Momenta - Audio Deepfake Detection Take-Home Assessment Submission

Part 1: Research & Selection

Approach 1: SafeEar (Acoustic Features + Transformer)

Key Technical Innovation:

- Splits audio into “what’s said” (semantic) and “how it’s said” (acoustic—like tone and rhythm). Uses only acoustic stuff with a Transformer model to spot fakes, keeping content private.

Reported Performance Metrics:

- Equal Error Rate (EER) of 2.02% on ASVspoof2019—super low, meaning it’s really good at telling real from fake.

Why It’s Promising:

- Privacy is huge for real conversations—no one wants their words leaked. It’s ace at detecting AI speech (tested on ASVspoof), and Transformers are smart enough to catch tricky fakes. Could be fast with some tweaks.

Limitations/Challenges:

- Needs precomputed features (like Hubert), which takes setup time. Real-time might need a lighter version—Transformers can be heavy

Approach 2: RawNet2 (End-to-End Raw Waveform Model)

- **Key Technical Innovation**:
    - Uses raw audio waveforms directly , feeding them into a deep convolutional network with residual blocks and GRU layers to spot fake patterns. It’s end-to-end, learning everything from scratch.
- **Reported Performance Metrics**:
    - EER of 4.31% on ASVspoof2019 LA dataset.
- **Why It’s Promising**:
    - Raw audio skips preprocessing, making it fast and adaptable for real-time—perfect for AI speech detection. Its deep learning

power catches subtle fakes in convos, and your task loves cutting-edge stuff.

- **Potential Limitations**:
  - Heavy computation—needs GPU power, so real-time might need a lighter version. Small datasets might overfit without tweaks.

.

Approach 3: AASIST (Anti-Spoofing with Integrated Spectro-Temporal Graph Attention)

- **Key Technical Innovation**:
  - Combines spectrograms with a graph attention network to model time and frequency together. It's like a brain linking sound patterns across a clip to catch fakes.
- **Reported Performance Metrics**:
  - EER of 1.77% on ASVspoof2019 LA (from "AASIST: Audio Anti-Spoofing using Integrated Spectro-Temporal Graph Attention" linked in the repo's research orbit).
- **Why It's Promising**:
  - Top-tier accuracy—nails AI-generated speech. The spectro-temporal focus is ace for convo analysis, and its efficiency hints at real-time potential with optimization. Momenta's gonna love this precision!
- **Potential Limitations**:
  - Complex setup—graph networks aren't simple to run live without streamlining. Might overfit on small datasets unless you scale it down.

Part 2: Implementation

- **Model**: "Deepfake-audio-detection" (wav2vec2-base), pre-trained and fine-tuned.
- **Dataset**: Hemg/Deepfake-Audio-Dataset (100 samples, split 90 train/10 test).
- **Fine-Tuning**: 10 epochs in Colab with GPU, best at Epoch 2 (train loss: 0.300600, val loss: 0.307129, accuracy: 90%).

- **Code Highlights**: Used transformers and datasets, added noise/time-shifting augmentation, resampled to 16 kHz, trained with Trainer.

Comparison with Selected Approaches:

SafeEar (Acoustic Features + Transformer):

Overview: Splits audio into semantic (content) and acoustic (tone, rhythm) parts, uses only acoustic features with a Transformer for detection. EER: 2.02% on ASVspoof2019.

Technical Differences:

- **Input**: SafeEar uses precomputed Hubert L9 features (acoustic-only), my model takes raw 16 kHz audio, processed internally by wav2vec2-base.
- **Architecture**: Both use Transformers, but SafeEar's codec splits data first—mine is end-to-end, no privacy split.
- **Preprocessing**: I used resample and augment (noise, shifting); SafeEar relies on external feature extraction, more complex.

RawNet2 (End-to-End Raw Waveform Model):

Overview: Takes raw audio straight into a CNN with residual blocks and GRU layers—no manual features. EER: 4.31% on ASVspoof2019.

Technical Differences:

- **Input**: Both use raw audio (yours at 16 kHz), but RawNet2 skips feature extraction, while wav2vec2-base has a CNN layer first.
- **Architecture**: RawNet2's CNN + GRU vs. my chosen model is CNN + Transformer—GRU's lighter, Transformer's deeper.
- **Preprocessing**: I added augmentation (noise, shifting) and resampling; RawNet2 takes audio as-is.

AASIST (Spectro-Temporal Graph Attention):

Overview: Uses spectrograms with a graph attention network to link time and frequency patterns. EER: 1.77% on ASVspoof2019.

Technical Differences:

- **Input**: AASIST uses spectrograms, wav2vec2-base does too (internally), but starts with raw audio.
- **Architecture**: AASIST's graph attention vs. wav2vec2-base is CNN + Transformer—graph's unique for time-frequency links, Transformer's broader context.
- **Preprocessing**: I resample and augment; AASIST needs spectrogram generation (via Librosa), so more manual.