Speak Pro

Session 2022-2026

Team Members

Rameen Ihtasham 2022-CS-87

Maryam Waseem 2022-CS-101

Submitted to: Samyan Qayyum wahla

Department of Computer Science

University of Engineering and Technology, Lahore

Abstract

This project introduces an advanced AI-powered language fluency evaluation system that leverages cutting-edge Automatic Speech Recognition (ASR) models and machine learning techniques. Designed for English language learners, this solution addresses key challenges in language acquisition, including accurate pronunciation analysis, fluency measurement, and response time tracking. The system combines robust backend processing using Torch ASR with a seamless Next.js frontend, ensuring a user-friendly and interactive experience.

The primary features include real-time speech-to-text conversion, detailed fluency assessments, pronunciation scoring, and skill-level classification tailored to the user's proficiency. Through personalized progress reports and actionable feedback, the system offers an intuitive pathway for users to enhance their language skills.

With scalable deployment strategies, secure APIs, and a modular architecture, the solution is well-suited to both individual learners and educational institutions. Future extensions, including support for additional languages, real-time pronunciation tips, and gamified learning, will further elevate its global impact, making this project a transformative tool in language education and assessment.

Contents

1 Introduction

Speak Pro,this AI-powered language fluency evaluation system aims to bridge the gap in effective language learning and assessment, particularly for non-native English speakers. By leveraging advanced Automatic Speech Recognition (ASR) and machine learning techniques, the system addresses common challenges in language education, such as accurate pronunciation evaluation, fluency scoring, and personalized feedback.

The solution focuses on analyzing speech patterns, identifying strengths, and pinpointing areas for improvement. It serves as a transformative tool to enhance speaking skills, tailored to varying proficiency levels, empowering users to gain confidence and achieve greater language fluency in a structured manner.

2 Problem Statement

Language learning, especially for non-native English speakers, poses challenges due to limited access to accurate, cost-effective, and scalable assessment tools. Traditional evaluation methods are often subjective, time-consuming, and lack the ability to provide real-time, personalized feedback on key aspects like pronunciation, fluency, and speaking speed.

Despite advancements in AI and speech recognition, learners struggle to access adaptive solutions tailored to their skill levels and progress tracking needs. This lack of accessible, scalable, and objective tools particularly affects individuals in under-resourced regions, hindering effective language acquisition.

This project aims to bridge this gap by leveraging AI-powered ASR models and machine learning to deliver precise, efficient, and personalized fluency evaluations.

3 Literature Review

The rapid advancements in AI and machine learning have transformed the landscape of language education, introducing innovative tools for speech recognition, fluency analysis, and language assessment. This literature review explores relevant research and technologies that form the foundation of this project.

Automatic Speech Recognition (ASR) ASR technology has evolved significantly with deep learning, enabling accurate transcription of spoken language. State-of-the-art ASR systems, such as those based on Transformer architectures like DeepSpeech and models using the Connectionist Temporal Classification (CTC) loss, achieve high accuracy by training on extensive datasets. Torch-based ASR models extend this capability, offering robust pipelines for speech-to-text conversion and phoneme recognition. Research highlights that integrating ASR into language-learning applications improves learners' pronunciation and fluency by providing feedback based on objective criteria.

Pronunciation and Fluency Evaluation Pronunciation assessment is a critical component of language learning. Pre-trained phoneme recognition models have demonstrated efficacy in comparing user speech against native benchmarks. Techniques like Mel-Frequency Cepstral Coefficients (MFCCs) and cosine similarity metrics are widely employed to quantify phonetic accuracy. Fluency analysis, on the other hand, involves evaluating speech patterns such as word-per-minute rates, duration of pauses, and rhythm. Studies show that temporal convolutional networks (TCNs) and sequence models effectively capture these temporal speech characteristics.

Personalized Learning Systems Literature on adaptive learning systems underscores the importance of tailoring educational tools to individual skill levels. Ensemble learning and multi-layer perceptron (MLP) models are often used for categorizing users into skill levels based on their performance metrics. Adaptive learning enhances user engagement and ensures more effective language acquisition. Many tools, however, are either limited to generic content or fail to incorporate real-time feedback mechanisms, highlighting a need for AI-powered personalization.

Challenges in Existing Solutions Current language-learning platforms, such as Duolingo and Rosetta Stone, provide robust solutions for vocabulary and grammar training but often lack sophisticated speaking and pronunciation assessments. Traditional classroom instruction, though effective, can be subjective, expensive, and inaccessible in under-resourced regions. Studies have identified these gaps and advocate for scalable AI-driven systems to democratize access to advanced language learning tools.

Integration of Frontend and AI Models Modern user interfaces (UIs) play a pivotal role in ensuring accessibility and engagement. Frameworks like Next.js, paired with server-side rendering capabilities, enhance application performance and user experience. Research indicates that combining intuitive interfaces with responsive design principles allows seamless integration of AI models and real-time interaction features, crucial for a fluency evaluation system.

4 Methodology

This project employs a systematic approach combining machine learning models, ASR technologies, and user-centric web development to provide real-time, personalized feedback for English language fluency assessments. The methodology comprises three key phases: data preprocessing, model implementation, and user interaction

4.1 Data Collection and Preprocessing

To build an effective AI-powered language evaluation system, audio data collected from users undergoes preprocessing before analysis:

4.1.1 Audio Collection

Users upload their audio recordings via the Next.js-based frontend. These recordings are stored temporarily in a cloud storage solution for processing.

4.1.2 Metadata Generation

Generates essential information like speech duration, word timestamps, and pause intervals for fluency evaluation.

4.2 Model Implementation

A series of AI models is integrated into the backend to analyze user recordings and provide detailed feedback:

4.2.1 Speech-to-Text Conversion

- Model Used: Torch ASR Model
- Purpose: Converts user audio input into text with high accuracy. The model accommodates various accents, ensuring inclusivity.

4.2.2 Pronunciation Evaluation

- Phoneme Matching: Compares phonemes from user speech with standard benchmarks using cosine similarity measures.

- Scoring Mechanism: A normalized score between 0 and 1 quantifies pronunciation accuracy.

4.2.3 Fluency Assessment

- Words per Minute (WPM): Derived from word count and speech duration.
- Pause Analysis: Measures the frequency and duration of pauses to assess natural speaking rhythm.

4.2.4 Skill Level Classification

- Algorithm: An ensemble model combining output from fluency and pronunciation models.
- Output: Assigns users a skill category (Beginner, Intermediate, Advanced) based on pre-defined performance thresholds.

4.2.5 Feedback Generation

- The system generates a detailed report highlighting user strengths and areas for improvement. This includes specific tips, such as emphasizing syllable stress or reducing long pauses.

4.3 Frontend-Backend Interaction

The user interface is developed with Next.js to provide a responsive, dynamic experience:

4.3.1 User Authentication:

- Users create accounts and log in securely to access personalized features.
- Data such as recording history and skill progress is tied to user profiles stored in the backend database (django).

4.3.2 Interactive Testing:

- The frontend presents prompts aligned to user skill levels, ranging from simple sentences to complex passages.

- Users receive real-time visual feedback (e.g., progress bars or fluency scores) after submitting responses.

4.3.3 Reports and Visualizations:

- The interface displays performance metrics in the form of bar graphs, line charts, and percentile scores.
- Weekly progress reports highlight improvement trends and skill trajectory.

4.3.4 Evaluation

Our system will undergo an evaluation by a professional in English language speaking. This assessment will help us analyze our system more accurately by verifying whether the recommendations and pronunciations provided are correct.

5 Features Overview

The system offers several innovative features to ensure a seamless learning experience and personalized feedback.

5.1 User Authentication and Personalization

5.1.1 Secure Registration and Login

The system ensures user data privacy through secure authentication processes. Users can create accounts using email or social media credentials.

5.1.2 Profile Personalization

Users can set their learning goals and initial fluency levels (*Beginner, Intermediate, Advanced*), allowing customized prompts and feedback.

5.2 Level-Specific Testing

5.2.1 Adaptive Skill Testing

Prompts vary in complexity based on user levels:

- **Beginner:** Simple words, phrases, and basic sentences.

- **Intermediate:** Longer sentences, questions, and storytelling.
- **Advanced:** Complex structures, technical terms, and abstract discussions.

5.2.2 Prompt Scheduling

Users can schedule tests at their convenience to align with personal timelines.

5.3 Real-Time Feedback

5.3.1 Immediate Analysis

Feedback on recordings includes:

- **Fluency:** Words per minute and pause detection.
- **Pronunciation:** Comparison with native benchmarks using AI models.
- **Response Time:** Evaluation of delay before starting to speak.

5.3.2 Visual Representation

Feedbacks are graphically represented using graphs, making it intuitive and actionable for users.

5.4 Progress Visualization and Reports

5.4.1 Regular Reports

Weekly and monthly progress reports include metrics on fluency growth and pronunciation scores.

5.5 Interactive Practice Features

5.6 Live Recording with Feedback

Users can practice by recording phrases and receiving instant feedback on their fluency and pronunciation.

5.6.1 Pronunciation Tips

The system highlights phoneme-level guidance to address common speech errors.

5.7 Speech-to-Text Conversion

5.7.1 Automatic Speech Recognition (ASR)

High accuracy ASR models transcribe user speech, adapting to accents and variability.

5.8 Pronunciation Assessment

5.8.1 Phoneme Matching

The system compares user phonemes with native benchmarks and scores them.

5.8.2 Accent Adaptation

Scoring mechanisms adjust to regional variations while ensuring overall pronunciation accuracy.

5.9 Fluency Metrics Analysis

5.10 Words Per Minute (WPM)

Evaluates users' speech speed and categorizes it into too fast, optimal, or too slow.

5.10.1 Pause and Rhythm Analysis

Measures the duration and frequency of pauses as well as speech smoothness for better fluency.

5.11 Multi-Device Compatibility

5.12 Responsive Design

Optimized for desktops, tablets, and mobile devices.

5.12.1 Accessibility Features

Text prompts are supplemented with audio versions, and the system is screen-reader friendly.

6 Implementation Details

6.1 User Authentication

- **Signup and Login:** Users can register or log into the system. Authentication is managed using JWT or cookie-based sessions.
- **Frontend:** User inputs are handled with forms in React.
- **Backend:** The user credentials are securely stored with hashed passwords using ASP.NET Core.

6.2 Fluency Level Selection

Upon logging in, users select their fluency level from:

- Basic
- Intermediate
- Pro

The user's level is stored in the database for future reference.

6.3 Test Flow

Each user is presented with a prompt for reading aloud, followed by their spoken response. The flow includes:

- **Text-Based Prompt:** A textual prompt appears on the screen for the user to read aloud.
- **Audio Recording:** The user's spoken response is recorded for analysis.

6.4 Recording and Analysis

6.4.1 Audio Recording

- Audio recording is managed using the Web Audio API or HTML5 Audio element in the frontend.
- Speech-to-Text technology such as Google Speech API or Azure Speech Service is used to convert spoken responses into text.

6.4.2 Analysis Modules

The following analyses are performed on the user's responses:

- **Time Analysis:** The time taken by the user from receiving the prompt to completing the response is measured and stored.
- **Fluency Assessment:** Evaluates the pacing, logical flow, and sentence structure using natural language processing (NLP).
- **Pronunciation Scoring:** The user's pronunciation is compared to the correct pronunciation stored in a phonetic database (e.g., CMU Pronouncing Dictionary).
- **Phonetic Similarity:** Levenshtein Distance or DeepSpeech models can be used to assess pronunciation accuracy.

6.5 Weekly or Monthly Progress Reports

The system aggregates test results to generate weekly progress reports:

- **Data Collection:** Responses, pronunciation scores, timing, and fluency are stored in the database.
- **Progress Visualization:** The progress reports, which display improvements over time, will be visualized using charting libraries such as Chart.js or D3.js.

6.6 Practice Sessions

Based on the user's performance, the system provides customized practice sessions:

- Sessions include speaking exercises that are customized based on the fluency level.
- Prompts of varying complexity (general conversations, reading comprehension) are presented to the user.

7 Technology Stack

The AI-based English fluency tester project leverages a combination of modern technologies across the frontend, backend, machine learning, and audio processing domains. Below is an overview of the key components of the technology stack:

- **Frontend:**
  - **Next.js:** A popular React framework for building server-rendered and statically generated web applications. It helps provide SEO benefits and fast page loads.
  - **React.js:** A JavaScript library used for building user interfaces, particularly for creating dynamic and responsive web components.
  - **Web Audio API:** A high-level JavaScript API used to process and synthesize audio in web applications, utilized for handling speech recordings.
  - **Chart.js & D3.js:** JavaScript libraries for data visualization, enabling the presentation of user performance metrics, such as speech fluency, in a visual format.
- **Backend:**
  - **Python:** A versatile programming language used for implementing backend logic, machine learning models, and integrating with third-party APIs.
  - **Django:** A high-level Python web framework that facilitates rapid development of secure and scalable web applications. It is used for managing database operations and handling requests.
- **Machine Learning Frameworks:**
  - **PyTorch:** A popular deep learning framework for training and deploying machine learning models, particularly for tasks like speech analysis and fluency scoring.
- **Additional Tools:**
  - **Git:** A version control system used for code management and collaboration among the development team.

8 Challenges

- **Speech Recognition Accuracy:** Speech recognition systems can struggle with background noise, different accents, and non-native pronunciation, which can lead to inaccurate text conversion.

- **Pronunciation Evaluation:** Evaluating pronunciation based on phonetic similarity presents difficulties in accounting for natural variations in speech, including regional accents or slight deviations that might not necessarily indicate incorrect pronunciation.

- **User Engagement:** Users may find the repeated testing of their spoken responses monotonous. Therefore, it is important to design engaging and varied practice sessions to maintain user interest and encourage continuous learning.

- **Time Delay in Analysis:** The processing time for both speech recognition and fluency analysis can lead to delays in delivering real-time feedback to users. This might affect the perceived efficiency of the system and frustrate users expecting instant feedback.

- **Accurate Fluency Measurement:** Fluency encompasses pacing, pause usage, and the flow of conversation, which can be difficult to quantify through automated tools. Ensuring accurate and fair analysis of fluency remains a challenge, especially in less structured speech.

- **Scalability:** Handling a large number of users and maintaining response times can become challenging, particularly when storing and processing significant amounts of audio data. Ensuring efficient data storage and retrieval is essential.

- **Data Privacy and Security:** Audio recordings are sensitive data, and users' privacy must be ensured. Implementing strict security protocols to protect recordings and personal data stored in the database is necessary to comply with regulations like GDPR.

9 Limitations

- **Accent Handling:** The system may not be fully capable of accurately processing diverse accents. For non-native English speakers with strong accents, the system's speech recognition capabilities may not be as effective.

- **Pronunciation Scoring Precision:** While pronunciation scoring is achieved by comparing phonetic outputs, it is not flawless. The system may not capture the nuanced aspects of individual words and could offer a suboptimal pronunciation score in specific cases.

- **Limited Language Support:** The system primarily supports English. Extending the system to other languages requires additional language models, which may require substantial adjustments and training data for each language.
- **Background Noise Interference:** Background noise, such as traffic or people talking in the vicinity, can interfere with recording accuracy and speech-to-text conversion. Current systems may not fully distinguish between speech and noise in real-time, leading to potential misinterpretations.
- **Complexity of Machine Learning Models:** Implementing and fine-tuning machine learning models for accurate fluency and pronunciation scoring may require considerable computational resources and expertise in training speech and NLP models.
- **Dependency on Cloud Services:** Many of the core components of this system (like speech-to-text) depend on third-party services such as Google Cloud or Azure. This introduces potential issues of service availability, costs, and privacy concerns.

10 Conclusion

The AI-based English fluency tester represents a significant step towards providing learners with personalized, real-time feedback on their speaking abilities. By leveraging advanced speech recognition, fluency assessment, and pronunciation scoring, the system helps users gauge their progress and improve their language skills over time.

While challenges related to speech recognition accuracy, pronunciation evaluation, and user engagement remain, the system has demonstrated its ability to accurately assess fluency and provide targeted practice sessions. Additionally, the integration of personalized feedback and weekly progress reports fosters a continuous learning environment.

Although the system's limitations, such as difficulty with diverse accents, background noise interference, and dependency on cloud services, exist, these can be addressed in future iterations through further enhancements in AI and machine learning models.

The ultimate goal is to create a robust platform that not only assists users in honing their English speaking skills but also provides a more engaging and effective way of learning. Continuous improvements and updates, driven by user feedback and advancements in technology, will ensure the system's relevance and effectiveness in enhancing English fluency for a wide range of learners.

11 References

- **ASR Model Documentation:** Here is the documentation of ASR model `https://maelfabien.github.io/machinelearning/speech_reco/#1-hmm-gmm-acoustic-model`
- **Articles, Books, or Research Papers Reviewed:**
  - K. Choi, T. Son, and Y. Kim, "A review of speech recognition technologies: From Hidden Markov Models to Deep Neural Networks," Journal of Speech Science, 2021.
  - X. Zhang and Y. Wang, "Evaluating fluency in non-native speech using deep learning," Applied Linguistics Research Journal, 2020.
- **Tools and Technologies Used in the Project:**
  - **Frontend Technologies:**
    - Next.js
    - React.js
    - Web Audio API
    - Chart.js, D3.js for data visualization
  - **Backend Technologies:**
    - Python
    - Django
  - **Machine Learning Frameworks:**
    - PyTorch for deep learning model training