**Audio Deepfake Detection - Momenta Assessment**

This document outlines the research, implementation, and analysis audio deepfake detection system, marking up the requirements.

**Project Overview:**

The goal of this project is to implement a robust audio deepfake detection system capable of identifying Human and AI-generated human speech, with a focus on real-time detection and real conversations. This project follows the instructions provided below.

**Part 1: Research & Selection**

For the task of detecting AI-generated human speech, particularly in real-time or near real-time scenarios involving real conversations, I've thoroughly reviewed the provided GitHub repository. After careful consideration, I've identified three promising approaches: RawNet2, LFCC-GMM, and Res Net-based Spectrogram Analysis.

**1. RawNet2: Direct Raw Waveform Analysis for Fine-Grained Spoofing Detection**

* **Key Technical Innovation:**
  + RawNet2 deviates from traditional approaches by directly processing raw audio waveforms. This eliminates the need for handcrafted features like MFCCs or LFCCs, which can potentially discard subtle but crucial information.
  + The architecture relies on a deep convolutional neural network (CNN) with Squeeze-Excitation (SE) modules. SE modules enable the network to learn channel-wise attention, effectively emphasizing informative frequency bands and suppressing irrelevant ones. This is critical for capturing minute discrepancies introduced by spoofing algorithms.
  + The use of Leaky ReLU activation functions is a key point, that help with the vanishing gradient problem, and help the model to learn more complex patterns.
* **Reported Performance Metrics:**
  + RawNet2 has demonstrated state-of-the-art performance on the ASVspoof 2019 Logical Access (LA) dataset, which is a standard benchmark for audio spoofing detection. This indicates its ability to effectively distinguish between genuine and spoofed speech.
* **Promise for Our Specific Needs (AI-Generated Human Speech, Real-Time, Real Conversations):**
  + **High Accuracy:** The high performance on ASVspoof datasets suggests strong potential for detecting AI-generated speech.
  + **Direct Raw Waveform Input:** The ability to process raw audio waveforms is particularly beneficial for analyzing real conversations, which often involve complex acoustic environments and variations. No need for pre-processing.
  + **Potential for Real-Time:** While the deep architecture can be computationally intensive, CNNs are generally well-suited for optimization and hardware acceleration. Techniques like model quantization and pruning can significantly reduce the computational cost.
* **Potential Limitations/Challenges:**
  + **Computational Complexity:** Real-time processing on resource-constrained devices may require significant optimization efforts.
  + **Generalization:** Real-world conditions can introduce significant variability, and the model's performance may depend heavily on the diversity of the training data.
  + **Data Hunger:** Neural networks require a lot of data to train.

**2. LFCC-GMM: Statistical Modelling for Resource-Efficient Detection**

* **Key Technical Innovation:**
  + This approach uses Linear Frequency Cepstral Coefficients (LFCCs), a traditional audio feature, to represent the spectral characteristics of speech.
  + Gaussian Mixture Models (GMMs) are then employed to model the statistical distribution of LFCCs for genuine and spoofed speech. By comparing the likelihood of an input utterance belonging to each distribution, the system can classify it as genuine or spoofed.
* **Reported Performance Metrics:**
  + LFCC-GMM has proven effective in scenarios with limited computational resources, offering a balance between performance and efficiency.
* **Promise for Our Specific Needs (AI-Generated Human Speech, Real-Time, Real Conversations):**
  + **Low Computational Cost:** The simplicity of LFCC extraction and GMM modeling makes this approach suitable for near real-time detection on devices with limited processing power.
  + **Robustness to Variations:** Statistical Modeling can be robust to variations in real conversations, as it focuses on the overall distribution of features rather than specific patterns.
* **Potential Limitations/Challenges:**
  + **Limited Feature Representation:** LFCCs may not capture the complex, subtle artifacts introduced by advanced spoofing techniques as effectively as deep learning models.
  + **Performance Dependence:** The performance of the system heavily depends on the quality of LFCC extraction and GMM training.

**3. Res Net-based Spectrogram Analysis: Leveraging Image Classification Techniques for Audio Spoofing**

* **Key Technical Innovation:**
  + This approach converts audio waveforms into spectrograms, which are visual representations of the audio's frequency content over time.
  + Res Net architectures, originally designed for image classification, are then used to learn discriminative features from these spectrograms. The idea is that spoofing artifacts manifest as subtle visual patterns in the spectrograms.
* **Reported Performance Metrics:**
  + Res Net-based spectrogram analysis has demonstrated strong performance on various audio spoofing datasets, indicating its ability to learn effective features.
* **Promise for Our Specific Needs (AI-Generated Human Speech, Real-Time, Real Conversations):**
  + **Powerful Feature Learning:** Res Net architectures are known for their ability to learn complex and hierarchical features, which can be beneficial for capturing subtle spoofing artifacts.
  + **Visual Representation:** Spectrograms provide a visually intuitive representation of audio, which can aid in understanding the model's decision-making process.
* **Potential Limitations/Challenges:**
  + **Spectrogram Computation Overhead:** Computing spectrograms add a pre-processing step, which can impact real-time performance.
  + **Parameter Tuning:** Spectrogram parameters (e.g., window size, hop length) need to be carefully tuned for optimal performance.
  + **Data Size:** Spectrograms increase the data size, which will increase the training time and inference time.

**Part 2: Implementation**

For this assessment, I chose to implement **RawNet2** due to its state-of-the-art performance and potential for real-time optimization. The code is based on the provided GitHub repository (<https://github.com/eurecom-asp/rawnet2-antispoofing.git>).

**Code Snippets:** Code is available on the provided Git hub repo.

**Git Hub Repo Link:** https://github.com/tarun02-git/Audio-Spoofing-Detection.git

**Implementation Details:**

* **Dataset:** ASV spoof 2019 LA dataset.
* **Code:** Python using Py Torch and TrainingLight fine-tuning was performed to adapt the model to the chosen dataset.

**Part 3: Documentation & Analysis**

**Implementation Process:**

* The RawNet2 model architecture was directly adapted from the GitHub repository.
* The ASV spoof 2019 LA dataset was used, and appropriate data loaders were created.
* The model was trained using the Adam optimizer and binary cross-entropy loss.
* Challenges encountered:
  + File loading errors due to corrupted audio files.
  + Solutions: Implemented error handling in the dataset loader to skip corrupted files.
* Assumptions:
  + The dataset is representative of real-world spoofing attacks.

**Analysis:**

* **Model Selection:** RawNet2 was chosen due to its ability to directly process raw audio, its state-of-the-art performance, and its potential for real-time optimization.
* **Model Functionality:** RawNet2 uses deep CNNs with SE modules to learn discriminative features from raw audio waveforms. The SE modules allow the network to focus on important frequency bands.
* **Performance Results:** The model achieved a reasonable AUC on the development dataset, demonstrating its ability to detect audio deepfakes. (Results will vary based on training).
* **Observed Strengths:**
  + Effective in capturing subtle spoofing artifacts.
  + Direct raw audio processing.
* **Observed Weaknesses:**
  + Computational complexity.
  + Dataset Dependency.
* **Future Improvements:**
  + Model quantization and adaptation for real-time deployment.
  + Data augmentation to improve generalization.
  + Exploring advance learning from larger audio datasets.