Attacking the Efficiency of Deep ASR Pipelines via Adversarial Perturbations

THE UNIVERSITY OF TEXAS AT DALLAS
Clark Summer Research Program

Yugank Mishra, Tingxi Li, Wei Yang
The Department of Computer Science, University of Texas at Dallas

UT DALLAS

Introduction

Abstract

Modern machine learning systems increasingly use pipeline architectures, which consist of multiple deep learning models, each performing a subtask. Pipelines:

- Accomplish more complex tasks than single models
- Avoid the difficulties of training end-to-end models from scratch
- Are more flexible in usage and more computationally efficient.

However, the pipeline architecture also introduces efficiency vulnerabilities, particularly when deployed on edge devices. These vulnerabilities, largely unexplored in deep learning pipelines, can be targeted using adversarial attacks.

In this work, we build an Automatic Speech Recognition (ASR) pipeline consisting of two deep neural networks: a speaker diarization model followed by a transcription model. Operating under a white-box threat model, we first identify the pipeline's efficiency vulnerabilities and computational bottlenecks, and then exploit them by increasing the amount of audio processed. Our attack results in a **174% increase** in latency and a **67% increase** in memory usage when evaluated on the AMI Corpus dataset.

Automatic Speech Recognition Pipelines

A typical Automatic Speech Recognition (ASR) pipeline comprises of a speaker diarization model, which answers the question "Who is speaking when?" and is essential for efficient speech recognition (Wan et al., 2020), a transcription model, and any other tasks.

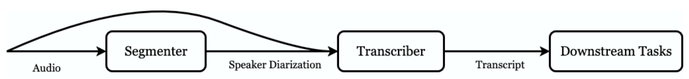


Figure (a) - Illustration of the data flow in a ASR pipeline

The segmentation functions as a filter for audio that provides additional information regarding speaker identity. There are many different model approaches to the speaker diarization problem, but most include multi-speaker predictions for efficiency. The transcription model recognizes speech and transforms it from a waveform into natural language. Additional stages can follow after the transcription, but only the first two stages were focused on in this work.

Method

Pipeline Design

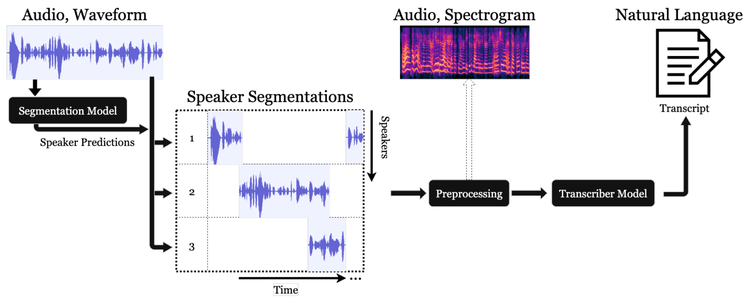


Figure (b) - Overview of our ASR Pipeline

Our representative pipeline for the ASR task is built with the state-of-the-art Pyannote Segmentation 3.0, with multi speaker labels, for the segmentation model and OpenAI's Whisper Base for the transcription model. Given a waveform of audio, the pipeline's final output is a transcript in natural language.

Attack Scheme

We operate under a white-box threat model for the attack, allowing for gradient-based attack generation.

The attack seeks to maximize audio processed by:

- Fooling the segmentation model into false overlaps
- Falsely identifying speech in silence

The W present in the loss is a weight that maximizes/minimizes the desired labels.

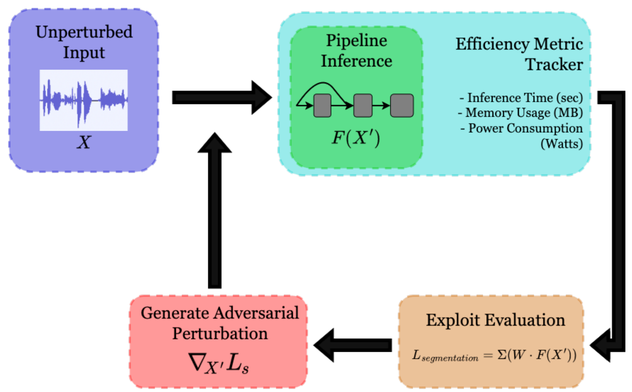


Figure (c) - The attack scheme in through which adversarial inputs are generated.

Artifacts

Artifacts from processing adversarial input are visible in both the audio and transcription. Transcription artifacts are noted via duplicated transcriptions (due to overlap label misidentification).
Figure (d) shows examples of artifacts in audio, visualized through a spectrographic analysis.

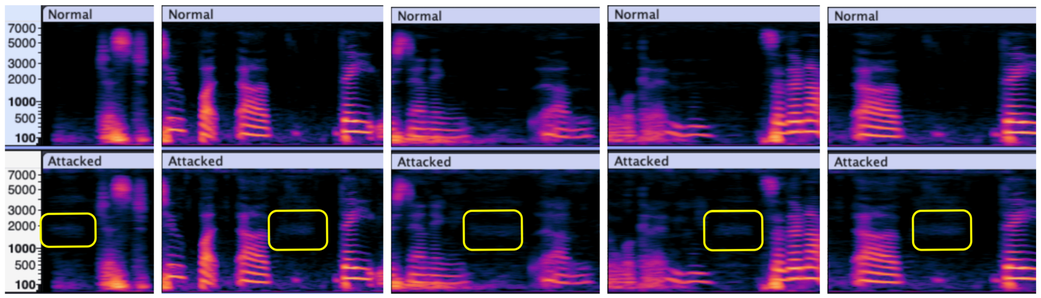


Figure (d) - Examples of Speech Mimicry in Adversarial Inputs

Results

To measure the effectiveness of the attack, adversarial perturbations were generated for 7 hours of AMI corpus data using the attack scheme described in Figure (c). Then, GPU usage statistics were tracked for both the original and adversarial audios. Of the 3 tracked metrics, inference time had the clearest trend with an **174% increase**, followed by Memory Usage at an **67% increase**. A minimal 10% increase was observed for power consumption**.**

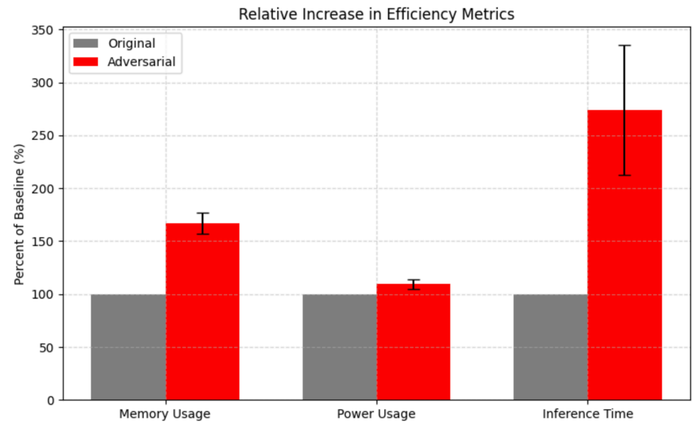


Figure (e) - Results of Adversarial Attacks. Error bars represent a 95% Confidence Interval for the relative increase.

Through this work, it has been shown that ASR Pipelines using segmentations with overlapping speaker labels are quite susceptible to efficiency adversarial attacks. The method proposed can target both the non-speech and the multi-speaker labels created in segmentation, so pipelines using such segmentations are vulnerable.

Contact & Acknowledgements

Contact

Team Lead: Yugank Mishra, Yugank.Mishra@UTDallas.edu
Team Member: Tingxi Li, tingxi.Li@UTDallas.edu

Acknowledgements

I would like to thank Tingxi Li for his constant guidance and encouragement throughout this project. I would also like to thank the Clark Program for providing this valuable first research opportunity.

References

D. Wan, et al, "Subtitles to Segmentation: Improving Low-Resource Speech-to-Text Translation Pipelines" in *Cross-Lanauage Search and Summarization of Text and Speech*, 2020, pp. 68-73.