**Aatithaya Paliwal**

**Part 2: Implementation Plan for Audio Deepfake Detection**

**1. Selected Approach for Implementation**

I will implement a **CNN-based approach using Mel-spectrograms** for audio deepfake detection. This approach has been widely used in speech and audio classification tasks and is effective for distinguishing real vs. spoofed speech.

**Comparison with Other Approaches**

| **Approach** | **Pros** | **Cons** |
| --- | --- | --- |
| **Raw Waveform-Based CNNs** | No need for handcrafted features | Requires large data and computing power |
| **Spectrogram-Based CNNs (Selected)** | Well-established, interpretable, robust against noise | Needs preprocessing (Mel-spectrogram computation) |
| **RNN/Transformers (LSTM/GRU/Wav2Vec2.0)** | Captures temporal dependencies well | Requires more training time and data |

I chose the **Spectrogram-Based CNN** because it balances efficiency and accuracy while working well with limited data.

**2. Using Existing Code and Implementation**

I will adapt existing **PyTorch-based CNN architectures** used for **audio classification**. The following steps will be implemented:

1. **Load dataset:** Read ASVspoof 2017 V2 audio files and labels.
2. **Preprocess audio:** Convert .wav files into **Mel-spectrograms**.
3. **Define CNN model:** Use a ResNet or a custom CNN for classification.
4. **Train and fine-tune:** Use the ASVspoof dataset for model fine-tuning.
5. **Evaluate:** Compute accuracy, Equal Error Rate (EER), and F1-score.

**3. Dataset Selection**

* **Primary dataset:** **ASVspoof 2017 Version 2**
* **Alternative dataset options:**
  + **ASVspoof 2021**: [Zenodo Link](https://zenodo.org/records/14498691)
  + **Datasets from Media-Sec Lab Repo**: [GitHub](https://github.com/media-sec-lab/Audio-Deepfake-Detection)

ASVspoof 2017 is selected because it provides well-structured **real and spoofed** speech samples with detailed metadata.

**4. Fine-Tuning and Model Training**

I will perform light fine-tuning on the dataset:

* Use **pretrained models** (e.g., ResNet18) initialized with weights from **ImageNet** for feature extraction.
* Train the model on **Mel-spectrogram representations** of the ASVspoof dataset.
* Fine-tune on **ASVspoof 2017 V2 train/dev set**, validating on the dev set.
* Use **AdamW optimizer** with **learning rate scheduling** for stable training.

**Expected Outcome**

* A trained CNN model that can classify genuine vs. spoofed speech.
* Comparison with baseline models (other approaches).
* Insights on model performance and feature extraction effectiveness.