrent neural networks," in *Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28*, 2013, p. III–1310–III–1318.

[60] X. Yuan, Y. Chen, Y. Zhao, Y. Long, X. Liu, K. Chen, S. Zhang, H. Huang, X. Wang, and C. A. Gunter, "Commandersong: A systematic approach for practical adversarial voice recognition," in *Proceedings of the 27th USENIX Conference on Security Symposium*, 2018, p. 49–64.

[61] D. Povey. Kaldi aspire chain model. Accessed: 2019. [Online]. Available: http://kaldi-asr.org/models.html

[62] J. Du, Y.-H. Tu, L. Sun, F. Ma, H.-K. Wang, J. Pan, C. Liu, and C.-H. Lee. The ustc-iflytek system for chime-4 challenge. Accessed: 2019. [Online]. Available: https://spandh.dcs.shef.ac.uk/chime_workshop/chime2016/presentations/CHiME_2016_Du_oral.pdf

[63] A. Hannun, C. Case, J. Casper, B. Catanzaro, G. Diamos, E. Elsen, R. Prenger, S. Satheesh, S. Sengupta, A. Coates, and A. Y. Ng. Deep speech 0.4.1. Accessed: 2019. [Online]. Available: https://github.com/mozilla/DeepSpeech/releases/tag/v0.4.1

[64] Z. Li, C. Shi, Y. Xie, J. Liu, B. Yuan, and Y. Chen, "Practical adversarial attacks against speaker recognition systems," in *Proceedings of the 21st International Workshop on Mobile Computing Systems and Applications*, ser. HotMobile '20, 2020, p. 9–14.

[65] Q. Wang, P. Guo, and L. Xie, "Inaudible adversarial perturbations for targeted attack in speaker recognition," in *Interspeech*, 2020.

[66] Y. Lin, W. H. Abdulla, Y. Lin, and W. H. Abdulla, "Principles of psychoacoustics," *Audio Watermark: A Comprehensive Foundation Using MATLAB*, pp. 15–49, 2015.

[67] G. Chen, S. Chen, L. Fan, X. Du, Z. Zhao, F. Song, and Y. Liu, "Who is real bob? adversarial attacks on speaker recognition systems," *2021 IEEE Symposium on Security and Privacy (SP)*, pp. 694–711, 2019.

[68] J. Fortuna, P. Sivakumaran, A. M. Ariyaeeinia, and A. S. Malegaonkar, "Open-set speaker identification using adapted gaussian mixture models," in *Interspeech*, 2005.

[69] Y. Gong and C. Poellabauer, "Crafting adversarial examples for speech paralinguistics applications," *arXiv preprint arXiv:1711.03280*, 2017.

[70] F. Kreuk, Y. Adi, M. Cisse, and J. Keshet, "Fooling end-to-end speaker verification with adversarial examples," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 1962–1966.

[71] H. Beigi, *Fundamentals of Speaker Recognition*. Springer Publishing Company, Incorporated, 2011.

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

[73] R. Taori, A. Kamsetty, B. Chu, and N. Vemuri, "Targeted adversarial examples for black box audio systems," *2019 IEEE Security and Privacy Workshops (SPW)*, pp. 15–20, 2018.

[74] S. Khare, R. Aralikatte, and S. Mani, "Adversarial black-box attacks for automatic speech recognition systems using multi-objective genetic optimization," *CoRR*, vol. abs/1811.01312, 2018.

[75] T. Ojala, M. Pietikäinen, and D. Harwood, "A comparative study of texture measures with classification based on featured distributions," *Pattern Recognition*, vol. 29, no. 1, pp. 51–59, 1996.

[76] D. Lowe, "Object recognition from local scale-invariant features," in *Proceedings of the Seventh IEEE International Conference on Computer Vision*, vol. 2, 1999, pp. 1150–1157 vol.2.

[77] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 1, pp. 886–893 vol. 1, 2005.

[78] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, "Speeded-up robust features (surf)," *Computer Vision and Image Understanding*, vol. 110, no. 3, pp. 346–359, 2008, similarity Matching in Computer Vision and Multimedia.

[79] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, pp. 1345–1359, 2010.

[80] A. S. Qureshi, A. Khan, A. Zameer, and A. Usman, "Wind power prediction using deep neural network based meta regression and transfer learning," *Applied Soft Computing*, vol. 58, pp. 742–755, 2017.

[81] A. Khan, A. Sohail, U. Zahoora, and A. S. Qureshi, "A survey of the recent architectures of deep convolutional neural networks," *Artificial Intelligence Review*, 2020.

[82] A. Kanagasundaram, R. Vogt, D. Dean, S. Sridharan *et al.*, "i-vector based speaker recognition on short utterances," in *Interspeech*, 2011.

[83] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in *ICASSP*, 2018.

[84] A. Nagrani, J. S. Chung, and A. Zisserman, "Voxceleb: a large-scale speaker identification dataset," *Interspeech*, 2017.

[85] A. Kurakin, I. J. Goodfellow, and S. Bengio, "Adversarial machine learning at scale," in *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017.

[86] H.-J. Shim, J.-W. Jung, H.-S. Heo, S.-H. Yoon, and H.-J. Yu, "Replay spoofing detection system for automatic speaker verification using multi-task learning of noise classes," in *TAAI*, 2018.

[87] M. Farrús Cabeceran, M. Wagner, D. Erro Eslava *et al.*, "Automatic speaker recognition as a measurement of voice imitation and conversion," *International Journal of Speech, Language and the Law*, 2010.

[88] Z. Wu, T. Kinnunen, N. Evans *et al.*, "Asvspoof 2015: the first automatic speaker verification spoofing and countermeasures challenge," in *Interspeech*, 2015.

[89] G. Chen, S. Chenb, L. Fan, X. Du *et al.*, "Who is real bob? adversarial attacks on speaker recognition systems," in *SP*, 2021.

[90] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *ICLR*, 2015.

[91] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *ICASSP*, 2015.

[92] K. Kumar, R. Kumar, T. de Boissiere, L. Gestin *et al.*, "Melgan: Generative adversarial networks for conditional waveform synthesis," in *Neurips*, 2019.

[93] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "Specaugment: A simple data augmentation method for automatic speech recognition," *arXiv preprint arXiv:1904.08779*, 2019.