The "2019 Automatic Speaker Verification Spoofing And Countermeasures Challenge" (ASVspoof) competition aimed to facilitate the design of highly accurate voice spoofing attack detection systems. The competition did not emphasize model complexity and latency requirements; such constraints are strict and integral in real-world deployment. Hence, most of the top performing solutions from the competition used an ensemble approach, and combined multiple deep learning models to maximize detection accuracy – this kind of approach would sit uneasily with real-world deployment constraints.

To design a lightweight system, we combined the notions of skip connection (from ResNet) and max feature map (from Light CNN), and evaluated the accuracy of the system using the ASVspoof 2019 dataset. With an optimized constant Q transform (CQT) feature, our single model achieved a replay attack detection equal error rate (EER) of 0.37% on the evaluation set, surpassing the top ensemble system from the competition that achieved an EER of 0.39%.

**OBJECTIVES**

Why Voice Spoofing Detection?

Voice Spoofing Detection, Physical Access (PA)

Voice Spoofing Detection, Logical Access (LA)

Test to Speech (TTS): synthetic speech
Voice Conversion (VC): converted voice

**METHODS**

Spoofing Classification
- Classifying Human or Speaker $f: x \rightarrow y (0,1)$
- Dimension for $x$ is $(n_{freq}, n_{time})$
- Data: $(i, x) \in R(n_{freq}, y \in (0, 1))$
- Learning objective: $L(f, x, y)$
- Objective is to minimize $C$, w.r.t $\theta$:
  $C = \sum L(f(x), y)$

What would be a good model ($f(x)$)?

Automatic Speaker Verification Spoofing And Countermeasures Challenge (ASVspoof 2015, 2017 and 2019)

Deep-learning based methods and ensemble solutions are dominating in voice liveness detection challenge

Non-speech part have information?

How to develop well performing light-weighted model?

The longer you listen the better the performance

**RESULTS**

High performance of ResMax

Non-speech part have information?

How to develop well performing light-weighted model?

The longer you listen the better the performance

**CONCLUSIONS**

Existing voice spoofing attack detection solutions have been designed without considering real-world model complexity and detection latency requirements, and often consist of multiple heavy and complex deep learning models. Such solutions would not be considered suitable given the tight model size and latency requirements. In comparison, our CQT-ResMax model used only a single deep learning model with far fewer model parameters to outperform the top performing PA solution from the competition, achieving an EER of 0.37% compared to the current best competition EER of 0.39%, which is an ensemble solution. As for the LA set, we rank third with an EER of 2.19%, just behind the second best ensemble solution that achieved an EER of 1.86% in the evaluation set. Among the single model systems, although CQT-ResMax used the least number of parameters, it demonstrated significant superiority in detection accuracy.

**ACKNOWLEDGMENT**

This work was conducted at Samsung Research. The authors would like to thank Samsung Research Security Team for the helpful discussions. IK was supported by the National Research Foundation of Korea (NRF) grant funded by Ministry of Science and ICT (2020R1C1A6013020).

**CONTACTS**

For any questions, feel free to ask me via ikwak2@cau.ac.kr