**1. Introduction**

In an era of digital language learning and automated assessments, evaluating a speaker's grammatical proficiency in spoken English presents a unique challenge. Traditional grammar checks focus on written language, often ignoring the intricacies of spoken responses, such as disfluencies, informal phrasing, and variable sentence structure.

This project presents a robust, scalable Grammar Scoring Engine capable of predicting grammar proficiency on a continuous 0–5 scale based on short (45–60 seconds) audio recordings of spoken English. The model emulates human scoring using a hybrid approach, combining speech recognition, NLP-based grammatical feature extraction, and ensemble machine learning models.

**2. Dataset Overview**

The dataset comprises:

* **444 training audio files** (.wav format) with human-rated grammar scores
* **195 test audio files**
* Labels are based on the **Mean Opinion Score (MOS)** Likert scale (0 to 5), representing a learner's grammatical control and sentence structure complexity.

Each audio response is 45 to 60 seconds in duration and may include disfluencies or informal speech patterns typical of non-native English speakers.

**3. System Architecture**

**3.1 Audio Preprocessing**

* Resampling to 16 kHz mono
* Silence trimming and volume normalisation

**3.2 Speech Transcription**

* Utilised **OpenAI Whisper (base)** model for ASR
* Output transcripts retain ungrammatical speech for scoring integrity

**3.3 Transcript Cleaning**

* Removal of disfluencies ("um", "uh", "like", etc.)
* Whitespace and punctuation correction
* Lowercasing and trimming false starts

**3.4 Feature Engineering**

Extracted from cleaned transcripts:

* **Grammar error counts** via language\_tool\_python
* **Syntactic complexity**: POS diversity, sentence length
* **GEC-based features**: number of edits via a T5-based grammar correction model
* **Derived metrics**: error rate per word, GEC edit ratio

**4. Model Design**

**4.1 DistilBERT Regression**

* Fine-tuned distilbert-base-uncased using Hugging Face's Trainer
* Regression task with MSE loss to predict scores directly from transcript
* Achieved strong correlation with human scores (r = 0.56)

**4.2 Feature-Based Ensemble**

* RandomForest, LightGBM, and Ridge Regression trained on engineered features
* Captures explicit grammar issues and structural patterns

**4.3 Hybrid Meta-Ensemble**

* Combined BERT predictions and feature ensemble outputs
* Trained a Linear Regressor as a meta-model
* Final validation metrics:
  + **MAE**: 0.763
  + **RMSE**: 0.911
  + **Pearson Correlation**: 0.625

**5. Test Set Prediction and Submission**

* Preprocessed and transcribed test audio files
* Extracted all features as with training data
* Ran inference through both BERT and feature models
* Applied the meta-regressor for final grammar score prediction
* Scores were rounded to nearest integer and saved as submission.csv

**6. Evaluation and Discussion**

The hybrid approach proved highly effective for small data settings. By leveraging both deep semantic understanding and explicit grammar features, the system emulates expert rater judgment. The final correlation of 0.625 approaches the level seen in commercial automated scoring systems with much larger training data.

Limitations:

* ASR may still misrepresent non-standard grammar
* GEC model might overcorrect learner speech
* Lacks prosodic/fluency features present in human scoring

**7. Future Work**

* Integrate prosodic and fluency cues (pauses, hesitation)
* Explore ordinal regression or quantised scoring
* Train end-to-end audio-to-score deep models (wav2vec2.0)
* Evaluate on multilingual learner speech and additional scoring dimensions (fluency, pronunciation)

**8. Conclusion**

This Grammar Scoring Engine demonstrates that even with a modest dataset, it is possible to build a highly effective and explainable model for automated spoken grammar assessment. The hybrid ensemble approach offers a balanced blend of interpretability, scalability, and linguistic depth, making it suitable for integration into language learning platforms or assessment tools.

* Tools: Python, Hugging Face Transformers, Whisper, spaCy, LightGBM, scikit-learn
* Runtime: Kaggle Notebook