RuHuman: A Resilient Multimodal AI-Powered Audio Verification System

Timothy Do

Tuesday, November 21st, 2023

ECE202A Embedded Systems @ UCLA

Motivation and Objectives

- Audio deep fakes are becoming more easily spoofable with the advent of generative AI.
- RuHuman evaluates existing audio liveness detection systems that exploit audio multimodalities.
- By the project deadline, this project will:
  - Evaluate different audio liveness detector architectures against state of the art voice cloners.
  - Develop a user interface to make gateway devices (e.g. phone) easily run with their microphones

Technical Approach and Novelty

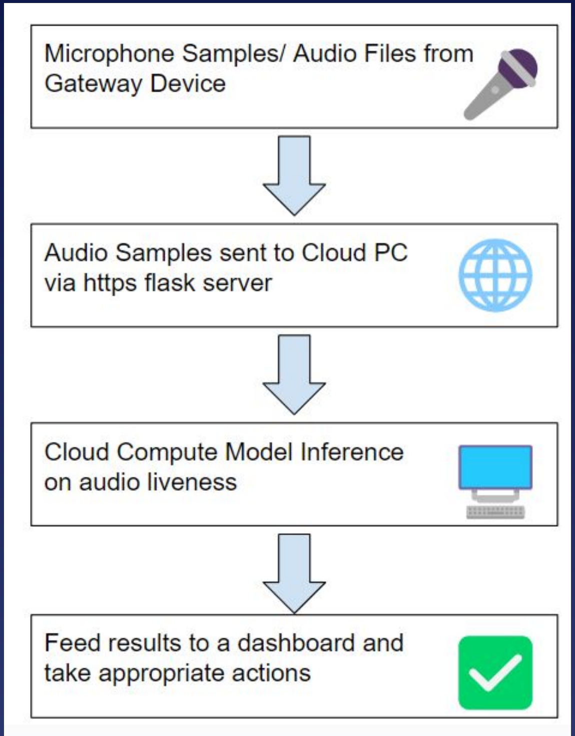
- ASVSpoof is a dedicated challenge that to develop audio spoofing classifiers. (RuHuman focuses on LA).
- Researchers take advantage of various encodings (STFT, MFCC, CQCC) to augment the training resolution of the audio sample.
- Datasets used for some submissions were in a controlled environment, insusceptible to ‘out in the wild’ samples.
- RuHuman considers additive sources (e.g. noise, multi-speaker) when evaluating the different model architectures in addition to fine tuning these models (if time permits)

ASVspoof

Automatic Speaker Verification and Spoofing Countermeasures Challenge

Methods

- Baseline Audio Encoding Algorithms:
  - MFCC: Mel-Frequency Cepstral Coefficients (in the human audible spectrum).
  - CQCC: Constant Q Cepstral Coefficients (spectro/temporal resolution in low/high frequencies).
  - Spectrogram: Heatmap plot of Frequency over time (computed through a filter bank).
- Previous submissions of ASVSpoof are implemented in MATLAB/Python.
- UI will be on Python Flask Web Server.
- The main dataset we will be the one provided by the ASVSpoof 2019 competition for LA (Logical Access)
  - 107 Distinct Speakers (46 M, 61 F).



Microphone Samples/ Audio Files from Gateway Device

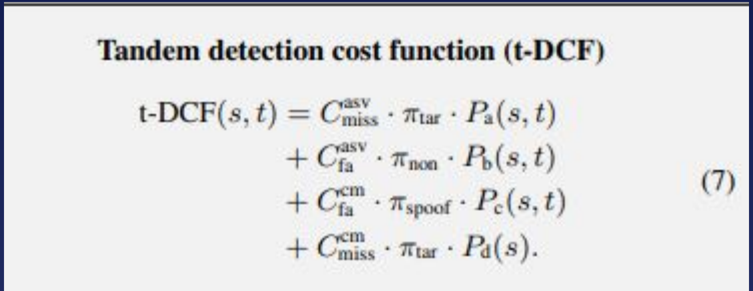
Audio Samples sent to Cloud PC via https flask server

Cloud Compute Model Inference on audio liveness

Feed results to a dashboard and take appropriate actions

Evaluation and Metrics

- t-DCF: A custom cost function developed by the ASVSpoof team that weighs between ASV and CM (Countermeasure) metrics
- EER: The rate at which the miss rate and the false alarm rate are equal each other.
- Computation Time: Liveness detection is inferred in a short time frame(i.e. >10s) so that actions can be done in real time.
- Objective is to minimize all metrics for accuracy and efficiency in live settings.



Tandem detection cost function (t-DCF)

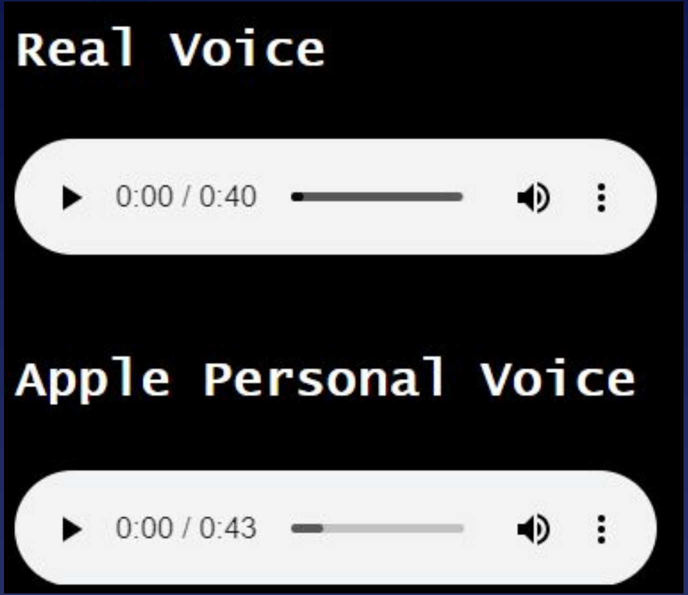
$$\begin{aligned}\text{t-DCF}(s,t) = {} & C_{\text{miss}}^{\text{asv}} \cdot \pi_{\text{tar}} \cdot P_{\text{a}}(s,t) \\ & + C_{\text{fa}}^{\text{asv}} \cdot \pi_{\text{non}} \cdot P_{\text{b}}(s,t) \\ & + C_{\text{fa}}^{\text{cm}} \cdot \pi_{\text{spoof}} \cdot P_{\text{c}}(s,t) \\ & + C_{\text{miss}}^{\text{cm}} \cdot \pi_{\text{tar}} \cdot P_{\text{d}}(s).\end{aligned} \qquad (7)$$

Current Status and Next Steps

- Compared State of the Art Voice Cloners with Real Voice: https://timothydo.me/RuHuman/progress.html
- Explored code pipeline for ASVSpoof 2021 baseline systems & ASVSpoof2019 from NESL.

Next Steps:

- Develop own custom evaluation set with additive sound sources
- Evaluate various ASVSpoof Submissions with evaluation set
- Develop a user interface where a device upload audios to determine liveness detection.
- Fine-tune ASVSpoof detection models for more advanced spoofing attacks (if time permits).



Real Voice

Apple Personal Voice

Thanks for Listening!
Any Questions?