Deepfake Audio Detection: Model Analysis

Table of Contents

1. Introduction

This report evaluates three **state-of-the-art deepfake audio detection models—RawNet2 + Sinc Filters, AASIST, and Wav2Vec2 + LCNN + Bi-LSTM**—by analyzing their **key innovations, performance metrics, strengths, and limitations**.

2. Deepfake Audio Detection Models

2.1 RawNet2 + Sinc Filters (AASIST)

Key Technical Innovations

- Uses **SincNet filters** instead of standard convolution layers for **more interpretable frequency feature extraction**.

- **RawNet2 CNN-based architecture** processes raw audio waveforms directly, eliminating the need for handcrafted features.

- **End-to-end learning** enables direct deepfake detection from raw waveforms.

Reported Performance Metrics

- **Equal Error Rate (EER): 0.033** (Logical Access dataset).

Why This Approach Is Promising?

- **End-to-end processing:** Works directly on raw waveforms without requiring handcrafted features.
- **Highly optimized for spoofing detection** with deep **CNN-based processing**.
- **Proven success** in **Logical Access deepfake detection** with a very low **EER**.

Potential Limitations or Challenges

- **Computationally expensive:** Deep CNN models require **high processing power**.
- **May not generalize well** to **unseen attack types** without retraining.

2.2 AASIST (Audio Anti-Spoofing with SincNet)

Key Technical Innovations

- Uses **SincNet-based convolution filters** to extract **high-quality frequency representations**.

- **Lightweight CNN-based model** designed for **efficient spoof detection.**

- Optimized for **robust generalization across datasets**.

Reported Performance Metrics

- **Equal Error Rate (EER): 0.034** (Logical Access dataset).

Why This Approach Is Promising?

- **Lightweight and efficient:** Can work in **real-time**.

- **Performs well across different datasets**, making it more **robust** to new attack types.
- **Easier to train and deploy** compared to deeper CNN architectures.

Potential Limitations or Challenges

- Might require **dataset fine-tuning** for **optimal performance** in unseen deepfake attacks.
- Performance is **slightly lower than RawNet2**, but still highly competitive.

2.3 Wav2Vec2 + LCNN + Bi-LSTM (Prosody & Phoneme-based Approach)

Key Technical Innovations

- **Combines three advanced components:**
    - **Wav2Vec2** → Self-supervised learning on raw waveforms (**captures deep audio patterns**).
    - **LCNN (Lightweight CNN)** → Detects **frequency distortions** in deepfake audio.
    - **Bi-LSTM (Bidirectional LSTM)** → Captures **prosodic and phoneme-level variations** in speech.
- **Unique Approach:** Uses **pronunciation-based detection**, focusing on **prosody and phonemes** to differentiate real vs. fake voices.

Reported Performance Metrics

- **Equal Error Rate (EER): 1.58** (Logical Access dataset).

Why This Approach Is Promising?

- **Captures high-level speech patterns:** Works beyond **waveform-level analysis**.
- Uses **Wav2Vec2**, which has strong **self-supervised learning capabilities**.
- **Good generalization ability:** Designed for **cross-dataset deepfake detection**.

Potential Limitations or Challenges

- **Computationally expensive:** Wav2Vec2 requires **high GPU power**.

- **No public implementation available**, so requires a **custom setup**.

3. Conclusion

This report evaluated three advanced deepfake audio detection models:

- **RawNet2 + Sinc Filters**: Offers **end-to-end processing and high accuracy** but is **computationally expensive**.
- **AASIST (SincNet-based CNN)**: **Lightweight and robust** for real-time applications but **requires fine-tuning** for unseen deepfakes.
- **Wav2Vec2 + LCNN + Bi-LSTM**: **Combines self-supervised learning and phoneme-based detection** but **needs a high computational budget** and lacks public implementations.