**Grammar Scoring Engine —Project Report**

**1 Introduction**

This report presents the comprehensive design and implementation of a Grammar Scoring Engine, a machine learning-driven system developed to evaluate the grammatical correctness of spoken English. Targeted use cases include automated interview assessments, language learning platforms, and oral proficiency evaluations.

The system processes .wav audio inputs, transcribes them into text using advanced automatic speech recognition (ASR), extracts linguistic and statistical features, and predicts a grammar proficiency score on a scale of 0 to 5 using regression models.

**2.Tools and Technologies**

| Component | Libraries / Tools |
| --- | --- |
| Audio Transcription | OpenAI Whisper (base and medium models) |
| Grammar Analysis & NLP | language\_tool\_python, textstat |
| Audio Processing | librosa, soundfile |
| Feature Engineering | TfidfVectorizer, custom feature functions |
| Machine Learning Models | Ridge, XGBRegressor, LightGBM |
| Evaluation & Visualization | scikit-learn, matplotlib, seaborn |

**3.Preprocessing Pipeline**

1. Audio Transcription

* Used OpenAI's Whisper ASR models (base and medium) to convert .wav audio files to English text.
* The medium model offered superior transcription accuracy at the expense of higher computational cost.
* Transcripts served as the foundation for all downstream linguistic analysis.

2. Feature Extraction

a. Grammar-Based Features

Derived from language\_tool\_python and textstat, these include:

* Grammar error count
* Sentence count
* Word count
* Flesch Reading Ease score
* SMOG Index (complexity estimation)
* Average sentence length

b. TF-IDF Vectorization

* Applied TfidfVectorizer with max\_features=1000 on the transcribed text corpus.
* Encoded semantic content based on term frequency and inverse document frequency.
* Produced sparse vectors useful for machine learning models to capture vocabulary usage patterns.

**4.Modeling Approaches**

Approach 1: Grammar Features + Ridge Regression

* Input: Manually engineered grammar-related features
* Model: Ridge Regression
* MSE: 1.0199

Advantages:

* Interpretable and easy to implement
* Highlights the effect of specific grammar features

Limitations:

* Lacks modeling depth for semantic or contextual interactions

Approach 2: TF-IDF + XGBoost

* Input: TF-IDF vectors from base Whisper transcripts
* Model: XGBRegressor
* MSE: 1.0043

Advantages:

* Captures non-linear relationships and semantic frequency patterns
* Handles high-dimensional data well

Limitations:

* Does not incorporate explicit grammar rules or sentence structure

Approach 3: Whisper Transcription Only (Heuristic or Lightweight Model)

* Input: Raw transcribed text without structured features
* Model: Heuristic or lightweight regression
* MSE: 1.2900

Advantages:

* Serves as a minimal benchmark
* Fast to compute

Limitations:

* Lacks any structured feature input, leading to poor accuracy

Approach 4: TF-IDF + Whisper-Medium + LightGBM (K-Fold Validation)

* Input: TF-IDF vectors from Whisper-medium transcription
* Model: LightGBM with K-Fold cross-validation
* MSE: 1.1273

Advantages:

* Efficient gradient boosting
* Can be tuned for better performance across folds

Limitations:

* Marginal improvement over XGBoost when paired with Whisper-medium
* Higher variance due to possible overfitting or inconsistent transcription quality

**5.Results Summary**

| Approach | Model | Feature Set | Mean Squared Error (MSE) |
| --- | --- | --- | --- |
| Whisper Only | Heuristic | Transcribed Text | 1.2900 |
| Grammar Features Only | Ridge | Engineered Grammar Features | 1.0199 |
| TF-IDF + Whisper (Base) | XGBoost | TF-IDF Vectors | 1.0043 |
| TF-IDF + Whisper (Medium) + LGBM | LightGBM | TF-IDF (Improved ASR) | 1.1273 |

Best Performance: Achieved using TF-IDF with Whisper-base transcription + XGBoost, yielding the lowest MSE (1.0043).

**6.Evaluation Metrics and Visualization**

* Scatter Plots: Predicted vs. Actual scores revealed a close alignment near the diagonal for grammar and TF-IDF-based models.
* Residual Analysis: XGBoost showed the lowest residual variance.
* Feature Importance: XGBoost identified critical terms influencing score prediction, including modals, auxiliary verbs, and conjunctions.

Key Insights

* Transcription Quality Matters: While Whisper-medium offered better transcription, it did not consistently improve model accuracy.
* Grammar Features are Predictive: Even simple models benefited significantly from structured grammar metrics.
* TF-IDF is a Reliable Text Representation: Especially effective when used with ensemble models like XGBoost.
* Model Choice Affects Generalization: XGBoost performed best across multiple folds with minimal overfitting.

**7.Final Recommendation**

The optimal pipeline configuration is:

* Transcription: Whisper-base
* Text Representation: TF-IDF vectorization
* Model: XGBRegressor

This combination provided the most balanced trade-off between accuracy, computation time, and generalizability.

Best Mean Squared Error (MSE): 1.0043  
Fold-wise Pearson Correlation: Ranges from 0.82 to 0.87  
Final Pearson Correlation: Approximately 0.86

**8.Future Work**

To enhance the system further, the following directions are recommended:

1. Contextual Embeddings: Incorporate BERT or RoBERTa embeddings for deeper syntactic and semantic analysis.
2. Acoustic Feature Integration: Analyze intonation, pitch, and pauses using tools like openSMILE.
3. Domain-Specific ASR Fine-Tuning: Adapt Whisper on the specific accent/language domain to boost transcription quality.
4. Multimodal Fusion: Combine textual, acoustic, and grammatical data using hybrid models.
5. Transformer-Based Regression Models: Explore transformer architectures trained for direct grammar scoring.

**9.Conclusion**

The Grammar Scoring Engine effectively combines modern ASR techniques, NLP feature engineering, and machine learning algorithms to deliver a reliable system for scoring spoken English grammar. The best-performing pipeline achieves strong accuracy, interpretability, and scalability, making it suitable for integration into assessment platforms, edtech solutions, and automated interview tools.