**Part 1: Research & Selection - Audio Deepfake Detection Approaches**

**Overview**

The GitHub repository [Audio-Deepfake-Detection](https://github.com/media-sec-lab/Audio-Deepfake-Detection) by media-sec-lab offers a curated collection of research papers and resources on audio deepfake detection. For Momenta’s use case—detecting AI-generated human speech, enabling real-time or near real-time detection, and analyzing real conversations—I’ve selected three promising forgery detection models: RawNet2, AASIST, and LCNN. These models were chosen from the repository’s references for their strong performance, computational efficiency, and applicability to real-world audio. Each summary below details the key technical innovation, reported performance metrics, reasons for promise, and potential limitations, including a focus on data augmentation as a critical factor.

**1. RawNet2: End-to-End Anti-Spoofing with Raw Waveform Processing**

* **Source**: "End-to-End Anti-Spoofing with RawNet2" (ICASSP 2021)
* **Key Technical Innovation**:
  + Processes raw audio waveforms using a SincNet-based filter bank, followed by residual CNN blocks and Gated Recurrent Units (GRUs) for temporal modeling.
  + Instead of first converting audio into a spectrogram or MFCCs, RawNet2 takes the  raw waveform as input. This allows it to capture very subtle patterns and artifacts thatmight be lost during traditional pre-processing.
  + Incorporates filter-wise feature map scaling (FMS) as an attention mechanism to highlight subtle synthetic speech artifacts.
  + End-to-end design eliminates traditional feature extraction, enabling direct learning from raw data.
* **Reported Performance Metrics**:
  + Equal Error Rate (EER): 2.19% on ASVspoof 2019 LA dataset.
  + Minimum t-DCF: 0.0594, reflecting robust spoofing detection performance.
* **Why Promising for Required Needs**:
  + **AI-Generated Speech Detection**: Captures low-level waveform artifacts (e.g., vocoder distortions) unique to AI-generated audio.
  + **Real-Time Potential**: Minimal preprocessing and lightweight architecture support near real-time inference with optimization.
  + **Real Conversations**: Temporal modeling via GRUs suits continuous speech, with proven robustness on ASVspoof datasets.
* **Potential Limitations/Challenges**:
  + Struggles with noisy or compressed audio without fine-tuning.
  + **Data Augmentation Dependency**: As noted in "Channel and Compression Augmentation for Synthetic Speech Detection," channel (e.g., stereo variations) and compression (e.g., MP3 artifacts) augmentations enhance robustness. Without these, performance may degrade in real-world conversational settings with variable audio conditions.

**2. AASIST: Spectro-Temporal Graph Attention for Spoofing Detection**

* **Source**: "AASIST: Audio Anti-Spoofing Using Integrated Spectro-Temporal Graph Attention Networks" (ICASSP 2022)
* **Key Technical Innovation**:
  + Combines SincNet raw waveform processing with a spectro-temporal graph attention network (GAT) to model time-frequency relationships.
  + Raw audio is processed to produce a high-level feature map using RawNet2.
  + Two separate pooling operations isolate the strongest frequency (spectral) and temporal signals, forming two graphs.
  + Graph attention is applied to these individual graphs to “highlight” important feature
  + The two graphs are fused into one heterogeneous graph that models cross-domain relationships.
  + Uses heterogeneous stacking graph attention layers (HS-GAL) to fuse multi-domain features, improving detection of complex spoofing patterns.
  + End-to-end system designed for generalization across diverse AI-generated audio attacks.
* **Reported Performance Metrics**:
  + EER: 0.83% on ASVspoof 2019 LA dataset, outperforming many baselines.
  + Minimum t-DCF: 0.0275, indicating top-tier spoofing countermeasure performance.
* **Why Promising for Required Needs**:
  + **AI-Generated Speech Detection**: Spectro-temporal focus excels at identifying intricate AI synthesis artifacts (e.g., phase inconsistencies).
  + **Real-Time Potential**: Efficiency can be optimized for near real-time use with model pruning or hardware acceleration.
  + **Real Conversations**: Graph attention adapts to variable-length audio, enhancing robustness for real-world dialogues.
* **Potential Limitations/Challenges**:
  + High computational complexity from GAT layers may require optimization for real-time deployment.
  + **Data Augmentation Dependency**: "Channel and Compression Augmentation for Synthetic Speech Detection" highlights that channel and compression augmentations improve generalization. Without proper implementation, AASIST may underperform on compressed or multi-channel conversational audio.

**3. LCNN: Light Convolutional Neural Network for Spoofing Detection**

* **Source**: Referenced as a baseline in ASVspoof challenges (repository listing)
* **Key Technical Innovation**:
  + Employs a lightweight CNN with handcrafted features (e.g., Linear Frequency Cepstral Coefficients, LFCCs) to detect spectral spoofing artifacts.
  + Uses max-feature-map (MFM) activation to reduce parameters and boost generalization.
  + Focuses on frequency-domain analysis for efficient synthetic speech detection.
* **Reported Performance Metrics**:
  + EER: ~2–3% on ASVspoof 2019 LA dataset (varies by implementation).
  + Low latency validated across spoofing detection benchmarks.
* **Why Promising for Momenta’s Needs**:
  + **AI-Generated Speech Detection**: Targets spectral anomalies (e.g., unnatural frequency patterns) in AI-generated audio.
  + **Real-Time Potential**: Extremely lightweight design enables real-time deployment, even on edge devices.
  + **Real Conversations**: Robust on ASVspoof datasets, adaptable to real audio with tuned features.
* **Potential Limitations/Challenges**:
  + Handcrafted features may miss raw waveform cues captured by end-to-end models.
  + Less flexible, requiring careful feature selection for diverse conditions.
  + **Data Augmentation Dependency**: Per "Channel and Compression Augmentation for Synthetic Speech Detection," channel and compression augmentations bolster performance under real-world distortions. Without these, LCNN may falter on multi-channel or compressed conversational audio.

**Thought Process Behind Selection**

From the repository’s diverse approaches—spanning traditional feature-based methods to advanced end-to-end models—I selected RawNet2, AASIST, and LCNN for their alignment with Momenta’s goals:

1. **Detection Accuracy**: All achieve low EERs (<3%) on ASVspoof 2019, with AASIST leading at 0.83%, ensuring reliable AI-generated speech detection.
2. **Real-Time Feasibility**: LCNN’s simplicity guarantees real-time capability, while RawNet2 and AASIST offer near real-time potential with optimization.
3. **Conversational Robustness**: RawNet2 and AASIST’s temporal modelling suits continuous speech, and LCNN’s spectral focus adapts with proper tuning.

The inclusion of data augmentation as a limitation stems from "Channel and Compression Augmentation for Synthetic Speech Detection," which shows these techniques are vital for handling real-world audio variations (e.g., stereo channels, MP3 compression). This dependency underscores a practical challenge for deployment in Momenta’s conversational context.

* **RawNet2**: A practical end-to-end baseline with raw waveform processing for rapid prototyping.
* **AASIST**: A cutting-edge model with superior generalization for complex scenarios.
* **LCNN**: A lightweight, scalable option for immediate real-time use.