PROJECT REPORT

Pronunciation Analyser

Submitted by :

Tanish Bhatia (102215058)

Hartinder Singh (102215064)

Navansh K. Goswami (102215193)

Sarthak Kala (102215213)

Group: 4O3B

Submitted to

Dr. Gaganpreet Kaur

&

Dr. Deepak Rakesh

THAPAR INSTITUTE
OF ENGINEERING & TECHNOLOGY
(Deemed to be University)

Department of Computer Science and Engineering

Thapar Institute of Engineering & Technology, Patiala

Aug - Dec (2025)

INTRODUCTION

The ability to be an effective communicator is essential and particularly important when learning a new language. However, a lot of people around the world struggle to be truly fluent in pronouncing words correctly. The meaning of a word can be entirely altered, or listeners can be left perplexed by even the most minute pronunciation mistakes. Since a majority of students lack a personal tutor to help them with every error, technology becomes a suitable ally.

Despite the rapid advancements in speech technology, most of the available tools are still unable to provide precise, useful feedback on pronunciation. Though it rarely assesses your speech quality, traditional speech recognition can at least transcribe what you say. We decided to fill this gap by developing a tool that actually focuses on providing proper pronunciation analysis rather than just listening.

We present a Speech Pronunciation Analyzer. It is intended to function as a virtual pronunciation coach. It uses sophisticated deep learning to compare users' speech with accurate pronunciations, whilst returning useful feedback to help them get better. Our goal is to improve the accuracy, accessibility, and enjoyment of pronunciation training.

Students learning English, speech therapists working with kids, and anyone wishing to improve their accent can all benefit from this tool. Our goal is to help users speak more confidently and clearly, one word at a time, by consolidating pronunciation scoring and speech embeddings.

DATASET

1. Audio Transcriptions using Whisper Base Model
The Whisper base model has been employed for automatic transcription of audio recordings. This model provides efficient speech-to-text conversion, which serves as the primary source of textual data for further phoneme analysis.

2.Montreal Forced Aligner (MFA) was used to align the audio signals with their corresponding transcription to generate a time aligned version of phoneme transcriptions. The alignment process leveraged the following resources:

Dictionary: english_us_arpa pronunciation dictionary, which provides phonetic transcriptions for English words.

Acoustic Model: english_us_arpa acoustic model trained for American English pronunciation patterns.
These components enabled accurate forced alignment and phoneme extraction from speech data.

3. Phoneme Conversion of Reference Text using NLTK Corpus
The Natural Language Toolkit (NLTK) corpus was utilized to generate the phonemic transcription of the reference text (ground truth). This phoneme conversion was crucial for comparing the phonemes derived from the MFA alignment with the reference phonemes, thus ensuring detailed analysis of pronunciation accuracy.

LITERATURE REVIEW

[1] **Cai et al. (2025)** – *Developing an Automatic Pronunciation Scorer: Aligning Speech Evaluation Models and Applied Linguistics Constructs*. A deep learning–based scorer trained on human-rated pronunciations, showing strong correlation with expert ratings and outperforming prior scorers researchgate.net. The model uses phone-level acoustic features and aligns with linguistic constructs to provide unbiased, real-time pronunciation feedback.

[2] **Zhang et al. (2022)** – *BiCAPT: Bidirectional Computer-Assisted Pronunciation Training with Normalizing Flows*. This Interspeech 2022 paper proposes a novel CAPT system that **simultaneously detects and corrects mispronunciations**. By combining ASR and TTS in a shared latent space via normalizing flows, the method generates corrected pronunciations in the student's speaking style while detecting errors isca-archive.org.

[3] **Korzekwa et al. (2022)** – *Computer-assisted Pronunciation Training – Speech Synthesis is Almost All You Need* (Speech Communication). The authors present three techniques (phoneme-to-phoneme, text-to-speech, speech-to-speech) for **generating synthetic mispronounced speech**. These synthetic data vastly improve error detection: the best method (speech-to-speech) boosted pronunciation-error detection AUC from 0.528 to 0.749, a 41% relative increase, achieving new state-of-the-art accuracy arxiv.org.

[4] **Ahn et al. (2025)** – *English Pronunciation Evaluation without Complex Joint Training: LoRA Fine-tuned Speech Multimodal LLM*. An arXiv (2025) study showing that fine-tuning a multimodal large language model (Phi-4) with **LoRA** yields integrated pronunciation assessment. The single model simultaneously performs automatic pronunciation assessment (APA) and mispronunciation detection/diagnosis (MDD) with high accuracy (PCC≈0.7 vs. human scores) and low WER/PER, all without heavy joint training arxiv.org.

[5] **Lakshminarayanan et al. (2025)** – *Automated Speech Therapy through Personalized Pronunciation Correction using Reinforcement Learning and Large Language Models*. A Results in Engineering (Elsevier) paper introducing an AI system that uses a custom **Reinforcement Learning (PPO)** algorithm for fine-grained pronunciation evaluation and a large language model to generate motivating, personalized feedback sciencedirect.com. Tested on datasets like CMU and TIMIT, it achieved ≈97.9% phoneme accuracy and shows how LLMs can produce rich, user-specific correction prompts.

[6] **Liu et al. (2023)** – *Leveraging Phone-level Linguistic-Acoustic Similarity for Utterance-level Pronunciation Scoring* (arXiv). Proposes measuring pronunciation deviation by explicitly computing the cosine similarity between reference phone embeddings and

acoustic embeddings of spoken phones arxiv.org. The paper introduces a GOP pre-training stage for better initialization and uses a Transformer scorer over these similarity features, significantly improving utterance-level scoring accuracy on L2 speech datasets arxiv.org.

[7] **Shahin et al. (2023)** – *Phonological-Level wav2vec2-based Mispronunciation Detection and Diagnosis Method* (arXiv). The authors propose a **speech-attribute–based** MDD approach rather than phoneme labels. They use a pre-trained wav2vec2 backbone and a multi-label CTC loss to detect articulatory features (e.g. voicing, place). This low-level attribute model significantly lowers false acceptance/rejection rates and diagnostic error rates compared to standard phoneme-level MDD ar5iv.labs.arxiv.org.

[8] **Bernhard et al. (2022)** – *Acoustic Stress Detection in Isolated English Words for CAPT* (Interspeech 2022). Addresses suprasegmental feedback by detecting **lexical stress** in single words. The pipeline segments speech into syllables, computes features (duration, intensity, pitch), and classifies each syllable as stressed/unstressed with a voting ensemble (SVM, NN, k-NN, Random Forest). This system achieved 94% F1 and 96% accuracy on English word recordings, highlighting the importance of prosodic cues in pronunciation training isca-archive.org.

[9] **Shekar et al. (2023)** – *Assessment of Non-Native Speech Intelligibility using Wav2vec2-based MDD and Multi-level Goodness of Pronunciation Transformer* (Interspeech 2023). Combines a Wav2vec2-based mispronunciation detector with a Transformer that ingests multi-level GOP features. By using an L2 speech dataset annotated for suprasegmental (prosodic) labels, the system jointly evaluates segmental and suprasegmental pronunciation. This multi-granular approach links automated scores to listener intelligibility and offers insight into aspects of L2 speech affecting comprehension isca-archive.org.

[10] **Ryu et al. (2023)** – *A Joint Model for Pronunciation Assessment and Mispronunciation Detection and Diagnosis with Multi-task Learning* (Interspeech 2023). Observing the high correlation between overall pronunciation scores and phonetic error rates, the authors train a **single Wav2Vec2-based model** with both CTC (for phoneme recognition/MDD) and cross-entropy (for scoring) heads. Multi-task learning yields mutual gains: APA scoring PCC improves by ~0.057 and MDD F1 by ~0.004 over separate models isca-archive.org, validating a unified approach to CAPT tasks.

METHODOLOGY

Our system evaluates spoken pronunciation by comparing user input(in the form of audio) against a reference text.
Below is the step-by-step outline of the checking procedure.

Data Pre-Processing:

To ensure accurate analysis, we first standardize and refine raw audio inputs:
Resampling: All recordings are converted to a 16 kHz mono format, optimizing compatibility with modern speech models.

Audio Transcription:

We have employed the Whisper model, developed by OpenAI. It is a powerful, end-to-end neural network designed for automatic speech recognition (ASR). It converts spoken audio into written text — that's transcription.

Features of Whisper :

- It takes raw audio input in the form of .wav, which is resampled at 16 kHz mono audio.
- It performs feature extraction internally, i.e., converts the raw waveform into a log-Mel spectrogram, a type of time-frequency representation that emphasises perceptually important features.
- Then applies an Encoder-Decoder Transformer, which is described below.
    - Encoder: Processes the log-Mel spectrogram and captures information about the audio content (e.g., speech, background).
    - Decoder: Uses the encoded representation to generate text, word by word, using language modeling and beam search to find the most likely transcription.
- Since Whisper is trained on 680,000 hours of multilingual and multitask supervised data. It can handle different languages, accents, noisy environments, and even perform language detection.

Phoneme Analysis

The input audio and the generated transcription from 4.2 are first aligned using the Montreal Forced Aligner. This generates spoken phonemes and is stored in a TextGrid file. This file is then parsed to correctly display the spoken phonemes with timestamps. The ideal phonemes are generated from the reference text using the Nltk corpus's cmudict. This is then used to give the required comparison and analysis.

Scoring: Measuring Progress

Lastly, the scores are generated in terms of three metrics: Accuracy, Fluency, and Overall.

- Accuracy tells how close the user's phonemes were to the ideal phonemes, which are generated from the phonemes of the reference text.
- Fluency is how close the speed of talking is to an average speaker.
- Overall returns the weighted average of accuracy and fluency.

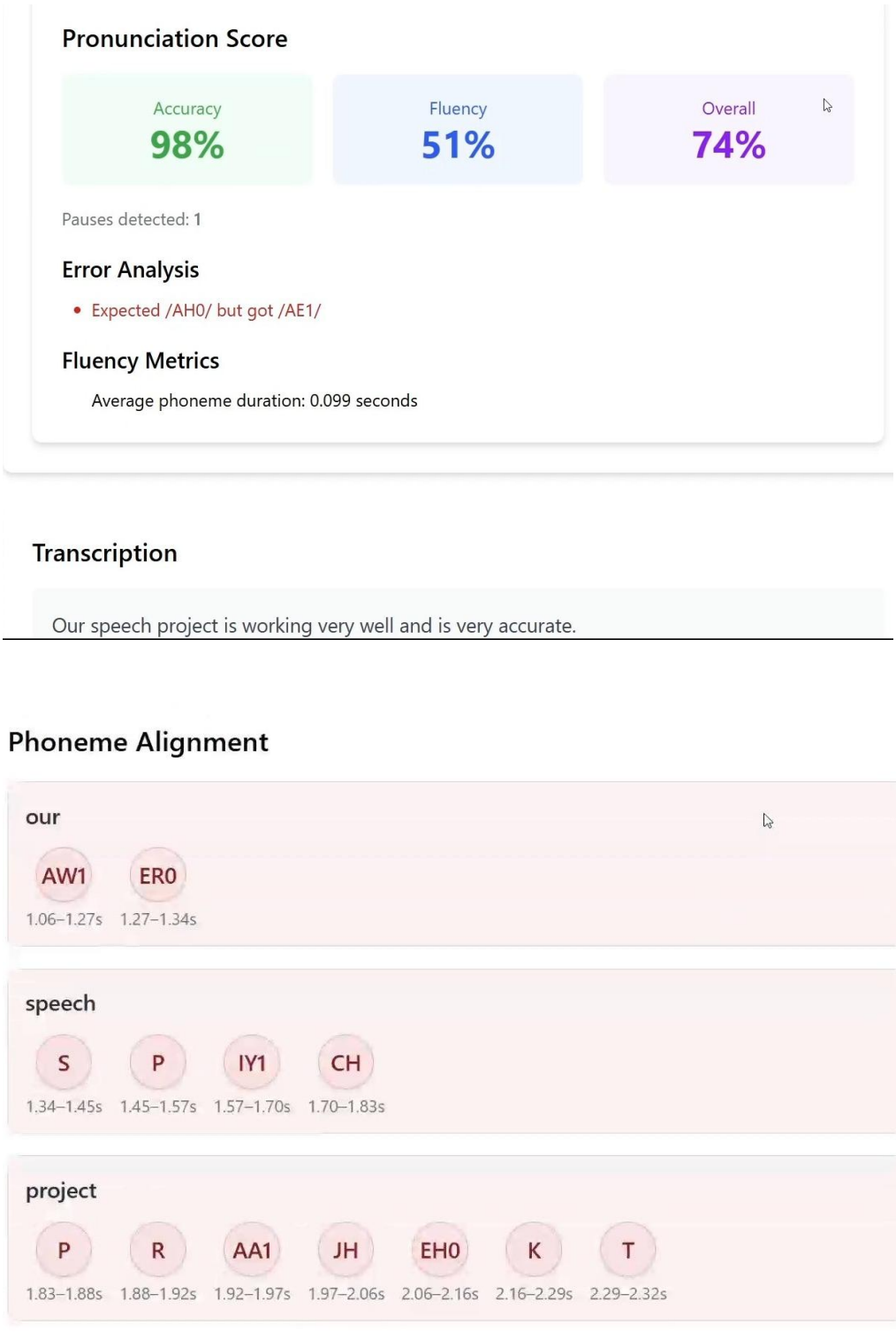
Feedback is also shown in terms of proper phoneme comparison. It displays what was expected and what was spoken.

Frontend and Backend Integration:

To provide a seamless user experience, we implemented a simple web-based interface where users can record and submit audio and receive feedback.

- The frontend was developed using Vite + React, offering a fast and an efficient development environment.
- Tailwind CSS is used to style the UI, providing a modern, responsive, and consistent interface.
- Users can record audio or upload a file, and provide a reference text for pronunciation evaluation.
- Upon submission, the frontend sends the audio and reference text to the backend using an HTTP POST request.
- Integration between frontend and backend is handled via RESTful API endpoints exposed by FastAPI. The React frontend uses Axios to interact with these endpoints asynchronously.
- The backend is built with FastAPI which receives the request and processes the input by performing audio preprocessing (such as resampling) and then running a speech evaluation model to compare the spoken audio against the reference text.
- The backend then returns a JSON response containing the overall pronunciation score along with detailed phoneme-level feedback.
- Then the frontend receives this data and dynamically renders the results using Tailwind-styled components.

RESULTS


Pronunciation Score
Accuracy
98%
Fluency
51%
Overall
74%
Pauses detected: 1
Error Analysis
Expected /AH0/ but got /AE1/
Fluency Metrics
Average phoneme duration: 0.099 seconds
Transcription
Our speech project is working very well and is very accurate.
Phoneme Alignment
our
AW1
1.06–1.27s
ER0
1.27–1.34s
speech
S
1.34–1.45s
P
1.45–1.57s
IY1
1.57–1.70s
CH
1.70–1.83s
project
P
1.83–1.88s
R
1.88–1.92s
AA1
1.92–1.97s
JH
1.97–2.06s
EH0
2.06–2.16s
K
2.16–2.29s
T
2.29–2.32s

is

IH1	Z
2.35–2.44s	2.44–2.61s

working

W	ER1	K	IH0	NG
2.61–2.70s	2.70–2.78s	2.78–2.84s	2.84–2.90s	2.90–3.17s

very

V	EH1	R	IY0
3.17–3.30s	3.30–3.38s	3.38–3.42s	3.42–3.55s

well

W	EH1	L
3.55–3.63s	3.63–3.68s	3.68–3.81s

and

AE1	N	D
4.16–4.37s	4.37–4.42s	4.42–4.45s

is

IH1	Z
4.45–4.66s	4.66–4.74s

very

V	EH1	R	IY0
4.74–4.85s	4.85–4.94s	4.94–4.98s	4.98–5.13s

accurate

AE1	K	Y	ER0	AH0	T
5.20–5.27s	5.27–5.40s	5.40–5.44s	5.44–5.58s	5.58–5.61s	5.61–5.65s

CONCLUSION

The Speech Pronunciation Analyzer project is shown to effectively utilize deep learning methods to analyze spoken language and offer informative pronunciation feedback. With the use of self-supervised speech models such as Whisper, we were able to develop a robust system that could compare user speech against reference pronunciations and hence, return a quantitative measure based on similarity. The system successfully identified pronunciation mistakes among n speakers, including improper articulation of sounds such as "th", "w", and "r". Using exact-phoneme-level matches offered an easy yet efficient method for scoring pronunciation quality. The analyzer can be easily extended to accommodate different languages, dialects, and phoneme-level feedback, and it can be a useful resource for educators, learners, and speech therapists alike.