International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

A Real-Time Speech Analysis System for Early Literacy Development using AI-Based Transcription and Levenshtein Similarity Evaluation

Manish Madhava Tripathi[1],Samad Rizvi[2], Sanskar Mishra[3]

Department of Computer Science and Engineering, Integral University, Lucknow, India

Abstract. Early literacy acquisition is a decisive factor in the future academic achievement, but the traditional pedagogical approach often lacks the means of giving instant feedback to an individual learner in a personalized and timely manner. The given paper introduces a Real-Time Speech Analysis System (RTSAS) of early literacy development, which is to assess and improve reading fluency and pronunciation accuracy of small learners by means of an interactive, AI-driven platform. The system suggested uses an audio pipeline, a state-of-the-art automatic speech recognition (ASR) model (OpenAI Whisper through Groq API), and text similarity evaluation buffer founded on Levenshtein edit distance to evaluate the correctness of pronunciation. Fluency in reading is also measured in a Words-Per-Minute (WPM) measure. The learner is guided by a multi-slide graphical user interface (GUI) created in Python Tkinter, which will take the learner through four phases of interaction, which include text input, speech recording, transcription, and display of results. The system is shown through experimental testing to be able to achieve a mean pronunciation accuracy of 87.3% and an average latency of 1.4 seconds per session on a group of 40 early learners, confirming that it is appropriate to be used in the real world. The system provides an objective, engaging and scalable alternative to manual assessment, and has extensive use in primary education, computer- assisted language learning (CALL) and speech therapy.

Keywords: Speech recognition · early literacy acquisition · levenshtein distance · pronunciation recognition · reading fluency · computer-aided language learning · automatic speech recognition.

1 INTRODUCTION

Ability to read fluently and pronounce the words correctly is the foundation of academic and cognitive growth of a child. The National Reading Panel identified phonemic awareness as one of the five pillars of reading proficiency, phonics, fluency, vocabulary, and comprehension. However, even after decades of progress in pedagogy, traditional classroom learning still has a problem of delivering individualized and real-time corrective feedback, especially in learning settings, which have limited resources and teacher to student ratios are still high.

The growth of natural language processing (NLP) and artificial intelligence (AI) technologies has triggered an educational technology paradigm shift. ASR systems which were initially limited to laboratory environments are now robust enough and available enough to drive consumer-facing applications. Specifically, ASR models based on transformers like OpenAI Whisper have shown near-human performance on multilingual transcription tests, allowing them to be used as interactive learning tools.

In this paper, a Real-Time Speech Analysis System (RTSAS) a desktop-based application is introduced, which uses AI-driven ASR and algorithmic text comparison to provide real- time, objective feedback on pronunciation quality and reading rate. The main contributions of this work are as follows:

- An end-to-end, modular pipeline containing audio capture, cloud-based ASR transcription, and similarity scoring based on the Levenshtein distance in a single interactive platform.
- A two-metric performance assessment system that involves pronunciation accuracy (%) and reading fluency (WPM) to give a holistic literacy measure.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

- Empirical assessment of system performance that was carried out on 40 early learners, showing high accuracy, low latency, and high user engagement.
- An open-source implementation written in Python with Tkinter, sounddevice, SciPy and the Groq Whisper API, and is reproducible.

The rest of this paper will be formatted in the following way. Section 2 is a review of related work in speech based educational systems. Part 3 outlines the system design and approach. The experimental setup and results are introduced in section 4. Section 5 explains the uses and restrictiveness and lastly Section 6 concludes with future research directions.

2 RELATED WORK

The cross-over between speech recognition and educational technology has been a subject
of continuous research. We conduct a literature review of the most pertinent previous research in three thematic areas: ASR in education, text similarity algorithms, and interactive literacy systems.2.1.1 Speech recognitions in education.

2.1 Automatic Speech Recognition in Education

The topic of automatic speech recognition has been widely investigated as part of Computer-Assisted Language Learning (CALL). The extensive survey by Neri et al. [1] showed that pronunciation training with ASR is an effective way to enhance the articulation of second-language learners. Later on, deep learning structures have replaced Hidden Markov Model (HMM)-based systems. Whisper [2] is a huge (weakly) supervised ASR model trained on 680,000 hours of multilingual data, and has state-of-the-art performance in common benchmarks. Tommerdahl et al. [3] conducted a review of several ASR applications in learning institutions and observed that one of the most effective affords of speech-based learning tools is real-time feedback.2.2 Text Comparison and Pronunciation Test.

2.2 Text Similarity and Pronunciation Evaluation

String similarity measures are a key parameter in the assessment of similarity between predicted and transcribed text. The Levenshtein distance [4] is defined as the smallest set
of single-character editing operations insertions, deletions and replacements that are needed to convert one string into another, and it is a principled metric of phonological distance. Jurafsky and Martin [5] place edit distance in context of NLP evaluation, and
observe that it is widely used in spell-checking, optical character recognition, and speech

assessment. Li et al. [6] showed that Levenshtein-based scoring is strongly correlated to expert human pronunciation ratings ($r = 0.84$), which confirms that it is a useful automated proxy of pronunciation quality.

2.3 Interactive Literacy Systems

A number of commercial and academic systems have tackled the issue of early literacy using interactive technology. Zhang et al. [7] did suggest a speech-based reading tutor among elementary school students and did attain important improvements in oral reading fluency in a six-week intervention. Wren and Iverson [8] compared a Phonics instruction system using interactive feedback and tablets and discovered that the learners who used the interactive feedback outperformed the controls by 18 percent in standardized tests. Despite these developments, the majority of current systems are proprietary, platform- based, or single-language deployment based. The proposed RTSAS provides these gaps by providing an open, extensible, and language-neutral platform that integrates ASR transcription, similarity scoring, and WPM measurement into a single interface.3 System design and Methodology.

3 SYSTEM DESIGN AND METHODOLOGY

The RTSAS is designed as a modular pipeline with seven loosely connected components as shown in Fig. 1. The design philosophy is focused on simplicity of interaction to the early learner with an eye to extensibility later on.

3.1 System Architecture

The general architecture is a sequential dataflow design: text input to audio acquisition to ASR transcription to

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

text comparison to metric calculation to feedback generation to result display. Each step is provided as a separate module with clear interfaces, allowing replacing single parts without altering the overall pipeline. It is written in Python 3.10 and tested on Windows 10/11 and Ubuntu 22.04.

Module	Technology / Library	Function
GUI	Python Tkinter	Multi-slide interactive interface
Audio Acquisition	sounddevice, SciPy	PCM recording at 44.1 kHz, WAV export
ASR Transcription	Groq Whisper API	Speech-to-text conversion
Text Comparison	python-Levenshtein	Similarity ratio computation
WPM Computation	Native Python	Reading fluency measurement
Feedback Engine	Rule-based classifier	Grade assignment and suggestions
Error Handler	Python try/except	API and recording fault tolerance

Table 1. System module summary.

3.2 Audio Acquisition Module

The sounddevice library is used to capture voice input at a rate of f s = 44,100 Hz with a constant recording window of T = 5 seconds. The 44.1 kHz sampling rate meets the Nyquist Shannon sampling theory of speech samples with spectral content that generally does not extend past 8 kHz that guarantees a lossless digital representation of the sample. The PCM stream captured is then serialize to WAV format with scipy.io.wavfile and saved into a temporary file to be later processed.

3.3 ASR Transcription Module

The WAV file is sent to the Groq cloud API which contains the Whisper large-v3 model. Whisper uses a transformer encoderdecoder network that is trained using weakly supervised multilingual data. Taking an audio input X of length T, the model generates a textual hypothesis H = argmax P(W | X) through beam search decoding with a beam width
of 5. The transcription is returned (a UTF-8 JSON payload) by the API, and extracted and normalized (lowercased, punctuation-stripped) before similarity is computed.3.4 Comparison and Accuracy Computation of the text.

3.4 Text Comparison and Accuracy Computation

The Levenshtein similarity ratio is used to measure pronunciation accuracy, and is defined as:

$$Sim(R, H) = 1 - lev(R, H) / max(|R|, |H|)$$

R is the reference text (user input), H is the ASR hypothesis, lev (R, H) is the Levenshtein edit distance, and |l| is the length of the string in characters. The ratio of similarity lies in [0, 1] and is transformed into a percentage accuracy score A = 100× Sim (R, H).3.5 Reading Fluency (WPM) Module Performance is categorized as: Excellent (A ≥ 90%), Good (75%≤ A < 90%), Needs Practice (50%≤ A < 75%), and Requires Significant Improvement (A < 50%).

3.5 Reading Fluency (WPM) Module

Reading fluency is gauged through Words Per Minute (WPM) which is calculated as:

$$WPM = (N\_w / T) \times 60$$

Nw is the size of the reference text in number of words and T is the time of recording in seconds. Fluency categories are defined as: Slow (WPM < 60), Adequate (60≤ WPM < 100), Good (100≤ WPM < 150), and Fast (WPM≥ 150).

3.6 Feedback Generation Module

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

The feedback engine is a rule-based classifier on the (A, WPM) tuples, which produces natural-language performance messages. Motivating feedback (e.g., "Great pronunciation! Keep it up.) is accompanied by specific improvement recommendations (e.g., "Read a little slower to better understand it.). This design is consistent with known concepts of formative assessment and positive reinforcement of education psychology.

4 EXPERIMENTAL EVALUATION

4.1 Experimental Setup

It was tested on a group of 40 early learners (5 to 9 years old) who were selected in two primary schools in Lucknow, India. The participants were separated into two groups:

Group A (n = 20) carried out the four-week intervention (20 minutes per session, three sessions per week) with the use of the RTSAS, and Group B (n = 20) was the control group that maintained a standard classroom-based learning. Both groups were assessed
on a standardized oral reading fluency test (ORF) in pre- and post-assessment scores. Besides, system-level measurements, including transcription accuracy, average session latency, and user satisfaction (measured on a 5-point Likert scale by supervising teachers) were taken in Group A during the intervention period.

4.2 Results

Table 2 summarizes the key system performance metrics recorded over the intervention period.

Metric	Value
Mean Pronunciation Accuracy (A)	87.3% (±3.6%)
Mean WPM (Group A, post-test)	98.4 (±8.1)
Mean WPM (Group B, post-test)	79.2 (±10.3)
Average Session Latency	1.4 s (API round-trip)
ORF Score Gain (Group A)	+22.7% (vs. pre-test)
ORF Score Gain (Group B)	+9.1% (vs. pre-test)
Teacher Satisfaction (Likert 1–5)	4.3 / 5.0

Table 2. Summary of system performance and intervention outcomes.

Group A showed statistically significant increase in ORF scores compared to Group B (22.7% vs. 9.1%, $p < 0.01$, paired t-test) supporting the efficacy of real-time speech feedback in enhancing early literacy acquisition. The average post test WPM of Group A (98.4) was higher than Group B (79.2) by 24.2, which reflected better reading fluency. The mean session latency of 1.4 seconds is sufficiently below the limit(≤ 3 s) widely used in literature on human-computer interaction as the requirement to sustain an interactive flow.

In the character error rate (1 -CER) the accuracy of the Whisper model in our experimental environment was 91.2%. The major causes of error were the background noise during classroom recordings (63 percent of the transcription errors) and code- switched speech where learners unintentionally combined Hindi and English words (29 percent).

5 DISCUSSION

The RTSAS can be used in various educational and clinical settings. It may be used in primary schools as a supplementary teaching tool allowing self-paced oral reading with real-time automated feedback, which will eliminate the need of a teacher to administer
oral tests individually. In second language pronunciation training CALL systems the reference corpus and ASR language model can be replaced. The WPM and accuracy measures are used in speech therapy to quantitatively longitudinally monitor patient improvement which supplements observations of the therapists.

5.2 Limitations

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

It has a number of shortcomings that should be mentioned. First, the system relies on cloud-based ASR, which brings variability of latency and connectivity requirements, limiting its usage to offline or low-bandwidth environments. Second, the Levenshtein similarity measure operates on the character level and fails to indicate phoneme-level articulatory errors, which can lead to inaccurately high semantically incorrect but
orthographically similar score. Third, the predetermined 5-second recording window might not be flexible enough to allow learners with much slower reading rates. Fourth, the existing implementation is inherently English-only, which restricts the extrapolation to the multilingual educational settings, which are common in India and other language- diverse areas.6 Conclusion and Future Work.

6 CONCLUSION AND FUTURE WORK

This paper introduced the Real-Time Speech Analysis System (RTSAS) a Python-based interactive early literacy system that combines artificial intelligence-driven ASR with Levenshtein-based pronunciation scoring and WPM measurement to provide real-time personalized feedback to young students. The study of a cohort of 40 primary school students found that oral reading fluency scores improved by a significant margin of 22.7% after a four-week intervention, compared to a control group of improvement by 9.1%. The system demonstrated a mean pronunciation accuracy of 87.3% and an average session latency of 1.4 seconds, indicating that the system is viable in real-world deployment in classrooms.

The identified limitations will be tackled in future work in three key directions. To begin with, we intend to incorporate an offline ASR model (e.g., Whisper.cpp or Vosk) in order to remove the dependency on connectivity and minimize the latency. Second, we will add the phoneme level analysis with the help of a forced-alignment toolkit (e.g., Montreal Forced Aligner) to allow more precise pronunciation feedback. Third, we will add to the system features of adaptive content selection, Hindi and Urdu support as well as a gamification layer with badges and progress tracking to maintain long-term engagement with learners.

REFERENCES

[1] Neri, A., Cucchiarini, C., Strik, H., Boves, L.: The pedagogy-technology interface in Computer-Assisted Pronunciation Training. Comput. Assist. Lang. Learn. 15(5), 441–467 (2002)
[2] Radford, A., Kim, J.W., Xu, T., Brockman, G., McLeavey, C., Sutskever, I.: Robust Speech Recognition via Large-Scale Weak Supervision. In: ICML 2023. PMLR (2023)
[3] Tommerdahl, J., et al.: Applications of Speech Recognition in Education: A Review. Int. J. Educ. Technol. 18(3), 55–78 (2021)
[4] Levenshtein, V.I.: Binary Codes Capable of Correcting Deletions, Insertions, and Reversals. Sov. Phys. Doklady. 10(8), 707–710 (1966)
[5] Jurafsky, D., Martin, J.H.: Speech and Language Processing, 3rd edn. Pearson, Hoboken (2020)
[6] Li, X., Qian, Y., Liu, J., Xu, H.: Deep Learning for Speech Recognition: A Review. IEEE Trans. Neural Netw. Learn. Syst. 32(6), 2397–2412 (2020)
[7] Zhang, C., et al.: Speech Recognition Technology and Its Applications in Education. Int. J. Comput. Sci. Inf. Technol. 14(2), 33–47 (2022)
[8] Wren, S., Iverson, G.: Tablet-Based Phonics Instruction and Oral Reading Fluency in Elementary School Children. Read. Res. Q. 57(1), 45–62 (2022)
[9] Virtanen, P., et al.: SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. Nat. Methods 17, 261–272 (2020)
[10] Young, S., Evermann, G., Gales, M., et al.: The HTK Book (for Hidden Markov Models). Cambridge University Engineering Department (2006)
[11] Shi, W., et al.: Edge Computing: Vision and Challenges. IEEE Internet Things J. 3(5), 637–646 (2016)
[12] Rabiner, L., Juang, B.H.: Fundamentals of Speech Recognition. Prentice Hall, Hoboken (1993)