Part 1: Research & Selection

Overview:

Selecting three different detection methods most suitable for:

1. AI detection of human speech
2. Real or near real-time assistance
3. Real conversation analysis

Method 1: Voice Spoofing Countermeasure to Detect Logical Access Attacks

Paper: [IEEE Document 9638512]

Key Technical Innovation:

- Extract handcrafted features such as **LFCC (Linear Frequency Cepstral Coefficients)**.
- Process features thereof with a **DBiLSTM (Deep Bidirectional Long Short-Term Memory)** network to identify temporal dependencies.

Performance Metrics reported:

- **EER (Equal Error Rate):** ≈ 0.74%
- **t-DCF (tandem Detection Cost Function):** ≈ 0.008

Why It's Promising:

- **High Accuracy:** Very high accuracy in authentic speech vs. speech synthesis distinction.
- **Efficiency:** Seamless hand-crafted feature extraction, thus suitable for use in real time.
- **Robustness:** It has been tested under spoofing-controlled conditions and is therefore an ideal candidate for structured conversational analysis.

Possible Limitations/Challenges:

- **Noise Sensitivity:** Would likely require further tuning in order to be effective in multihomogeneous noisy real-world scenarios.
- **Channel Variability:** Can be tuned when transitioning to other conversation conditions.

Method 2: End-to-End Anti-Spoofing on RawNet2

Paper: [IEEE Document 9414234]

Key Technical Contribution:

- Operates raw audio end-to-end directly with **Sinc filter** front-end and **RawNet2** architecture.
- Eschews hand-designed feature extraction by directly learning feature representations from waveform data.

Published Performance Metrics:

- **EER (Equal Error Rate):** ≈ 1.12%
- **t-DCF (tandem Detection Cost Function):** ≈ 0.033

Why It's Valuable:

- **Efficient Pipeline:** Streamlines processing, which is deployable in real-time.
- **Direct Feature Learning:** Retains subtle AI-synthesized speech artifacts through end-to-end learning.
- **Flexibility:** Simple to port to various recording environments without hand-tuning.

Potential Limitations/Challenges:

- **Implementation Difficulty:** High computational cost and careful hyperparameter tuning.
- **Data Dependency:** Dies when applied to other real conversational speech audio.

Method 3: End-to-End Dual-Branch Network Towards Synthetic Speech Detection

Paper: [IEEE Document 10082951]

Technical Innovation:

- Operates on a dual-branch multi-feature representation based on **LFCC** and **CQT (Constant-Q Transform)** with **HFCC**.
- Applies a multi-task learning approach to the learning of complementary spectral and temporal cues simultaneously.

Performance Measures as quoted:

- **EER (Equal Error Rate):** ≈ 0.80%
- **t-DCF (tandem Detection Cost Function):** ≈ 0.021

Why It's Promising:

- **Robust Feature Fusion:** Wider range of deepfake signals learned, required for advanced conversational audio.
- **Enhanced Robustness:** Multi-tasking allows for improved generalizability to various audio environments.
- **Balanced Performance:** Offers a balanced trade-off in detection performance vs. real-time processing capability with some extra optimization fine-tuning.

Potential Limitations/Challenges:

- **Model Complexity:** Higher complexity will affect inference rate if not optimized well.
- **Implementation Overhead:** Requires proper tuning and combination of both branches to provide maximum synergy.

Reasoning Summary:

Detection Capability:
All the solutions can identify subtle differences between human and AI language using different approaches—ranging from hand-coded feature extraction to aggregating multiple features and end-to-end learning.

Real-Time Feasibility:

- Hand-coded approaches are sparse.
- End-to-end models reduce pre-processing overhead.
- Multi-task models will have to be further optimized for real-time usage.
- Multi-task and end-to-end approaches are well placed to analyze natural conversational speech because they learn directly from unprocessed, mixed data and maintain local and global patterns.