**Introduction**

This project addresses the challenge of evaluating textual coherence and grammatical quality from spoken audio inputs by transforming them into feature-rich representations. The objective was to build a system capable of predicting human-aligned labels on a scale from 0 to 5, combining audio processing, transcription, and advanced ensemble learning.

By implementing a meta-ensemble approach that leverages both handcrafted linguistic features and transformer-based representations, this pipeline aims to maximize prediction accuracy and correlation with human-labeled ground truth.

**Dataset and Preprocessing**

The dataset consists of:

Audio files with corresponding label scores (Likert-style scale from 0–5)

Transcriptions generated using OpenAI Whisper (base)

Derived transcripts were cleaned by removing disfluencies and normalizing structure

Audio processing steps included:

Resampling to 16 kHz

Silence trimming and volume normalization

Speech detection, enabling selective transcription to optimize resources

**Feature Engineering**

From the cleaned transcripts, the following features were extracted:

Grammar Errors: Identified using custom rule-based logic

Syntactic Structure: POS tag diversity, average sentence length, word count

GEC Metrics: Number of grammar correction edits using T5-based model

Derived Metrics: Error rate per word, GEC edit rate, lexical diversity

Readability: Flesch-style readability score

These were used to build structured representations of the spoken responses.

**Model Architecture**

**Feature-Based Ensemble**  
Three regressors were trained on the engineered features:  
Random Forest  
LightGBM  
Ridge Regression  
**Validation Results**:  
 MAE: 0.820  
 RMSE: 1.016  
Pearson Correlation: 0.516

**DistilBERT Regression**The cleaned transcripts were also fed into a DistilBERT model fine-tuned for regression. Texts were tokenized, embedded, and trained to predict scores directly using Hugging Face's Trainer API.The model was trained for 12 epochs with early stopping disabled, leveraging GPU acceleration where available. **Training Loss: 0.095000**

**Meta-Ensemble Fusion**To combine strengths of both approaches, a Linear Regressor was trained on stacked predictions from the DistilBERT and feature ensemble outputs. This final layer provides the most balanced and robust prediction.

**Final Validation Results:**MAE: 0.599RMSE: 0.769 Pearson Correlation: **0.751**This high correlation illustrates the model's strong alignment with human annotation.

**Evaluation and Discussion**The meta-ensemble strategy proved highly effective in leveraging both linguistic features and semantic depth from pretrained transformers. It demonstrates significant improvements in all metrics compared to feature-only models, especially in Pearson correlation (0.751 vs. 0.516), highlighting its reliability in reflecting score trends.

**Observations:**  
Non-speech samples were effectively handled using median imputation.  
Feature engineering captured surface-level errors while BERT captured deeper semantics.  
The linear fusion added a final optimization layer for prediction refinement.

**Conclusion**This pipeline demonstrates how hybrid modeling—rooted in both traditional features and deep learning—can achieve strong results even on moderately sized datasets. The final Pearson correlation of 0.751 suggests high agreement with human scoring and positions the model as a scalable solution for automated spoken language assessment.