WINLAB | Wireless Information Network Laboratory

Adversarial Machine Learning Against Voice Assistant Systems

Celina Zhou, David Lau, Saurabh Bansal
Advisor: Dr. Yingying (Jennifer) Chen
2020 WINLAB Summer Internship, Rutgers University

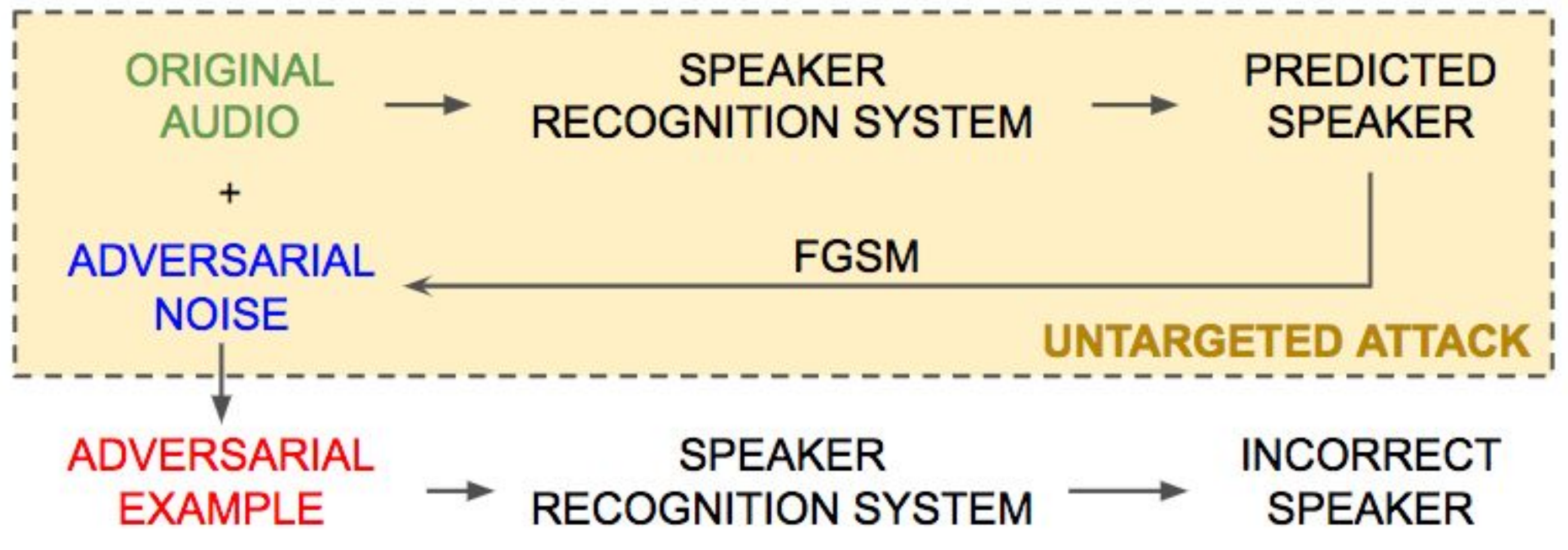
Abstract

Voice assistant systems have become increasingly more commonplace, transcending personal use and expanding into corporate environments. Speaker recognition systems, specifically, often serve as another form of biometric authentication. Our research targets the vulnerabilities of such speaker recognition systems, particularly the X-Vector model. By adding small perturbations to audio input, we perform untargeted and targeted attacks, which aim to produce misclassification of audio and imitate another user, respectively.

Background

- **Voice Assistant Systems**
  - Often used to authenticate users via voice recognition
  - Examples: Alexa, Google Home, Siri, Cortana, etc.
- **Adversarial Attacks**
  - Added perturbations to produce misclassifications
  - Attacker intentionally designs these inputs to fool the model into making a mistake
  - Shows that many modern machine learning algorithms can be broken

Untargeted Attack Overview



ORIGINAL AUDIO → SPEAKER RECOGNITION SYSTEM → PREDICTED SPEAKER

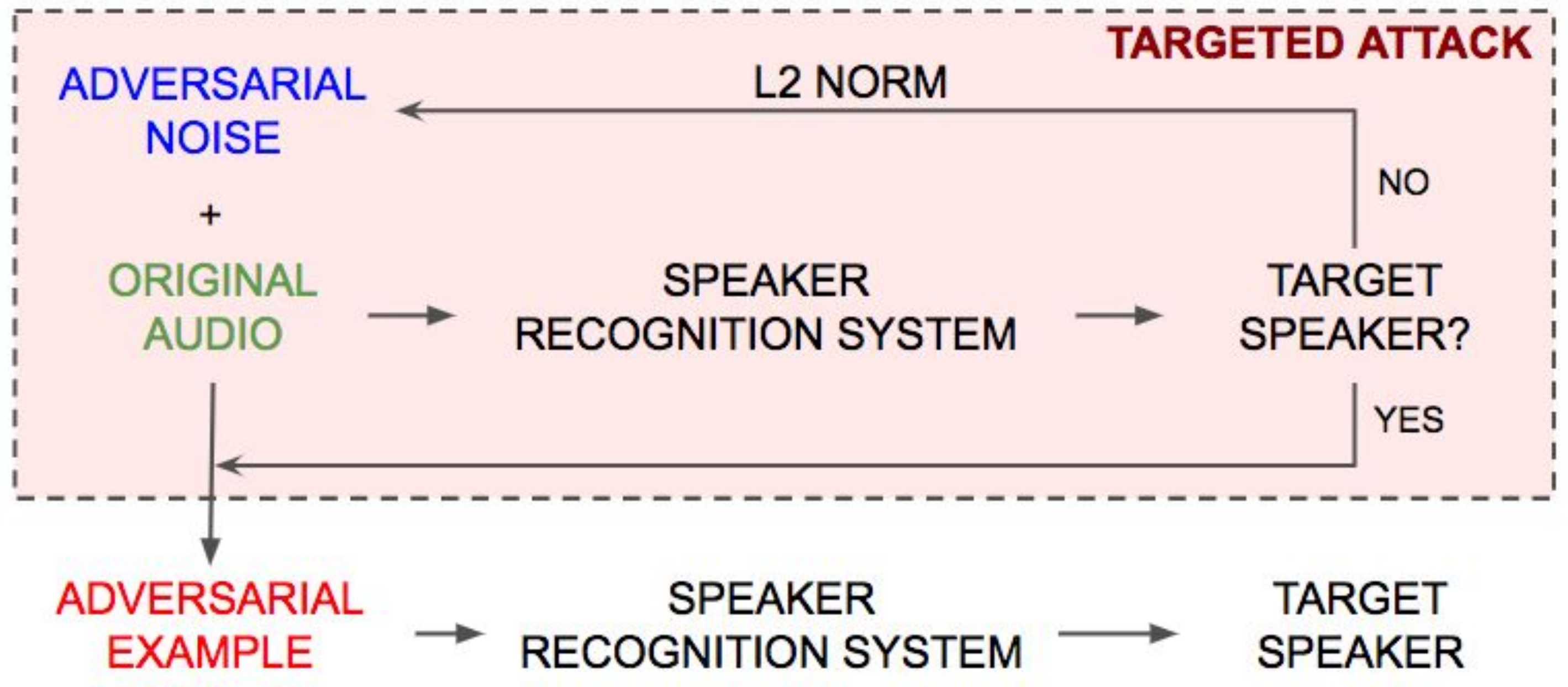
+

ADVERSARIAL NOISE ← FGSM

UNTARGETED ATTACK

ADVERSARIAL EXAMPLE → SPEAKER RECOGNITION SYSTEM → INCORRECT SPEAKER

Targeted Attack Overview



TARGETED ATTACK

ADVERSARIAL NOISE ← L2 NORM ← NO

+

ORIGINAL AUDIO → SPEAKER RECOGNITION SYSTEM → TARGET SPEAKER?

YES

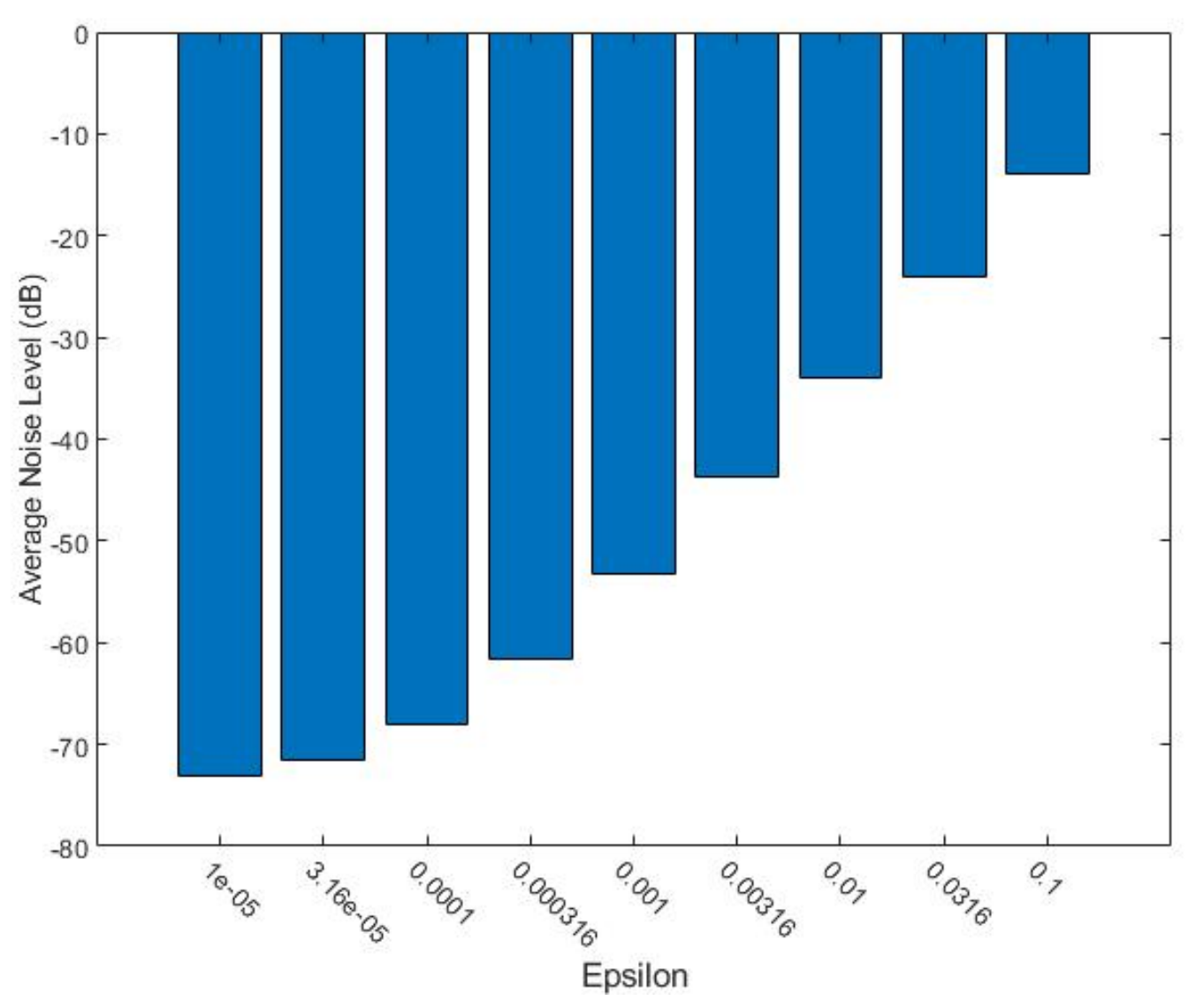
ADVERSARIAL EXAMPLE → SPEAKER RECOGNITION SYSTEM → TARGET SPEAKER

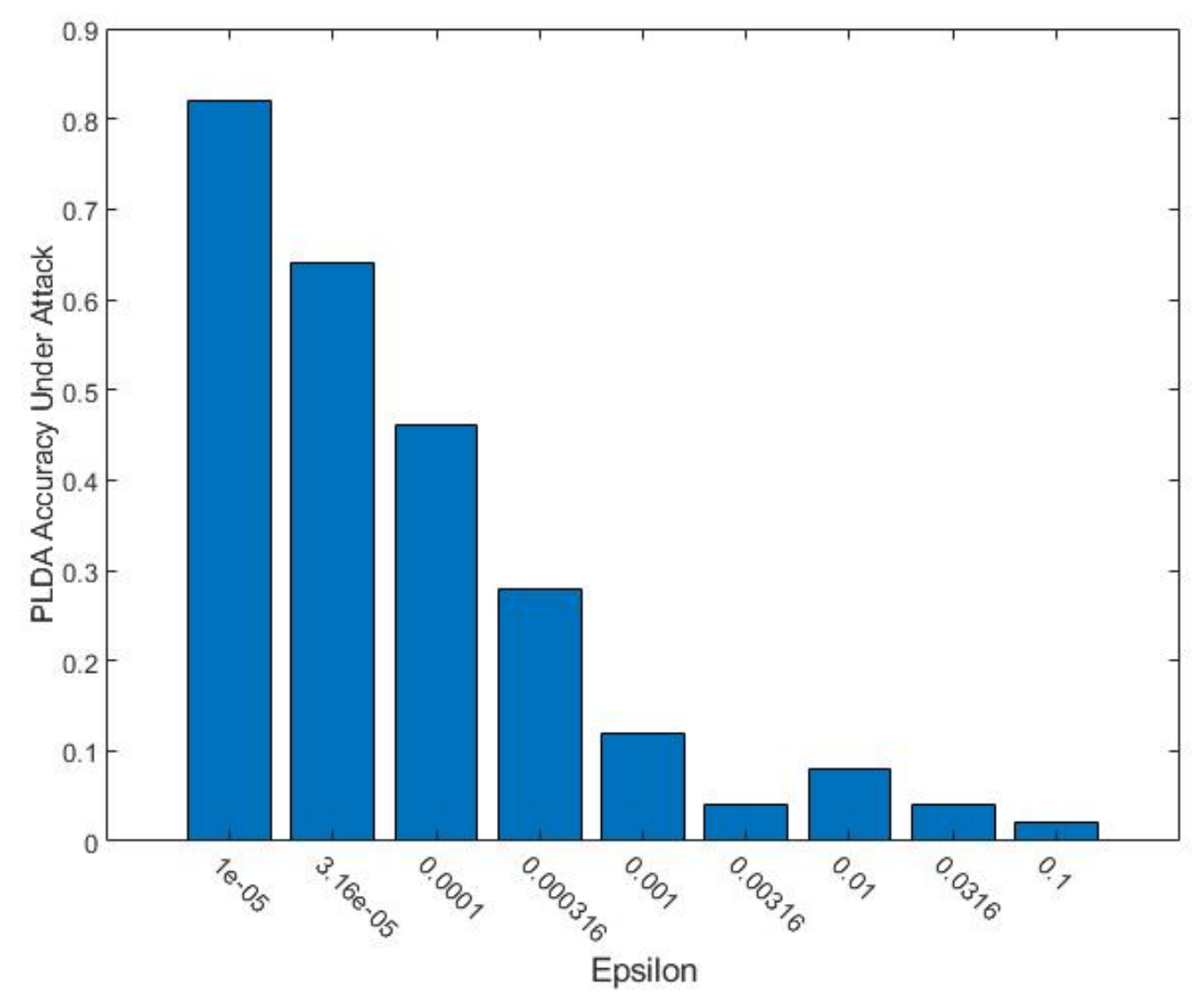
Materials and Methods

- **Implemented X-Vector model in Tensorflow**
  - Extract Mel-frequency cepstral coefficients (MFCC) from the audio to serve as lower-dimensional representation
  - Construct model using time-delay neural networks (TDNN) to capture contextual details of the audio
    - MFCC serve as input for TDNN
  - Train probabilistic linear discriminant analysis (PLDA) classifier for robust speaker recognition
- **Untargeted Attack** **Equation:** $X' = X + \in sign(\nabla x\ (-y{*}log(f(X))))$
  - Misclassify audio sample as an incorrect speaker
  - Generate voice feature embeddings from the X-Vector model
  - Take the gradient of the cross-entropy cost function
  - Add linear perturbation to original signal using fast gradient sign method (FGSM)
  - Produce samples of varying strength, scaled by a given epsilon value
- **Targeted Attack** **Equation:** (*minimize*) $-y_t{*}log(f(X+\delta))+c||\delta||_2$
  - Augment an audio sample to imitate a specific speaker
  - Iteratively modify perturbation according to the gradient descent direction where the possibility of target speaker increases
  - Minimize L2 norm of perturbation

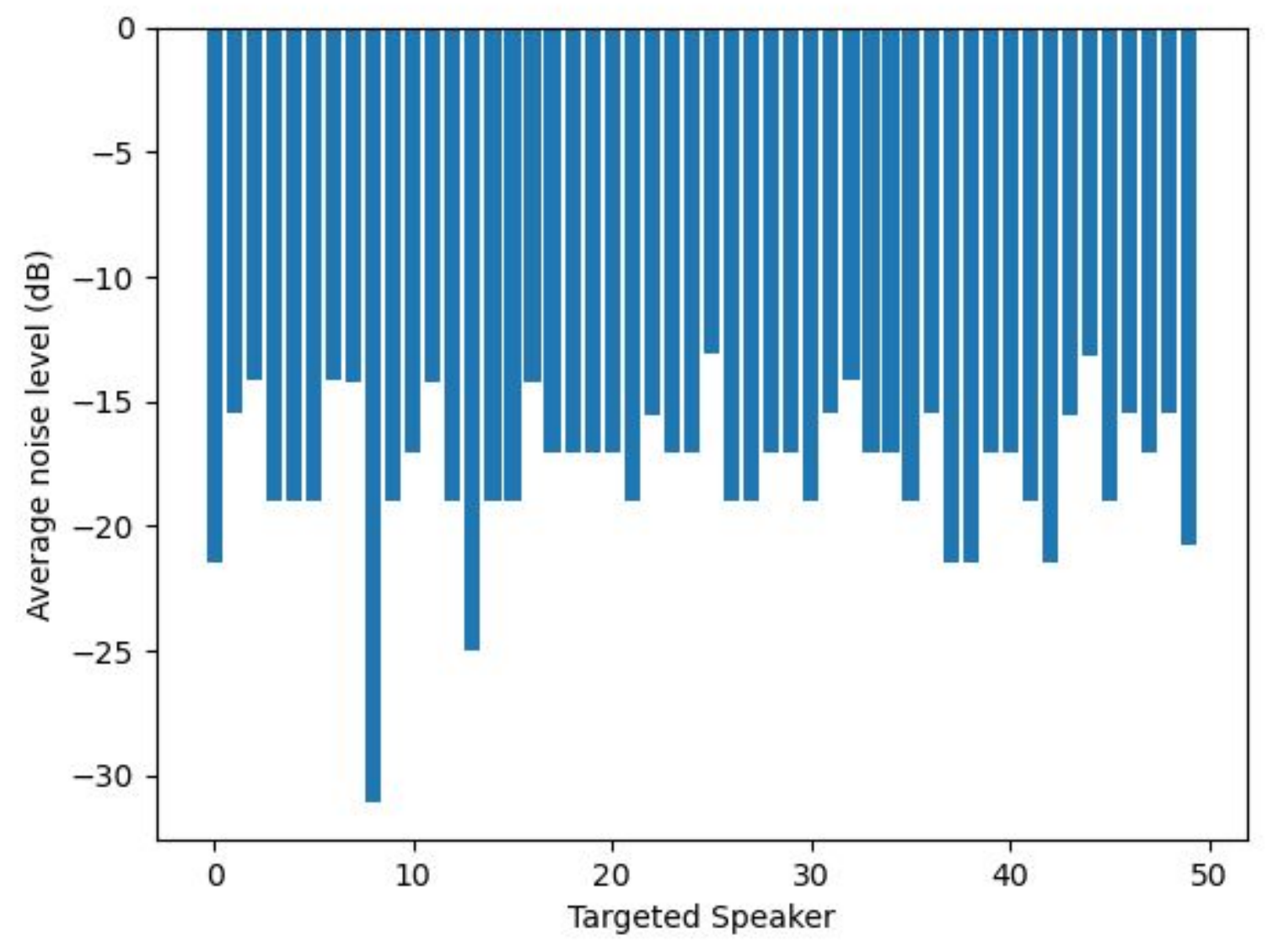
Results

Untargeted Attack results:



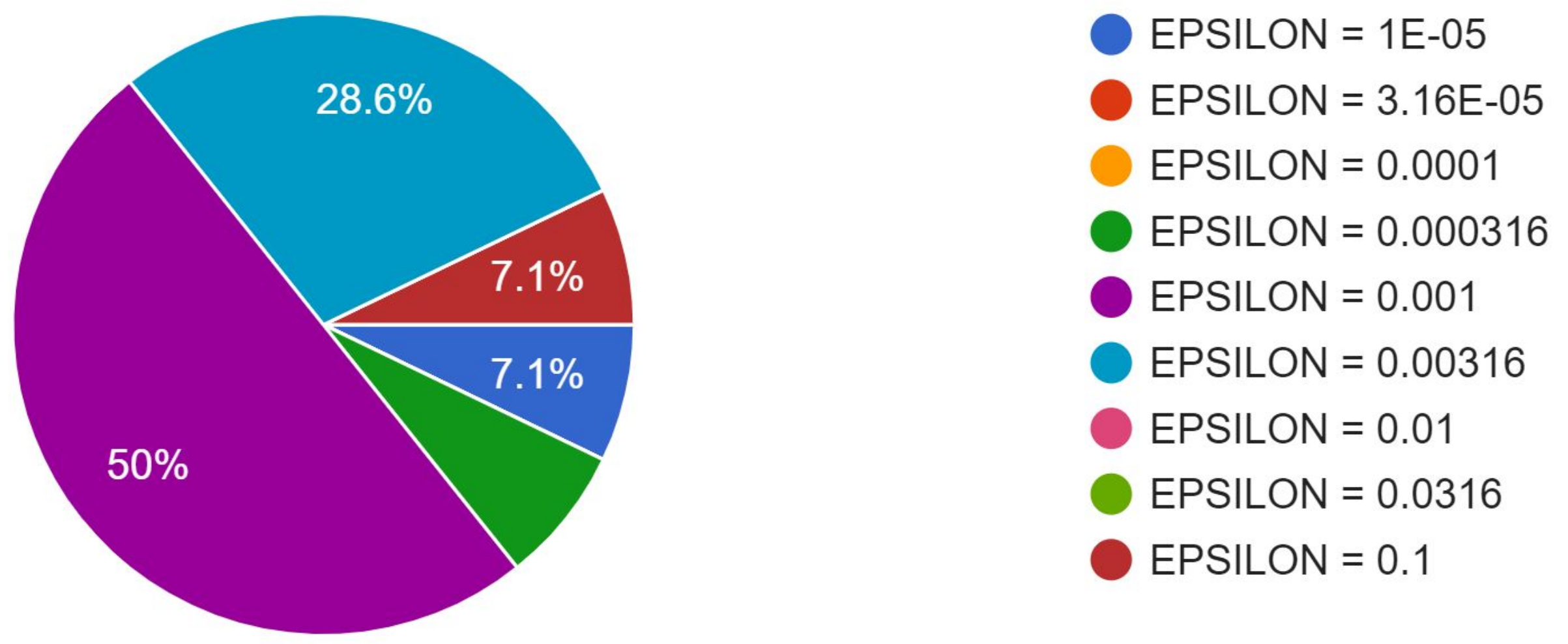


Targeted Attack results:



Discussion

Untargeted Attack: The left plot shows different epsilon values and their corresponding noise level in decibels (dB). The right plot shows different epsilon values and the corresponding PLDA accuracy under attack.



We created a survey to determine the epsilon value at which the human ear could first detect our added perturbations. The results are displayed in the pie chart above.

Targeted Attack: The bar graph depicts the average noise level in decibels (dB) for adversarial samples targeting each speaker.

Conclusions and Future Work

- **Untargeted attack:**
  - ε = 1E-5 distortion was inaudible to human ear but decreased PLDA accuracy by ~20%
  - Earliest discernable distortion occurred at ε = 0.001, with the PLDA accuracy decreased by almost 70%
- **Targeted attack:**
  - Average noise level of -21 dB, ~ ε = 0.0316 in untargeted
- **Future Work:**
  - Launch over-the-air attacks and account for the room impulse response
  - Better disguise attacks to produce greater unpredictability

References

[1] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey and S. Khudanpur, "X-Vectors: Robust DNN Embeddings for Speaker Recognition," 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, 2018, pp. 5329-5333
[2] Li, Zhuohang & Shi, Cong & Xie, Yi & Liu, Jian & Yuan, Bo & Chen, Yingying. (2020). Practical Adversarial Attacks Against Speaker Recognition Systems. 9-14. 10.1145/3376897.3377856.