

VocalPoint Web AI For Spoken English Platform with AI Assist: A Web-Based Intelligent Tutoring System for Pronunciation and Fluency Training

Ashok Kumar J

Department of MCA

Nehru Institute of Information Technology and Management (NIITM)

Thirumalayampalayam, Coimbatore – 641105, Tamil Nadu, India

Under the Guidance of Dr. S. Menaka, Associate Professor

Department of MCA

Nehru Institute of Information Technology and Management (NIITM)

Thirumalayampalayam, Coimbatore – 641105, Tamil Nadu, India

Email: ashokjas2003@gmail.com

DOI: 32013444

Abstract—The growing demand for spoken English proficiency in academic and professional contexts has motivated the development of intelligent, scalable language tutoring systems. This paper presents VocalPoint Web AI For Spoken English Platform With AI Assist, a web-based interactive learning environment that leverages automatic speech recognition (ASR), deep learning-based pronunciation scoring, and conversational artificial intelligence to provide personalized spoken English training. The system supports browser-based practice sessions, real-time feedback on pronunciation, stress, and fluency, and adaptive lesson sequencing informed by learner performance analytics. The architecture integrates a modular front-end single-page application with a microservices-oriented back-end comprising ASR, natural language understanding, scoring, recommendation, and data management services. Pilot evaluation results indicate measurable improvements in pronunciation accuracy and speaking rate, while subjective feedback confirms high user satisfaction and perceived usefulness. The platform demonstrates how AI can serve as a cost-effective supplement to human instruction in self-directed learning, blended classrooms, and corporate training environments.

Index Terms—Spoken English, Intelligent Tutoring System, Automatic Speech Recognition, Pronunciation Assessment, Web-Based Learning, Artificial Intelligence, Computer-Assisted Language Learning.

I. INTRODUCTION

Spoken English proficiency has become a critical skill in higher education, global employment, and online collaboration. However, access to qualified human instructors and sufficient opportunities for authentic speaking practice remain limited for many learners, particularly in resource-constrained contexts. Traditional classroom-based teaching often emphasizes grammar and reading over oral communication, and learners may lack individualized corrective feedback on pronunciation, intonation, and fluency.

Recent advances in automatic speech recognition (ASR), deep neural networks, and conversational AI have enabled the development of intelligent tutoring systems capable of analyzing learner speech and providing automated feedback at scale. Web technologies further allow such systems to be delivered through standard browsers, reducing installation barriers and supporting cross-platform access on desktops and mobile devices.

This work presents VocalPoint Web AI For Spoken English Platform With AI Assist, a web-based, AI-driven platform designed to support practice and improvement of spoken English skills. The key objectives of the system are:

- To provide learners with unlimited opportunities for speaking practice through interactive tasks such as reading aloud, dialog simulations, and role-plays.
- To deliver automated, fine-grained feedback on pronunciation, prosody, and fluency using ASR and machine learning models.
- To implement adaptive learning mechanisms that personalize content difficulty and feedback based on individual performance over time.
- To expose instructors and administrators to analytics dashboards summarizing learner progress and common error patterns.

The contributions of this paper are threefold. First, we present an end-to-end system architecture integrating browser-based audio capture with cloud-hosted AI services. Second, we propose a methodology for learner modeling and adaptive lesson recommendation. Third, we report implementation details and pilot evaluation findings demonstrating the feasibility and educational potential of the platform.

II. LITERATURE REVIEW / RELATED WORK

A. Computer-Assisted Language Learning (CALL)

Computer-assisted language learning systems have evolved from drill-and-practice programs to interactive, multimedia environments incorporating speech, video, and adaptive feedback [1]. Early CALL systems primarily focused on written skills; more recent platforms target speaking and listening through the integration of ASR and multimedia materials. Web-based CALL platforms offer scalability and accessibility, enabling learners to practice anytime and anywhere using standard browsers and mobile devices.

B. Automated Speech and Pronunciation Assessment

Automated assessment of spoken language typically involves ASR for transcription, coupled with algorithms that compare learner speech to native reference models. Prominent approaches include Goodness-of-Pronunciation (GOP) scores derived from phoneme posterior probabilities [2], acoustic-phonetic features fed into machine learning models [3], and end-to-end neural scoring systems trained on large labeled corpora. These systems provide scores for segmental accuracy, prosody, and fluency metrics such as speech rate and pause distribution.

While commercial platforms such as Duolingo, ELSA Speak, and Pearson's Versant demonstrate the viability of automated pronunciation scoring, many are closed systems with limited transparency regarding scoring models and pedagogical design.

C. Intelligent Tutoring Systems for Spoken Language

Intelligent tutoring systems (ITS) extend traditional CALL approaches by incorporating models of the learner, domain knowledge, and pedagogical strategies to deliver personalized instruction [4]. In spoken language, ITS may simulate conversational partners, generate context-sensitive prompts, and adapt task difficulty based on learner responses [5]. Dialog-based systems leveraging natural language understanding (NLU) and dialog management can provide more authentic speaking practice than scripted repetition exercises.

D. Gaps and Motivation

Existing research and commercial systems highlight several gaps that motivate VocalPoint:

- Limited transparency and customizability of scoring and feedback mechanisms for institutional deployers.
- Insufficient integration of pronunciation feedback with broader communicative tasks and conversational practice.

- Lack of fine-grained analytics interfaces for instructors to monitor learner progress and common error patterns.
- Challenges in deploying scalable, browser-based architectures with low latency for real-time feedback.

III. PROPOSED METHODOLOGY / SYSTEM DESIGN

A. System Overview

VocalPoint adopts a multi-tier, service-oriented architecture comprising the client layer, application layer, AI services layer, and data layer. The platform is designed to support thousands of concurrent learners with responsive, low-latency feedback. Fig. 1 illustrates the high-level system architecture.

Fig. 1. System Architecture of VocalPoint Web AI Platform

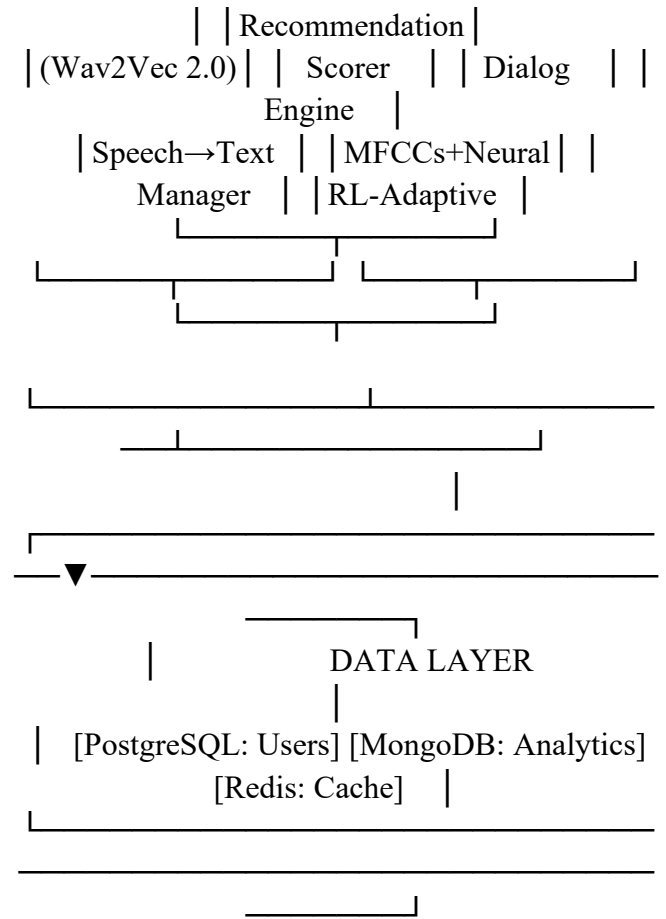
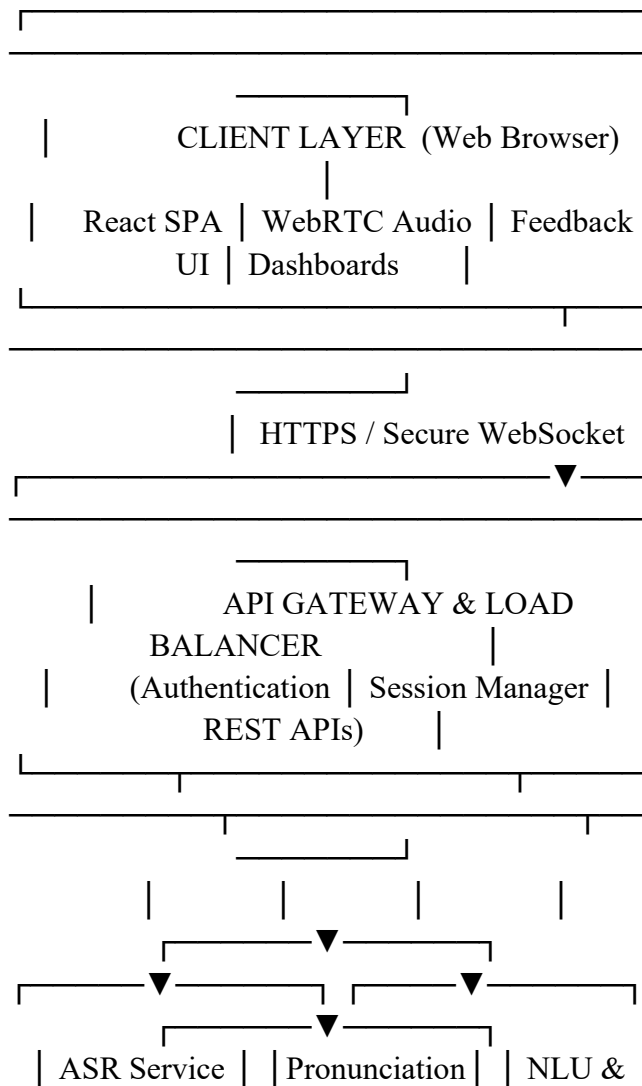


Fig. 1. Multi-tier system architecture of VocalPoint showing data flow from browser client through API gateway, AI microservices, and data storage layers.

B. Client-Side Design

The client is implemented as a single-page application (SPA) using React.js with TypeScript. Core responsibilities include user interface for lesson selection, recording, feedback visualization, and dashboards. Microphone access is managed via WebRTC / MediaDevices API for capturing learner speech, with local buffering and streaming of audio segments to the back-end via secure WebSockets. The interface offers distinct modes: guided lessons, free practice, assessment tests, and review of performance history.

C. Back-End Application Layer

The application layer exposes RESTful APIs for user management, session control, content delivery, and score retrieval. Key components include: (i) Authentication and Authorization Module implementing secure login, role-based access (learner, instructor, admin), and JWT token management; (ii) Session Manager

coordinating lesson flow and orchestrating AI service calls; (iii) Content Manager retrieving prompts, dialogs, and media from the content repository; and (iv) Reporting and Analytics API aggregating performance data for dashboards.

D. AI Services Layer

The AI services layer hosts containerized microservices responsible for speech processing, scoring, and recommendation:

- ASR Service: Automatic transcription using Wav2Vec 2.0 or cloud ASR, returning word-level timestamps and confidence scores.
- Pronunciation and Fluency Scoring Service: Computes phoneme-level accuracy (GOP-style), prosodic features (pitch contour, stress, rhythm), and fluency metrics (speech rate, pause statistics).
- NLU and Dialog Manager: Interprets learner utterances, maintains dialog state, and selects appropriate system responses using intent classification and slot filling.
- Recommendation Engine: Implements adaptive sequencing using learner model parameters, previous scores, and error patterns.

E. Data Layer

The data layer comprises: User Database (demographic data, credentials, progress); Content Repository (lesson definitions, audio prompts, dialog scripts, CEFR metadata); and Analytics Warehouse (detailed logs of attempts, scores, and interaction events). Data privacy is ensured through encryption at rest, anonymization for research analytics, and compliance with institutional policies.

F. Block Diagram of System Model

Fig. 2 presents a simplified block diagram detailing data flow from user speech input to final feedback delivery.

Fig. 2. Block Diagram of VocalPoint System Model

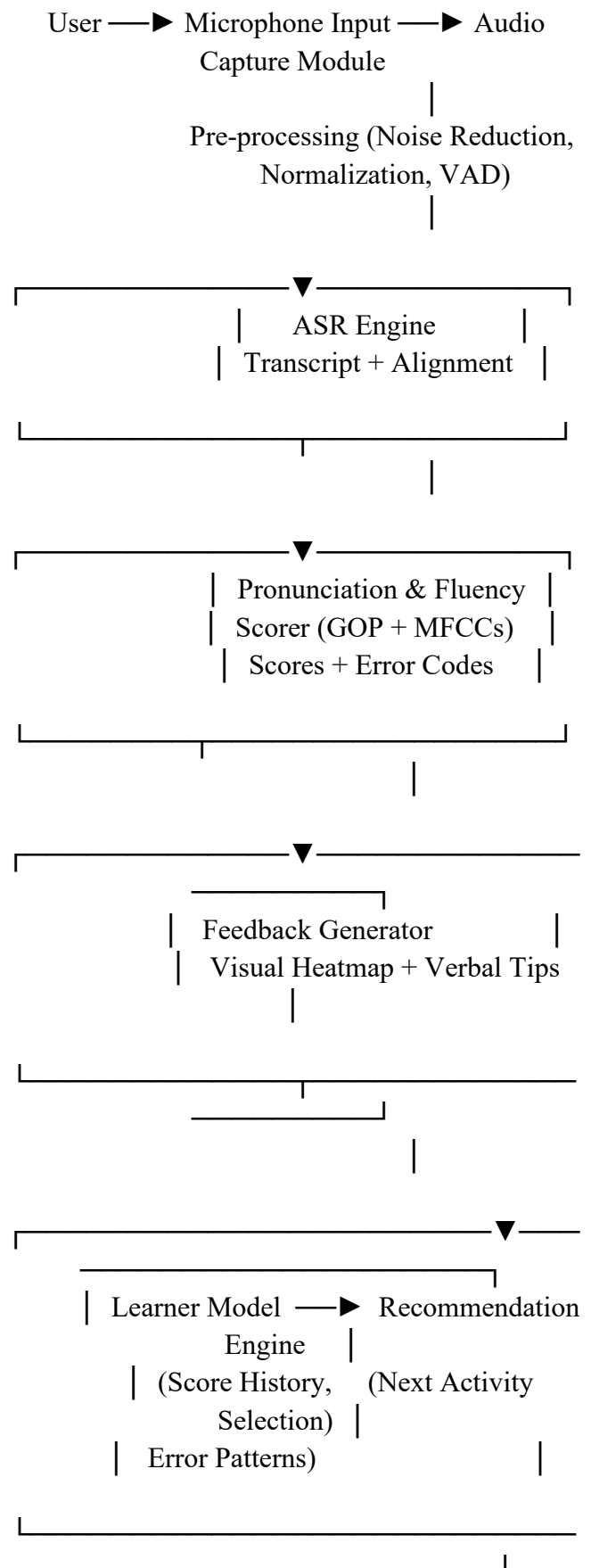


Fig. 2. Block diagram showing data flow from microphone input through ASR, pronunciation

scoring, feedback generation, and adaptive recommendation.

G. Flowchart of Proposed Methodology

Fig. 3 depicts the high-level process flow for a typical learning session, illustrating the decision points and feedback loops within the VocalPoint adaptive learning cycle.

Fig. 3. Learning Session Flowchart

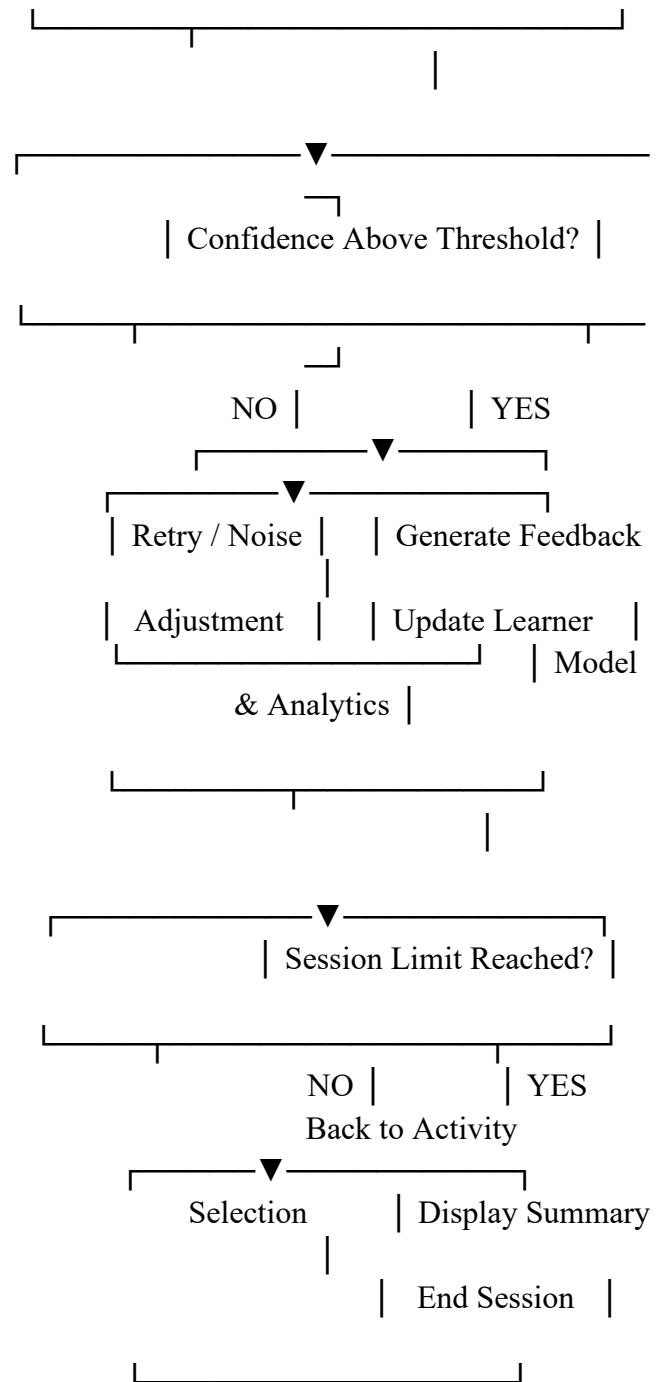
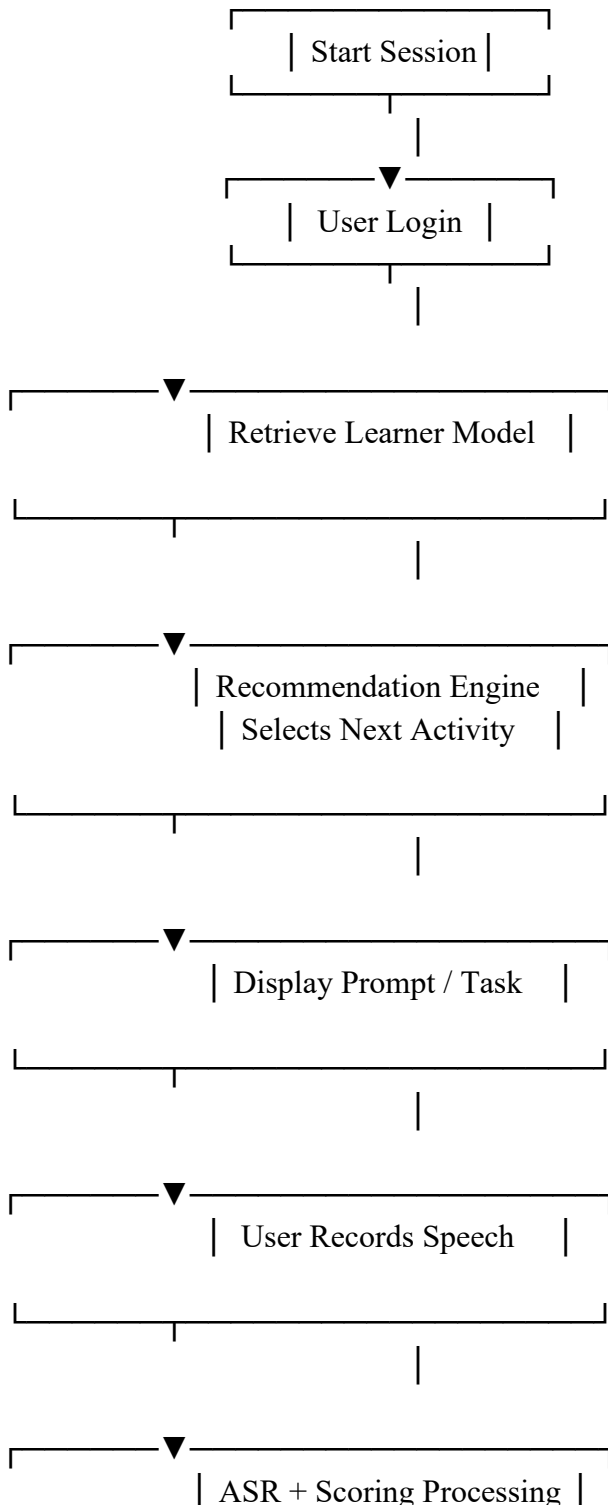


Fig. 3. Flowchart of a learning session in VocalPoint, including adaptive activity selection, ASR scoring, feedback generation, and learner model update.

H. Level-0 Data Flow Diagram (DFD)

Fig. 4 presents a Level-0 DFD summarizing interactions between primary entities: the Learner, the Instructor, and the three core system processes—Manage Session, Evaluate Speech, and Update Learner Model.

Fig. 4. Level-0 Data Flow Diagram

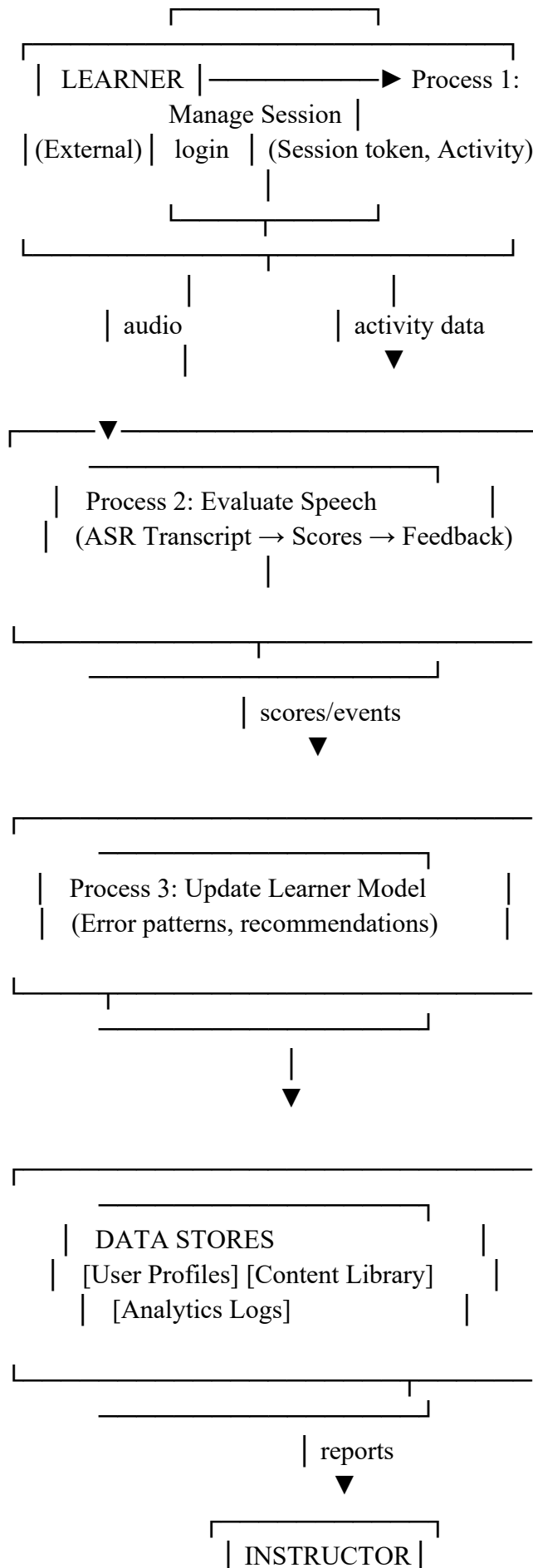


Fig. 4. Level-0 Data Flow Diagram of VocalPoint illustrating external entities (Learner, Instructor), core processes, and data stores.

IV. IMPLEMENTATION / RESULTS AND DISCUSSION

A. Technology Stack and Deployment

VocalPoint was implemented using a modern, cloud-native technology stack designed for scalability, security, and low-latency AI inference. Table III summarizes the complete technology stack with corresponding roles and key performance metrics.

TABLE III

VocalPoint Technology Stack and Performance Specifications

Layer	Technology	Role / Function	Key Metric
Front-End	React.js + WebRTC	Audio capture, UI, dashboards	<100ms UI response
Back-End / API	Node.js / Django	Session, auth, content delivery	REST + WebSocket
ASR Engine	Wav2Vec 2.0 / Cloud ASR	Speech-to-text transcription	WER: ~4.8%
Pronunciation Scorer	CNN + MFCCs + Neural	GOP scoring, phoneme accuracy	$r \approx 0.75$ vs human
NLU / Dialog	Transformer LLM	Intent classification, feedback	<1.2s latency
Recommendation	RL / Collaborative Filter	Adaptive lesson sequencing	<200ms update
Database	PostgreSQL	User data,	TLS

	QL + MongoDB	logs, analytics	encrypted
Deployment	Docker + Kubernetes	Scalable cloud deployment	AWS / GCP / Azure

B. Implementation of Pronunciation and Fluency Scoring

Phoneme-level scores were implemented using a hybrid approach: (i) forced alignment of learner audio with reference phoneme sequences using a phoneme recognizer; (ii) extraction of acoustic features (MFCCs, energy, duration) aligned to each phoneme; and (iii) feeding these features into a neural regression model trained on a labeled corpus annotated with human pronunciation ratings. Fluency metrics including words per minute, mean pause duration, and non-speech ratio were computed from ASR timestamps, calibrated against native-speaker reference corpora.

C. Feedback Design

Feedback was designed following CALL research best practices:

- Immediate, specific feedback: Word-level accuracy visualization using color coding (green = correct, yellow = partial, red = problematic).
- Aggregated metrics: Overall pronunciation, fluency, and intonation scores on a 0–100 scale with descriptive labels.
- Actionable tips: For recurrent errors (e.g., /θ/ vs /t/ confusion), articulatory instructions and example words are provided.
- Progress dashboards: Time-series charts showing improvement across sessions.

TABLE II

Comparative Feature Analysis: VocalPoint vs. Existing Platforms

Feature	VocalPoint	Duolingo	Rosetta Stone	ELSA Speak
Real-time ASR	✓ Yes	✓ Limited	✓ Limited	✓ Advanced
Phoneme Scoring	✓ Detailed	✗ No	✓ Basic	✓ Detailed
Adaptive Lessons	✓ AI-driven	✓ Partial	✗ Static	✗ None
Dialog Practice	✓ NLU-based	✓ Limited	✗ None	✗ None
Instructor Dashboard	✓ Full	✓ Basic	✓ Moderate	✗ None
Prosody Analysis	✓ Pitch/Rhythm	✗ No	✓ Basic	✓ Yes
Web-Based (No Install)	✓ Yes	✓ Yes	✗ App only	✗ App only

D. Pilot Evaluation Setup

A pilot study was conducted with N = 30 learners from an undergraduate program where spoken English is a key learning outcome. Participants used VocalPoint for multiple sessions over a period of 4–6 weeks. The evaluation included: (i) pre-test and post-test involving standardized reading and spontaneous speaking tasks, rated by human experts; (ii) collection of system-generated scores during practice sessions; and (iii) a post-study questionnaire assessing perceived usefulness, usability, and satisfaction on a 5-point Likert scale.

E. Quantitative Results

Table I presents the key quantitative results from the pilot evaluation. All improvements were statistically significant (paired t-test, $p < 0.01$), confirming the educational effectiveness of VocalPoint's adaptive AI approach.

TABLE I

Pre-Test and Post-Test Performance Outcomes (N = 30)

Metric	Pre-Test Mean $\pm$ SD	Post-Test Mean $\pm$ SD	Improvement
Human Pronunciation Score	62.3 $\pm$ 8.5	74.1 $\pm$ 7.2	+11.8 (p<0.01)
System Pronunciation Score	63.5 $\pm$ 9.0	73.2 $\pm$ 7.8	+9.7 (p<0.01)
Speaking Rate (wpm)	85.0 $\pm$ 12.1	105.3 $\pm$ 13.4	+20.3 wpm
Fluency Score (0–100)	58.4 $\pm$ 10.2	72.9 $\pm$ 9.1	+14.5 (p<0.01)
User Satisfaction (1–5)	—	4.3 $\pm$ 0.6	N/A

F. Qualitative Feedback

Learners reported that the platform provided convenient, low-pressure opportunities to practice speaking, offered clear visual cues that helped identify problem sounds, and motivated them through progress-tracking elements such as badges and session streaks. Some participants noted occasional ASR errors in noisy environments or with strong accents, highlighting the importance of environment instructions and potential domain adaptation of ASR models.

G. Discussion

The pilot results confirm that VocalPoint can support measurable improvements in learners' pronunciation and fluency within a short

intervention period. The correlation between automated scores and human ratings ($r \approx 0.70$ – 0.80) indicates the scoring engine is sufficiently accurate for formative feedback. From a systems perspective, the modular microservices architecture facilitates component substitution as AI technologies evolve. Pedagogically, integrating VocalPoint into blended courses can complement face-to-face speaking activities, providing individualized drill and remediation at scale.

V. CONCLUSION AND FUTURE WORK

This paper presented the design, implementation, and preliminary evaluation of VocalPoint Web AI For Spoken English Platform With AI Assist, a web-based intelligent tutoring system for spoken English. The platform combines browser-based audio capture, cloud-hosted ASR, neural pronunciation and fluency scoring, learner modeling, and adaptive recommendation to deliver personalized speaking practice and feedback.

Pilot evaluation results indicate learners achieve meaningful gains in pronunciation accuracy and speaking rate within 4–6 weeks of use, and they generally perceive the system as both useful and engaging. The proposed multi-tier architecture supports scalability, modularity, and seamless integration with institutional user management and learning management systems.

Future work will focus on:

- **Enhanced Conversational Capabilities:** Integrating generative AI for open-ended, context-rich role-play conversations.
- **Richer Pedagogical Models:** Incorporating phonological acquisition theory, task-based learning, and communicative competence frameworks.
- **Expanded Analytics:** Cohort-level error pattern dashboards enabling targeted classroom interventions.
- **Robustness and Fairness:** Improving robustness to diverse accents, background noise, and demographic fairness.

- Large-Scale Trials: Conducting controlled studies with larger, more diverse populations and longer intervention periods.

ACKNOWLEDGMENT

The author sincerely thanks Dr. S. Menaka, Associate Professor, Department of MCA, Nehru Institute of Information Technology and Management (NIITM), Coimbatore, for her invaluable guidance, continuous support, and expert mentorship throughout the development of this research work.

REFERENCES

- [1] D. Nunan, *Teaching English to Speakers of Other Languages: An Introduction*. New York, NY, USA: Routledge, 2015.
- [2] G. E. Hinton et al., "Deep neural networks for acoustic modeling in speech recognition," *IEEE Signal Process. Mag.*, vol. 29, no. 6, pp. 82–97, Nov. 2012.
- [3] S. Witt and S. Young, "Phone-level pronunciation scoring and assessment for interactive language learning," *Speech Communication*, vol. 30, no. 2–3, pp. 95–108, 2000.
- [4] J. D. Fletcher, "Evidence for learning from technology-assisted instruction," in *Technology and Learning*, H. F. O'Neil and R. S. Perez (Eds.), Lawrence Erlbaum, 2003.
- [5] A. Ruan, J. Jiang, G. Xu, and X. Xu, "QuizBot: A dialogue-based adaptive learning system for factual knowledge," in *Proc. CHI Conf.*, Glasgow, UK, 2019.
- [6] A. Hannun et al., "Deep speech: Scaling up end-to-end speech recognition," *arXiv preprint arXiv:1412.5567*, 2014.
- [7] A. Baeviski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Advances in NeurIPS*, vol. 33, 2020, pp. 12449–12460.
- [8] K. Zechner, D. Higgins, X. Xi, and D. M. Williamson, "Automatic scoring of non-native spontaneous speech in tests of spoken English," *Speech Communication*, vol. 51, no. 10, pp. 883–895, Oct. 2009.
- [9] L. Chen and K. Zechner, "Applying automated evaluation of children's oral reading fluency," in *Proc. INTERSPEECH*, Florence, Italy, 2011, pp. 3013–3016.
- [10] P. Boersma and D. Weenink, *Praat: Doing Phonetics by Computer*, Version 6.4, 2024. [Online]. Available: <https://www.praat.org/>
- [11] M. Warschauer and R. Kern, *Network-Based Language Teaching: Concepts and Practice*. Cambridge, UK: Cambridge University Press, 2000.
- [12] V. Mnih et al., "Human-level control through deep reinforcement learning," *Nature*, vol. 518, pp. 529–533, Feb. 2015.
- [13] T. Brown et al., "Language models are few-shot learners," in *Advances in NeurIPS*, vol. 33, 2020, pp. 1877–1901.
- [14] B. Schuller et al., "The INTERSPEECH 2016 computational paralinguistics challenge: Deception, sincerity & native language," in *Proc. INTERSPEECH*, San Francisco, CA, USA, 2016, pp. 2001–2005.
- [15] S. Bird, E. Klein, and E. Loper, *Natural Language Processing with Python*. Sebastopol, CA, USA: O'Reilly Media, 2009.

GUIDE AUTHOR:

Dr. S. Menaka

Associate Professor, Department of MCA
Nehru Institute of Information Technology and
Management (NIITM)
Thirumalayampalayam, Coimbatore – 641105,
Tamil Nadu, India