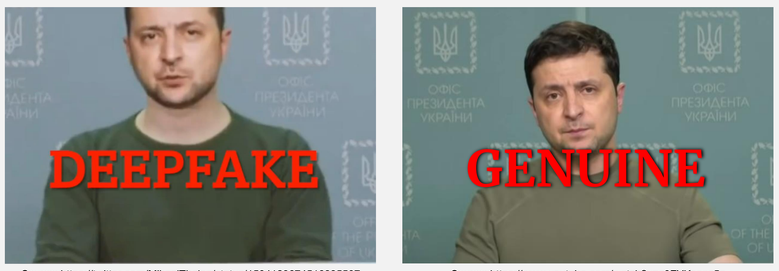
Can DeepFake voices steal high-profile identities?

Bence Mark Halpern[1,2], Finnian Kelly[3]
[1]University of Amsterdam, Netherlands Cancer Institute (b.m.halpern@uva.nl)
[2]Technical University of Delft
[3]Oxford Wave Research (finnian@oxfordwaveresearch.com)

Introduction



Source: https://twitter.com/MikaelThalen/status/1504123674516885507

Source: https://www.youtube.com/watch?v=c0FYKsqzq5g

RQ: Can we apply an LR framework to detect "in-the-wild" DeepFakes of high-profile identities?

- DeepFakes are becoming more convincing every day
- The recent case of a Zelenskyy DeepFake highlights possible malicious use of this technology
- There is a growing need for tools to reliably detect malicious use of DeepFakes, aka spoofing

Dataset

- 30 audio DeepFakes of high-profile celebrities collected from online sources
- The DeepFakes were likely created using a Tacotron-2 model, which can synthesise high-quality speech using 3 hours of training data
- For each of the 30 DeepFakes, a corresponding genuine recording was also sourced

Ask to listen some samples!

Spoof detector

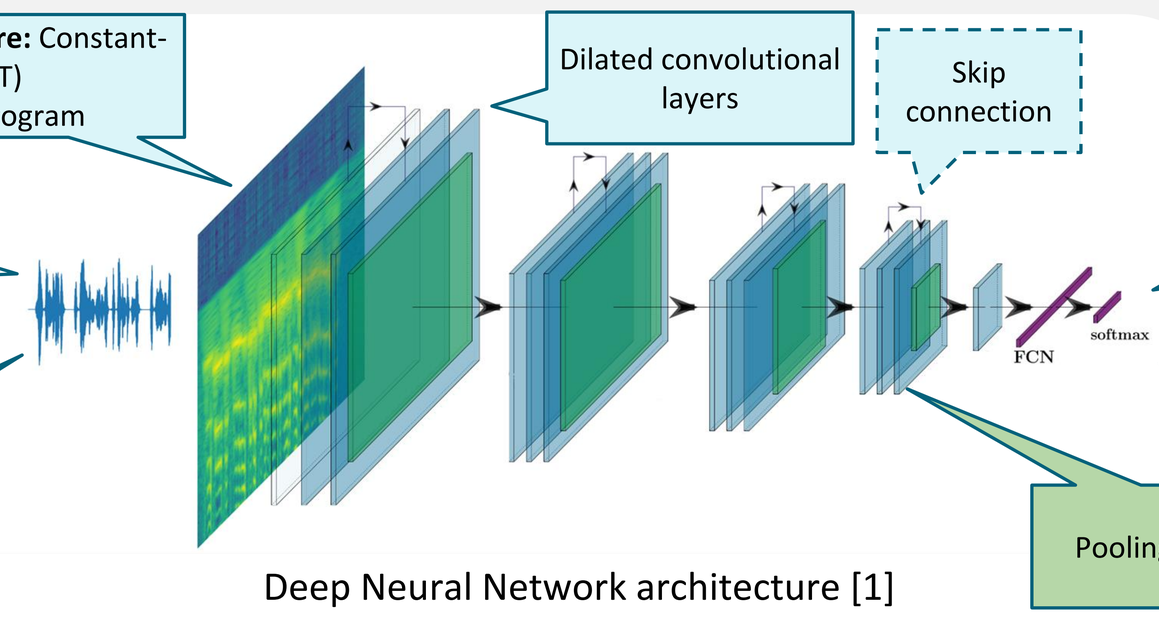
Input: 6-second speech ***chunks*** sampled at 16 kHz, loudness normalised, with (●), and without (✕) voice activity detection (VAD) [3]

Noise level: estimated per-***chunk*** using WADA SNR [2]

Feature: Constant-Q (CQT) spectrogram

Dilated convolutional layers

Skip connection



Output: Softmax output scores for each ***chunk*** are averaged to obtain a detection score for the recordings

Pooling layers

Fully connected layer

Deep Neural Network architecture [1]

Likelihood ratio (LR) framework

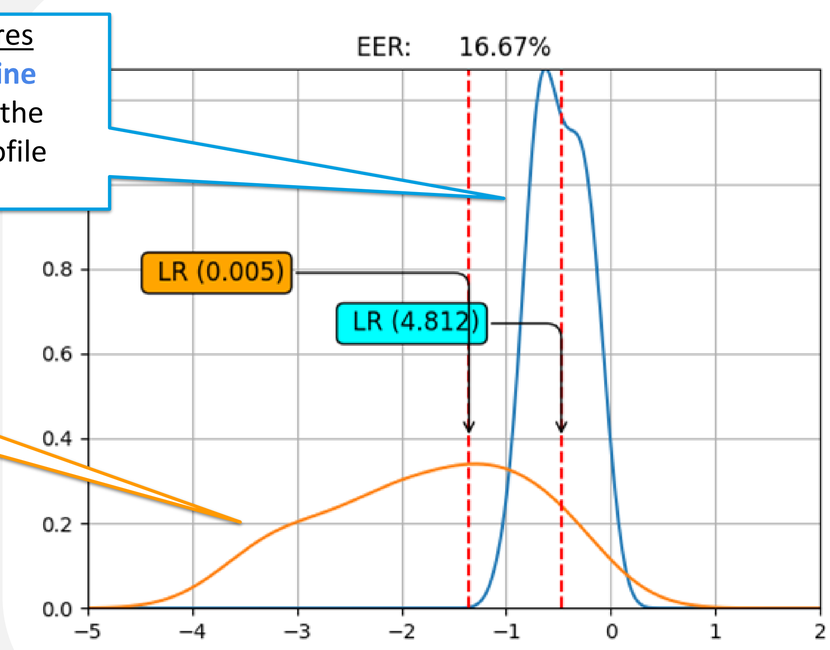
- We propose to apply a likelihood ratio framework to spoofed speech detection
- Kernel density estimates are obtained from the detection scores of **30 genuine (H0)** and **30 spoof recordings (H1)**
- We calculate a genuine/spoof LR for the Zelenskyy recordings given **H0** and **H1**

$$LR = \frac{p(H_0|x)}{p(H_1|x)}$$

- A **Genuine recording** of Zelenskyy produces and LR>1 and the **Zelensky DeepFake** produces an LR<1
- **RQ:** The LRs provide correct support in both DeepFake and genuine cases, demonstrating that this approach can be successfully applied to "in-the-wild" audio

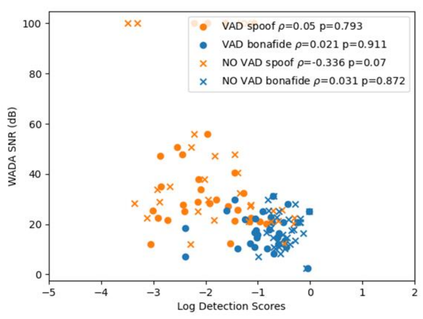
detection scores from **30 genuine recordings** of the same high-profile celebrities

detection scores from **30 spoof recordings** of high-profile celebrities



Log Detection scores

Post-hoc noise analysis



- The effect of noise is investigated using a correlation analysis between the WADA SNR and the detection scores; no strong correlations are observed
- The detector is therefore robust to noise, but qualitative analysis indicates that reverb influences the detection scores
- VAD does not improve the equal error rate (EER), consistent with [4, 5]
- Silence is important in spoof detection [5] but leads to less noise-robust detectors
- Future experiments should focus on reverb and data augmentation

References

[1] Halpern, B. M., Kelly, F., van Son, R., & Alexander, A. (2020). Residual networks for resisting noise: analysis of an embeddings-based spoofing countermeasure.
[2] Kim, C., & Stern, R. M. (2008). Robust signal-to-noise ratio estimation based on waveform amplitude distribution analysis. In Ninth Annual Conference of the International Speech Communication Association.
[3] Bredin, H., Yin, R., Coria, J. M., Gelly, G., Korshunov, P., Lavechin, M., ... & Gill, M. P. (2020, May). Pyannote. audio: neural building blocks for speaker diarization. In ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 7124-7128). IEEE.
[4] Lai, C. I., Chen, N., Villalba, J., & Dehak, N. (2019). ASSERT: Anti-spoofing with squeeze-excitation and residual networks. arXiv preprint arXiv:1904.01120.
[5] Müller, N. M., Dieckmann, F., Czempin, P., Canals, R., Böttinger, K., & Williams, J. (2021). Speech is silver, silence is golden: What do ASVspoof-trained models really learn?. arXiv preprint arXiv:2106.12914.