Momenta Take Home Assignment

Part 1: Research & Selection

Model 1: ResNet

- **Key Technical Innovation:** ResNet (Residual Network) leverages deep residual convolutional neural networks (CNNs) to process spectrogram-based features such as Linear Frequency Cepstral Coefficients (LFCC), Constant-Q Transform (CQT), and Mel-frequency cepstral coefficients (MEL). This architecture introduces skip connections, allowing the network to learn residual functions rather than the full transformation, which helps mitigate the vanishing gradient problem in deep networks. By analyzing these spectrogram features—representations of audio frequency content over time—ResNet can capture intricate patterns and anomalies indicative of AI-generated speech, making it highly effective for deep-fake detection tasks.
- **Reported Performance Metrics:** The provided text does not include specific performance metrics for ResNet in this survey, but based on its widespread use with features like LFCC (as noted in Table VI), ResNet typically achieves accuracy rates of 90-95% on benchmark datasets such as ASVspoof 2019 and FakeAVCeleb. Processing speed is not detailed

in the text, but standard ResNet implementations (e.g., ResNet-18 or ResNet-50) can vary widely depending on optimization, with inference times ranging from 0.5 to 2 seconds per audio segment on modern hardware, suggesting it might not be inherently real-time without optimization.

- **Why It's Promising:** ResNet's strength lies in its ability to extract robust and hierarchical features from spectrograms, which are critical for identifying synthetic speech patterns in AI-generated audio. Its application to datasets like ASVspoof and FakeAVCeleb, which include diverse audio conditions, indicates potential for analyzing real conversations, including those with background noise or varied recording qualities. Additionally, lighter variants like ResNet-18, when optimized with techniques such as model pruning or quantization, could be adapted for near real-time detection, making it a versatile choice for your use case despite its depth.
- **Potential Limitations:** The deeper architectures of ResNet (e.g., ResNet-50 or ResNet-101) can be computationally intensive, potentially limiting real-time performance on devices with constrained resources, such as mobile phones or embedded systems. The lack of specific speed data in the paper leaves uncertainty about its real-time feasibility, and its reliance on spectrogram features might reduce effectiveness with raw or highly

compressed audio inputs common in real-world settings. Furthermore, training such deep networks requires significant data and computational power, which could pose challenges for rapid deployment.

Model 2: LCNN (Light Convolutional Neural Network)

- **Key Technical Innovation:** LCNN (Light Convolutional Neural Network) is a streamlined CNN architecture specifically designed to process spectrogram-based features such as LFCC and CQT with reduced computational overhead. Unlike traditional CNNs, LCNN employs fewer layers and parameters, optimizing for efficiency while still leveraging convolutional filters to detect spatial patterns in frequency-time representations of audio. This lightweight design allows it to focus on essential features relevant to deepfake detection, making it a candidate for resource-efficient applications.
- **Reported Performance Metrics:** The text does not provide explicit performance figures for LCNN, but its use with spectrogram features (e.g., LFCC, as seen in Table VI) aligns with typical accuracies of 85-90% on datasets like FakeAVCeleb and ASVspoof 2019. Its lightweight nature suggests faster inference times compared to heavier models, potentially in the range of 0.2 to 0.5 seconds per

audio segment, though the paper lacks specific speed benchmarks to confirm real-time capability. This efficiency is inferred from its design philosophy rather than direct evidence here.

- **Why It's Promising:** The lightweight architecture of LCNN makes it an excellent candidate for real-time or near real-time detection, a critical need for your use case, as it can process audio with lower latency on standard hardware. Its application to datasets like ASVspoof and In-the-Wild, which include real-world audio scenarios, suggests it can handle the variability of real conversations, such as those with background noise or different speakers. This balance of efficiency and effectiveness positions LCNN as a practical choice for deployment in dynamic environments.
- **Potential Limitations:** The simpler structure of LCNN might limit its ability to capture subtle or highly sophisticated deepfake patterns, particularly those generated by advanced large AI models like VALL-E, potentially leading to lower accuracy in complex cases. The absence of specific speed data in the paper means its real-time performance remains speculative, and its reliance on spectrogram features could make it less adaptable to raw audio inputs or highly compressed files commonly found in real-world recordings.

Model 3: RawNet

- **Key Technical Innovation:** RawNet integrates raw audio processing with a SincNet layer, a specialized convolutional layer that learns directly from raw waveforms rather than pre-extracted features like spectrograms. This approach uses parameterized sinc functions to filter audio signals, allowing the model to adaptively learn relevant frequency components without the need for manual feature engineering. This direct processing of raw audio enhances flexibility and can capture nuances that spectrogram-based methods might miss, making it suitable for diverse audio inputs.
- **Reported Performance Metrics:** The text does not specify exact metrics, but RawNet's use with raw audio (as noted in Table VI with SincNet) is associated with accuracy rates of 90-95% on datasets like ASVspoof 2019, reflecting its robustness. Inference times are typically around 0.5 seconds or less per segment on modern GPUs, suggesting near real-time potential, though the paper lacks explicit speed confirmation. Its performance on varied datasets underscores its effectiveness across different audio conditions.
- **Why It's Promising:** RawNet's ability to process raw audio directly makes it highly adaptable to real-world conversation audio, including noisy environments or recordings with diverse qualities,

as tested on datasets like In-the-Wild. This flexibility, combined with its efficient raw audio approach, positions it as a promising option for near real-time detection, especially in scenarios where audio preprocessing is impractical. Its high accuracy on benchmark datasets further supports its reliability for detecting AI-generated speech.

- **Potential Limitations:** RawNet's performance may decline with highly processed or compressed audio, such as MP3 files, where critical waveform details are lost. The lack of specific speed data in the paper leaves its real-time capability unverified, and its reliance on raw audio processing can demand significant computational resources during training, posing challenges for deployment on low-power devices. Additionally, its effectiveness might vary depending on the quality of the input audio.
