Chaitravi Reddy  
16010422274

***REPRESENTATION LEARNING TO CLASSIFY AND DETECT ADVERSARIAL ATTACKS AGAINST SPEAKER AND SPEECH RECOGNITION SYSTEMS.***

Researchers used Deep Learning to create attack signatures (it captures important characteristics of audio data). It helps to identify the type of attack.  
DARPA RED (Defense Advanced Research Projects Agency Reverse Engineering of Deceptions) Program was used.

**Adversarial Attacks: -**

They involve making small changes to an audio waveform to deceive speech or speaker recognition systems. The attacker adds imperceptible perturbations to the audio data to alter the system's decision without being noticeable to human ears.

1. **Threat Models**

Threat models in adversarial attacks are specific methods or rules used by attackers to manipulate audio systems. To ensure imperceptibility to human ears, researchers employed distance metrics, measuring the magnitude of changes introduced to the audio waveform. These metrics reveal how attackers modify audio data without detection by human listeners.

1. **Attack Algorithms**PGD (Projected Gradient Descent) is an attack that introduces slight alterations to the audio waveform to deceive the classifier by iteratively maximizing misclassification. FGSM (Fast Gradient Sign Method) and Iter-FGSM are simplified versions of PGD, making small changes to fool the classifier. Carlini-Wagner attack aims to find tiny modifications to make the attack harder to detect, enhancing its effectiveness in evading detection.

**Attack Representation Learning: -**

1. **X-Vectors:** They use neural networks to create unique codes from speech, representing the identity or type of attack. The process involves capturing features, combining them into a single code, and calculating attack probabilities. The network learns to recognize various attack types during training.
2. **Application of Attack Signatures:**

a. Attack Classification: Identify normal or attacked audio based on the attack method or threat level using signature extractor networks and unique codes. Matched codes indicate known attack types, while unmatched ones may require additional classifier training.

b. Attack Verification/Detection: Verify if two audio recordings have the same attack technique. For known attacks, use PLDA to compare likelihood. Detect unknown attacks and add them to the defense database for future protection, improving system resilience against novel attacks.

**Speaker and Speech Recognition Tasks: -**

For speaker recognition, VoxCeleb1 and VoxCeleb2 datasets were used. The system calculates similarity scores between speaker audio samples to identify matches. However, the vulnerable x-vector architecture can be manipulated by adversarial attacks, affecting speaker recognition accuracy.

For ASR, the Librispeech dataset was employed to train the Espresso ASR system. It converts spoken words into text. Dataset division into training and test subsets allowed evaluation on unseen data.

**Adversarial attack generation: -**

To test system performance, various attack algorithms (FGSM, PGD) and threat models were employed to create tricky examples that deceive the system, leading to incorrect classification. This process helps develop more robust systems capable of detecting and defending against such attacks.

**Attack Signature Extraction Networks: -**

The researchers used Thin-ResNet34 x-Vector architecture to detect attack properties like the attack algorithm and threat model. The networks were trained on both attacked and benign audio samples and evaluated on various attacks in speaker classification, verification, and ASR systems. The aim was to assess their ability to identify and classify different adversarial attacks.

**Attack Classification in Speaker Recognition: -**

The researchers achieved high accuracy in classifying various attacks, except for distinguishing between L1 and L2 threat models, which caused confusion. Attacks with higher SNR were more challenging to classify. The model effectively detected adversarial attacks and identified their properties in speaker recognition systems.

**Attack Classification in Speech Recognition: -**

Initially, the system had 5% accuracy, but it improved using PLDA. Distinguishing PGD-L1 and PGD-L2 attacks was challenging. The system worked well with low SNR, but detecting speech recognition attacks in some scenarios remained difficult.

**Attack Classification in Speech Recognition: -**

The system checked if two speakers are the same using PLDA, resulting in low error rates when attacks were known. However, handling unknown attacks needs improvement, which can be achieved by training the model with more diverse attack types.

**Unknown Attack Detection: -**

In two experiments, using attacks not seen during training, the error rate for unedited + edited attacks were 37.3%, while for all edited/adversarial attacks, it was 19%. (Error rate indicates how effectively attackers were caught.)

**Conclusion: -**

X-vectors were effective in detecting attacks, achieving high accuracy with common attacks like FGSM, PGD, and Carlini-Wagner (90%). Attack signatures aided in classification. However, detecting unknown attacks showed 19% accuracy (measuring how well the system can detect adversarial attacks).

|  |  |  |  |
| --- | --- | --- | --- |
| Detect/Mitigate | Universal | Prior Knowledge | Only Defensive? |
| Detects | Yes. | Yes, during training phase. | Helps with detection and classification |

|  |  |  |  |
| --- | --- | --- | --- |
| Add Ons | Modified Network | Test on Attack Knowledge | Attack |
| Not mentioned | - | Not mentioned if attacker knows about the system. | FGSM, PGD, Carlini Wagner, Iter-FGSM. |

|  |  |  |
| --- | --- | --- |
| Architecture | Data Set | Accuracy |
| X-vector, PLDA (Probabilistic Linear Discriminant Analysis), Signature Extraction Networks. | VoxCeleb1, VoxCeleb2, Librispeech | Classification of known attacks = 90% Unknown attacks = 19% |