PART 1: Research & Selection

Selected Models for Forgery Detection

1. Hybrid Feature-based Forgery Detection

- Reference: One-class learning towards synthetic voice spoofing detection.
- Technical Innovation:
    - Uses LFCC (Linear Frequency Cepstral Coefficient) as the primary feature extraction and representation method.
    - Implements ResNet18 as the backbone of the network for classification
    - Utilizes OC-Softmax loss, which helps in the model to learn one-class classification efficiently.
- Performance Metrics:
    - Reported high accuracy while detecting AI-Generated voices and audio.
    - Effective in identifying unseen synthetic speech attacks.

- Why use this model?
    - Uses a combination of feature extraction and deep learning techniques leading to an improved generalization.
    - Works well in real-time due to lightweight ResNet18 architecture.

- Challenges:
    - Performance may degrade on highly sophisticated deep fake audio samples.
    - One-class learning approach may require extensive fine-tuning to avoid false positives.

2. End-To-End Forgery Detection

- Reference: Spoofing attacker also benefits from self-supervised pretrained model.
- Technical Innovation:
    - Uses HuBERT and WavLM for feature extraction - both are self-supervised models trained on large speech corpora.
    - Implements Residual blocks and Conv-TasNet for robust classification.
    - Uses AAM-Softmax loss for enhancing feature discrimination.

- Performance Metrics:
    - Achieves state of the art performance on benchmark datasets.
    - Highly effective in generalizing across different types of spoofing attacks.

- Why use this model?
  - Uses self-supervised learning, which allows better feature extraction from raw audio.
  - Suitable for real-world applications where AI-generated voices constantly evolve.

- Challenges:
  - Computationally expensive due to large models.
  - Requires significant hardware resources for real-time detection.

3. Feature Fusion-based Forgery Detection **(Implemented Model)**
    - Reference: Voice spoofing countermeasure for synthetic speech detection.
    - Technical Innovation:
      - Extracts features using **GTCC**, **MFCC(Mel-Frequency Cepstral Coefficients)**, S**pectral flux**, **Spectral centroid**.
      - Uses **Bi-LSTM** as the primary network structure for sequential audio modeling.

- Performance Metrics:
  - Show high detection accuracy by leveraging multiple feature types.
  - Captures both short-term and long-term dependencies in speech.

- Why use this model?
  - More interpretable compared to end to end deep learning models.
  - Works well for analyzing real human speech and AI-generated variations.

- Challenges:
  - Bi-LSTM can be computationally intensive compared to CNNs.
  - May require feature engineering adjustment for different datasets.

PART 2: Implementation

Selected Model: Feature Fusing-based Forgery Detection

Dataset used:

- SceneFake
  https://www.kaggle.com/datasets/mohammedabdeldayem/scenefake

```
datasset_path/
    |- train
        |- real
        |- fake
    |- dev
        |- real
        |- fake
    |- eval
        |- real
        |- fake
```

Implementation Steps:-

1. Feature Extraction
    a. Extracted GTCC, MFCC, Spectral flux, and Spectral centroid for each audio file.
    b. Used librosa for feature computation.

2. Model Architecture
    a. Implemented a Bi-LSTM network with:
        - Input layer: Concatenated extracted features.
        - LSTM layers: Two Bi-LSTM layers for temporal modeling.
        - Dense layer: Fully connected with Softmax for classification.

3. Training Process
    a. Used cross-entropy loss for classification.
    b. Optimized using Adam optimizer.
    c. Trains for 100 epochs with batch size of 64.

PART 3: Documentation & Analysis

Challenges Encountered

- Feature Selection: Choosing the right combination of spectral features was crucial for improving classification.
- Computational Cost: Bi-LSTM models are slower than CNN-based models, making real-time processes difficult.
- Data Imbalance: This particular dataset had more real or fake images and not equal.

Model Analysis

Why This Model Was Selected?

- Balances interpretability and performance.
- Feature fusion ensures robustness to different AI-generated speech types.
- Suitable for real conversations and generalizes well.

How the Model Works

- Extracts spectral and temporal features from input speech.
- Uses Bi-LSTM to analyze sequential dependencies in features.
- Outputs a probability score for detecting AI-generated vs. real speech.

Model Performance: After 100 epochs, the model achieved the following results:-

- Training Accuracy: **96.16%**
- Validation Accuracy: **92.70%**
- Final Model Accuracy on Test Set: **95.25%**

These above results indicate that the Bi-LSTM model effectively captures the distinguishing characteristics of real vs fake speech.

Observed Strengths and Weaknesses

- Strengths:
  - Strong feature fusion improves robustness.
  - Bi-LSTM captures contextual dependencies in audio.
- Weaknesses:
  - Slow inference time compared to CNN-based models.
  - Struggles with extremely high-quality AI-generated voices.
  -
- Key Observations:
- The high accuracy suggests that the selected feature set (GTCC, MFCC, Spectral Flux, Spectral Centroid) is effective in detecting AI-generated

speech.
The **validation accuracy (92.70%)** is slightly lower than **training accuracy (96.16%**), suggesting a small generalization gap, which could be further improved with data augmentation or regularization.
The final **test accuracy of 95.25%** confirms strong generalization of the model on unseen data.

- Future Improvements
- Experiment with additional feature sets like chroma features or prosody-based features for better differentiation.
  Optimize Bi-LSTM architecture (e.g., tuning hidden layer size, dropout rates) for even better generalization.
  Consider real-time processing optimizations to improve inference speed for real-world deployment.

Reflection Questions

1. What were the most significant challenges in implementing this model?
   - Optimizing feature selection for best performance.
   - Training Bi-LSTM efficiently on large datasets.
2. How might this approach perform in real-world conditions vs. research datasets?
   - Might struggle with real-time applications due to LSTM overhead.
   - Requires fine-tuning for different languages and accents.
3. What additional data or resources would improve performance?
   - More diverse synthetic audio samples from different AI generators.
   - Using self-supervised models like HuBERT for feature extraction.
4. How would you approach deploying this model in a production environment?
   - Optimize for low-latency inference by replacing LSTM with a CNN-based model.
   - Use quantization techniques to reduce model size.
   - Deploy as an API service for real-time voice authentication.

Conclusion

This project explored three state-of-the-art forgery detection approaches for AI-generated speech detection. The Feature Fusion-based Bi-LSTM model was implemented, demonstrating high accuracy but also highlighting challenges in real-time detection. Future work can explore transformers and attention-based methods to improve efficiency and performance.