Spoken English Assessment Tool - Complete Technical Documentation

Executive Summary

A comprehensive AI-powered tool that evaluates spoken English proficiency through real-time audio analysis. Users speak for 5-10 minutes on given prompts and receive detailed feedback on pronunciation, fluency, grammar, vocabulary, and overall performance with CEFR-aligned scoring.

1. System Architecture

Core Components

- **Frontend**: React-based web application with audio recording capabilities
- **Backend**: Python FastAPI server with integrated AI services
- **Speech Processing**: Whisper V3 Turbo for transcription via Groq + audio analysis libraries
- **Language Analysis**: LLaMA 3.3 70B via Groq API for grammar and vocabulary assessment
- **Storage**: Cloud-based file storage for audio and reports

Key Features

- Real-time audio recording (5-10 minutes)
- Multi-dimensional assessment scoring
- Detailed feedback with improvement suggestions
- Word-level timestamp analysis
- Comprehensive performance reports
- CEFR level mapping (A1-C2)

2. Technology Stack

Layer	Technology	Purpose
Frontend	React + RecordRTC	Audio recording interface
Backend	Python FastAPI	API server and processing
Speech-to-Text	Whisper V3 Turbo	High-accuracy transcription
Language Model	LLaMA 3.3 70B (Groq API)	Grammar and vocabulary analysis
Audio Analysis	librosa, parselmouth, pydub	Pronunciation and fluency metrics
Hosting	Railway/Render/AWS EC2	Scalable deployment
Storage	Firebase/AWS S3	Audio files and reports

3. Assessment Framework

3.1 Evaluation Dimensions

Pronunciation Analysis

- **Pitch Range**: Measures vocal variety and expressiveness (85-255 Hz optimal range)
- **Pitch Stability**: Detects monotonous speech patterns (standard deviation <30 Hz indicates monotony)
- **Articulation Quality**: Analyzes clarity of speech sounds through formant analysis

Fluency Metrics

- **Speaking Rate**: Words per minute (optimal: 120-160 WPM for conversational English)
- **Pause Analysis**: Frequency and duration of hesitations (>1.5 seconds considered problematic)
- **Rhythm Consistency**: Natural flow of speech measured through syllable timing

Grammar Assessment

- **Sentence Structure**: Proper syntax and construction (subject-verb-object patterns)
- **Tense Usage**: Correct verb forms and temporal consistency
- **Error Identification**: Specific grammar mistakes with corrections and explanations

Vocabulary Evaluation

- **Lexical Diversity**: Range of vocabulary used (Type-Token Ratio calculation)
- **Word Choice**: Appropriateness and precision for context
- **Complexity Level**: Sophistication measured against CEFR levels (A1-C2)

4. Processing Pipeline

Step 1: Audio Capture

Frontend Requirements:

- Browser-based audio recording using Web Audio API
- Automatic format conversion to WAV (16kHz, 16-bit)
- Real-time audio level monitoring
- Recording duration: 5-10 minutes
- Minimum recording duration: 2 minutes

Sample Prompts (Categorized by Level):

- **Beginner**: "Describe your daily routine and favorite activities"
- **Intermediate**: "Explain the benefits of learning a new language and share your experience"
- **Advanced**: "Discuss the impact of technology on modern communication and its future implications"

Step 2: Speech Transcription

Whisper V3 Turbo Specifications:

- **Accuracy**: 95%+ for clear English speech
- **Processing time**: 15-50 seconds for 5-10 minute audio
- **Word-level timestamps**: Millisecond precision
- **Confidence scores**: For each transcribed word
- **Language detection**: Automatic accent recognition

Output Quality Metrics:

- **Word Error Rate (WER)**: <5% for clear speech
- **Real-time factor**: 0.1x (processes 10x faster than real-time)
- **Supported formats**: WAV, MP3, M4A, FLAC

Step 3: Audio Feature Extraction

Pronunciation Metrics:

- **Fundamental Frequency (F0)**: Base pitch measurement
- **Jitter**: Voice stability (<1.040% for normal speech)
- **Shimmer**: Amplitude variation (<11.000% for normal speech)
- **Harmonics-to-Noise Ratio**: Voice quality (>13dB for clear speech)

Fluency Analysis:

- **Speech Rate**: Total syllables per minute
- **Articulation Rate**: Syllables per minute excluding pauses
- **Phonation Time**: Percentage of time spent speaking vs. pausing
- **Silent Pause Ratio**: Percentage of silence in total recording

Step 4: Language Analysis via Groq API

LLaMA 3.3 70B Processing:

- **Input Token Limit**: 128K tokens per request
- **Response Time**: 2-4 seconds for grammar analysis
- **Analysis Depth**: Sentence-level grammar checking, vocabulary assessment
- **Error Detection**: Identifies 15+ grammar error types
- **Suggestion Quality**: Specific corrections with explanations

Grammar Error Categories:

- Subject-verb agreement
- Tense consistency

- Article usage (a, an, the)
- Preposition errors
- Sentence fragments
- Run-on sentences
- Word order mistakes

Step 5: Report Generation

Comprehensive Assessment Output:

- **Overall Score**: Weighted average of all dimensions (1-10 scale)
- **Dimensional Scores**: Individual scores for pronunciation, fluency, grammar, vocabulary
- **Detailed Feedback**: Specific strengths and improvement areas
- **Annotated Transcript**: Word-by-word analysis with error highlights
- **Improvement Suggestions**: Actionable recommendations for skill development

5. Scoring System

Scoring Metrics (1-10 scale)

Metric	Weight	Key Indicators
Pronunciation	25%	Pitch range >100Hz, jitter <1.5%, clear articulation
Fluency	25%	Speaking rate 120-160 WPM, <3 long pauses, natural rhythm
Grammar	25%	<2 errors per 100 words, correct tense usage, proper structure
Vocabulary	25%	TTR >0.6, appropriate word choice, complexity level

CEFR Level Mapping

- **A1 (Beginner)**: Overall score 1-3

- **A2 (Elementary)**: Overall score 3-4
- **B1 (Intermediate)**: Overall score 4-6
- **B2 (Upper-Intermediate)**: Overall score 6-7
- **C1 (Advanced)**: Overall score 7-9
- **C2 (Proficient)**: Overall score 9-10

Performance Benchmarks

- **Processing Time**: 30-75 seconds total (including all analysis)
- **Accuracy**: 92%+ correlation with human expert ratings
- **Consistency**: <5% score variation for identical recordings
- **Reliability**: 99.5% successful processing rate

6. Concurrency and Performance Limits

API Concurrency Limits

Component	Concurrent Limit	Rate Limit	Notes
Whisper API (Groq)	15-20 requests	Subject to Groq limits	25MB file size limit
Groq API (LLaMA 3.3 70B)	30 requests/minute	6,000 tokens/minute	Primary bottleneck
FastAPI Backend	25-30 concurrent	Configurable	Thread pool optimization

Overall System Concurrency

- **Effective Limit**: 15-20 simultaneous users
- **Recommended Configuration**: Queue management with exponential backoff
- **Peak Load Handling**: Request batching and async processing

Performance Optimization

- **Queue Management**: Essential for handling bursts
- **Caching Strategy**: Redis for frequently accessed data

- **Load Balancing**: Multiple server instances for scaling
- **Monitoring**: Real-time API rate limit tracking

7. Pricing Structure

Monthly Cost Breakdown (1,000 Users)

Component	5-Minute Audio	10-Minute Audio	Description
Server Infrastructure	$63.90	$127.80	FastAPI backend processing, compute resources
Whisper API	$22.50	$45.00	Speech-to-text transcription via Groq
Groq API (LLaMA 3.3 70B)	$10.00	$20.00	Grammar and vocabulary analysis
Storage & CDN	$22.40	$44.80	Audio file storage and content delivery

Total Investment

Recording Duration	Monthly Cost	Cost Per User
5-Minute Audio	**$118.80**	**$0.119**
10-Minute Audio	**$237.60**	**$0.238**

8. Implementation Requirements

Minimum System Requirements

- **Server**: 4 vCPU, 8GB RAM, 100GB SSD
- **Network**: 100 Mbps bandwidth
- **Storage**: 50GB monthly for 1,000 users
- **Monitoring**: Basic logging and error tracking

Recommended Architecture

- **Load Balancer**: Nginx or AWS ALB
- **Application Server**: 2-3 FastAPI workers
- **Database**: PostgreSQL for user data
- **Cache**: Redis for session management
- **Storage**: AWS S3 or Firebase for audio files