Comparative Study of Acoustic Injection Masking Methods for Protecting Automatic Speech Recognition Systems

Aishwarya Trickangode Kaliyamardhanan
UFID : 75626032

Sai Kiran Kokkiligadda
UFID: 69295959

Priya Lakshmi Theagarajan
UFID : 92980120

Srikruth Reddy Puram
UFID : 70678614

1 Field and topic of study

Our study falls within the cybersecurity domain because it addresses vulnerabilities in automatic speech recognition (ASR) systems [1], which are a crucial part of many voice-controlled and speech-to-text applications.

The topic of the study is acoustic injection attacks and masking techniques to prevent or mitigate them. Acoustic injection attacks are a type of spoofing attack that uses sound waves to interfere with the sensor readings of voice-controlled devices, such as smartphones, drones, or smart speakers. By resonating the sensors at their natural frequencies, attackers can inject malicious commands or data into the devices, causing them to malfunction or execute unauthorized actions.

Masking techniques are countermeasures that aim to protect voice-controlled devices from acoustic injection attacks by either detecting or recovering from compromised sensor values. Detection techniques try to identify the presence of an attack by analyzing the sensor signals or the acoustic environment. Recovery techniques try to restore the benign sensor values by filtering out the noise or using other supplementary sensors.

2 Problem Specification

The core problem this research seeks to address is the vulnerability of automatic speech recognition (ASR) systems to acoustic injection attacks. These attacks can introduce unauthorized audio commands, potentially compromising the security and integrity of ASR systems. The specific focus of this research is to evaluate and compare the effectiveness of various acoustic masking techniques, such as spectral subtraction, temporal masking, and audio smoothing, in mitigating these attacks. The aim is to identify which technique or combination of techniques provides the most robust defense against acoustic injection attacks under varying conditions, thereby enhancing the security of ASR systems.

3 Motivation

Previous studies seem to concentrate on specific masking methodologies, leaving a void in our understanding of the broader spectrum of potential techniques. A comprehensive analysis encompassing various methods is essential for developing a holistic defense strategy. By surveying and evaluating a spectrum of defense techniques, we aim to contribute valuable insights to the broader field of securing ASR systems, ensuring their robustness in real-world applications.

In our motivation for this study, it is essential to highlight the current gap in the existing literature. Notably, our thorough review revealed a scarcity of research specifically surveying software-based defense techniques. Our research focuses on evaluating the strengths and weaknesses of the proposed defense strategies like Audio Turbulence and Audio Squeezing in CommanderSong [2] and usage of SVMs in DolphinAttack [3]. Through systematic experimentation, we aim to identify the robust aspects and potential limitations of these defenses, contributing essential insights to enhance their effectiveness in securing Automatic Speech Recognition systems.

4 Proposed solution

Our proposed solution aims to defend against the effects of the acoustic injected commands within audio samples, as highlighted in the "CommanderSong: A Systematic Approach for Practical Adversarial Voice Recognition" paper [2] By leveraging advanced masking techniques, our approach seeks to enhance the security and reliability of ASR systems by mitigating the impact of these injected commands.

4.1 Dataset

We will utilize audio samples with acoustic injected commands from the "Commander Song" paper [2] as our dataset. These adversarial audio samples have the capability of being transferable, and is effective on black box commercial

ASR systems with potential to affect millions of users. Hence, these samples are representative of real-world scenarios and provide a suitable testing ground to evaluate our proposed solution.

4.2 Methodology

We have identified three masking techniques that will be applied on the audio samples with injected commands. These techniques are selected for their effectiveness in reducing the presence of unwanted signals while preserving the integrity of the original audio content.

4.2.1 Masking techniques

Spectral Subtraction : This technique involves transforming audio signals from the time domain to the frequency domain using Fourier Transform. Subsequent steps include estimating and subtracting noise spectra from the audio spectrum, followed by inverse Fourier Transform to restore the signals to the time domain [4].

Temporal Masking: This technique identifies segments with loud sounds and reduces the volume or removes sounds within a masking threshold to mitigate the presence of injected commands [5].

Audio Smoothing (Low Pass Filtering): By selecting a cut-off frequency and applying a low pass filter, this technique attenuates high frequencies, thereby smoothing the audio signals [1].

4.2.2 Implementation details

Our proposed solution will be implemented using Python libraries SciPy, Numpy, SoundFile (Python package that provides an interface to read and write audio files in various formats), Librosa (Python library for analyzing and manipulating audio signals).

4.2.3 Evaluation Metrics and Expected results

The effectiveness of each of the masking technique will be calculated using metrics three metrics. Signal-to-Noise Ratio (SNR), Word Error Rate (WER), and Phoneme Error Rate (PER). These metrics will quantify the performance of the masking techniques in reducing the presence of injected commands within the audio samples. We anticipate that our proposed solution will significantly reduce the impact of acoustic injected commands within the audio samples, as evidenced by improvements in WER and PER metrics. These outcomes will validate the efficacy of our masking techniques in enhancing the security of audio-based systems.

5 Planned milestones and tasks to achieve

Data collection and Analysis (W1 : Feb 14 – Feb 20)

- Collect audio samples from Commander Song paper
- Analyze quality of audio
- Apply pre-processing techniques, if needed

Experiment Design (W2 : Feb 21 – Feb 27)

- Implement Spectral Subtraction technique.
- Adjust values of variables like Overlap ratio, size of the Fast Fourier Transform (FFT), length of the smoothing window etc. and save resulting audio

Data collection and Analysis (W3 : Feb 28 – Mar 5)

- Measure the effectiveness of Spectral Subtraction using metrics like SNR, WER, and PER.
- Identify the strengths and weaknesses of the Spectral Subtraction technique

Experiment Design (W4 : Mar 6 – Mar 12)

- Implement the Temporal masking technique.
- Modify the size of the masking window, adjust masking threshold curves, adaptation time etc. and save resulting audio.

Mid-term review (W5: Mar 13 – Mar 19)

- Present the experiment results of Spectral subtraction and Temporal masking

Data collection and Analysis (W6 : Mar 20 – Mar 26)

- Measure the effectiveness of Temporal masking using metrics like SNR, WER, and PER.
- Identify the strengths and weaknesses of the Temporal masking technique

Experiment Design (W7 : Mar 27 – Apr 2)

- Implement the Audio Smoothing technique.
- Modify variables such as filter order, type of filter, and transition width and save resulting audio.

Data collection and Analysis (W8 : Apr 3 – Apr 9)

- Measure the effectiveness of Audio Smoothing using metrics like SNR, WER, and PER.
- Identify the strengths and weaknesses of the Temporal masking technique

Discussion (W9 : Apr 10 – Apr 16)

- Analyze outcomes of each technique.
- Validate the findings from each experiment

Review results (W10 : Apr 17 – Apr 23)

- Create visualizations using Matplotlib, charts or tables.
- Identify the best masking technique

Final Project submission (W11 : Apr 18 – Apr 30)

- Create project paper.
- Present the findings as part of the experiments

References

[1] "Learning Mask Scalars for Improved Robust Automatic Speech Recognition," in *IEEE Conference Publication*, 2022. [Online]. Available: https://ieeexplore.ieee.org/document/10022813

[2] X. Yuan et al., "CommanderSong: A Systematic Approach for Practical Adversarial Voice Recognition," in *Proceedings of the 27th USENIX Security Symposium (USENIX Security 18)*, 2018. [Online]. Available: https://www.usenix.org/conference/usenixsecurity18/presentation/yuan-xuejing

[3] G. Zhang et al., "DolphinAttack: Inaudible Voice Commands," in *Proceedings of the ACM SIGSAC Conference on Computer and Communications Security (CCS '17)*, 2017. [Online]. Available: https://dl.acm.org/doi/abs/10.1145/3133956.3134052

[4] "Array-based Spectro-temporal Masking For Automatic Speech Recognition," *Carnegie Mellon University Repository*, 2022. [Online]. Available: https://kilthub.cmu.edu/articles/thesis/Array-based_Spectro-temporal_Masking_For_Automatic_Speech_Recognition/6714845

[5] "A perceptual masking approach for noise robust speech recognition," *EURASIP Journal on Audio, Speech, and Music Processing*, 2012. [Online]. Available: https://asmp-eurasipjournals.springeropen.com/articles/10.1186/1687-4722-2012-29