RawNet2 Audio Deepfake Detection

This repository presents a lightweight implementation of RawNet2, a deep learning model for detecting audio deepfakes using raw waveform data. It demonstrates a simplified pipeline with dummy data and lays the groundwork for future enhancements using real-world datasets and architectures.

Research & Model Selection

After reviewing deepfake detection techniques, we shortlisted three viable approaches:

1. RawNet2 (ICASSP 2020)

- End-to-end deep neural network that operates directly on raw waveforms
- Combines residual Conv1D blocks with GRU layers for temporal modeling
- Achieved **EER of ~1.08%** on ASVspoof 2019 dataset
- Avoids handcrafted features for better generalization
- **Limitations:**
    - Requires high GPU memory for training
    - Sensitive to noise in real-world environments

2. LFCC/CQCC + GMM

- Uses cepstral coefficients (LFCC or CQCC) as handcrafted features
- Gaussian Mixture Models (GMM) used for classification
- Lightweight and suitable for real-time/embedded systems
- **EER ~17.55%** with LFCC-GMM baseline
- **Limitations:**
    - Poor generalization to unseen spoofing methods
    - Not as accurate as deep models

3. AASIST (INTERSPEECH 2022)

- Multi-stream architecture using:
    - 2D CNNs (Res2Net)
    - 1D temporal attention
    - Spectro-temporal features
- Achieved **state-of-the-art EER ~0.63%** on ASVspoof 2019
- Robust in noisy, real-world conditions
- **Limitations:**
    - Computationally intensive
    - Requires optimization for deployment

Implementation Process

Challenges Faced

- Lack of real audio deepfake data
- GRU training instability with synthetic data
- Model overfitting due to small dummy dataset

How We Addressed Them

- Simulated a binary classification task using alternating labels
- Reduced model size (RawNet2Lite) to improve training stability
- Assumed:
    - Audio clips were 1-second @ 16kHz
    - Balanced dataset (bonafide vs. spoof)
    - Focused on demonstrating feasibility, not real-world performance

Analysis

Why RawNet2

- Processes raw waveforms directly (no feature engineering required)
- Performs well in public benchmarks like ASVspoof
- GRU effectively captures long-term speech patterns

Model Workflow

- **Input:** 1D raw audio waveform
- **Conv1D layers:** Extract low-level acoustic features
- **GRU layer:** Captures temporal dependencies
- **FC + Sigmoid:** Outputs binary classification (bonafide/spoof)

Performance (Dummy Dataset)

- Accuracy: ~60–70% (depending on randomness)
- Confirms the functional pipeline, not real-world readiness

Strengths

- Modular and extendable architecture
- Ideal for experimenting with raw audio-based deepfake detection

Weaknesses

- Dummy data lacks variability and realism
- GRU struggles with noise or longer sequences without attention

Suggestions for Improvement

- Use real datasets (e.g., ASVspoof 2019 LA, WaveFake, TIMIT)
- Add noise robustness and domain adaptation techniques
- Explore LSTM + attention or Transformer-based architectures
- Integrate pretrained RawNet2 models
- Extend pipeline with:
    - Confidence scoring
    - Post-processing
    - Voice activity detection (VAD)

Reflection

Key Challenges

- Limited access to real labeled audio
- Simplifying RawNet2 without losing core benefits

Real-World Considerations

- Real data introduces compression artifacts, accents, and background noise
- Clean training datasets may not generalize well without augmentation

Additional Data/Resources Needed

- Labeled spoofed speech from various synthesis methods
- Augmented data covering diverse speaking styles and noise levels
- More compute to support deeper RawNet2 versions

Production Deployment Plan

- **Model Serving:** TorchScript or ONNX for optimized inference
- **API Integration:** Use FastAPI or Flask for audio uploads and predictions
- **Preprocessing:** Normalize input audio and apply VAD
- **Monitoring:** Log confidence scores and route low-confidence predictions to human reviewers