



Luminya - AI-Powered Conversational Language Learning Systems with Interactive 3D Virtual Characters

Soham Mestry^{a,*}, Swagat Kochrekar^a, Shubham Maldikar^a, Prof. Vaishali Yeole^b

^a Department of Computer Engineering, Smt. Indira Gandhi College of Engineering, India

^b Mentor, Department of Computer Engineering, Smt. Indira Gandhi College of Engineering, India

ABSTRACT :

While digital language education has expanded rapidly, bridging the gap between passive vocabulary acquisition and fluent spoken conversation remains a persistent pedagogical challenge. Recent developments in deep learning and web graphics make it possible to pair spoken dialogue systems with real-time, embodied virtual tutors directly in the browser. In this survey, we critically examine AI-driven conversational language learning systems, focusing on how interactive 3D virtual characters facilitate spoken dialogue practice. We synthesize progress across six foundational areas: Intelligent Computer-Assisted Language Learning (ICALL) frameworks, speech recognition and synthesis, large language models tailored for interactive tutoring, browser-based 3D avatar animation with speech-driven lip synchronization, cognitive spaced repetition scheduling, and natural language technologies for Indic languages. Our comparative analysis of ten leading academic and commercial platforms identifies an unresolved systemic deficit: current platforms either isolate conversational dialogue from visual embodiment, require cumbersome hardware, or overlook under-resourced Indic languages like Hindi and Marathi. To address these limitations, we evaluate architectural trade-offs, detail ongoing engineering hurdles—chiefly end-to-end latency and phonetic alignment—and propose a roadmap toward building accessible, browser-native conversational tutors.

Keywords: AI-Powered Language Learning, 3D Virtual Characters, Conversational AI, Speech Recognition, Adaptive Learning, Indian Languages, Lip Synchronization, Intelligent Tutoring Systems

1. Introduction

Developing true conversational fluency requires continuous, interactive speaking practice—an experience that conventional classroom settings struggle to deliver at scale because of teacher-to-student ratios and limited contact hours (Settles & Meeder, 2016). When Computer-Assisted Language Learning (CALL) emerged in the 1980s, it promised automated individualized instruction; however, early software relied almost exclusively on static grammar drills and rote exercises that offered negligible pedagogical adaptability (Kannan & Munday, 2018).

Recent breakthroughs in generative AI and browser technologies have fundamentally altered what computer-based tutoring can accomplish. Modern Large Language Models (LLMs) such as GPT-4 and LLaMA can manage open-ended, pedagogically structured dialogue without rigid scripting (Brown et al., 2020). Simultaneously, multilingual Automatic Speech Recognition (ASR) engines like Whisper transcribe spoken utterances across dozens of languages with high fidelity (Radford et al., 2023), while neural Text-to-Speech (TTS) synthesizers produce expressive speech with adjustable pacing (Tan et al., 2021). Parallel advances in WebGL and scene-graph libraries (such as Three.js and React Three Fiber) allow developers to render textured, animated 3D characters directly inside consumer web browsers without client-side installations.

Yet, modern language learning tools rarely combine these complementary components into a cohesive system. Commercial applications remain starkly compartmentalized: Duolingo blends spaced repetition with roleplay text chat but lacks any embodied 3D visual presence; Mondly offers virtual reality environments but depends on rigid decision trees and expensive headsets; general-purpose chatbots such as ChatGPT support open dialogue but lack systematic curricula, pronunciation assessment, and memory retention tracking (Kohnke et al., 2023). Most glaringly, the vast majority of commercial platforms treat Indian languages as an afterthought, offering rudimentary Hindi vocabularies while entirely ignoring regional languages such as Marathi.

Motivated by the design of Luminya—an open-source, browser-native language platform combining real-time 3D embodied avatars, voice-driven lip synchronization, LLM dialogue orchestration, and native support for Hindi and Marathi—this survey investigates how modern technological pillars can be unified into an accessible learning environment. Specifically, our objectives are fourfold: (1) evaluate the state of the art across six foundational technological disciplines; (2) benchmark ten prominent commercial and academic systems across critical pedagogical and technical capabilities; (3) identify core engineering bottlenecks; and (4) outline research directions for scalable, embodied language tutoring systems.

The paper is organized around these objectives: Section 2 traces the transition from early CALL to Intelligent Tutoring Systems. Section 3 evaluates speech recognition and neural synthesis pipelines. Section 4 examines LLM-driven conversational prompting. Section 5 details 3D avatar rendering and

phoneme lip-syncing. Section 6 assesses spaced repetition algorithms. Section 7 surveys the Indic NLP landscape. Section 8 presents a comparative analysis of existing systems, followed by key open challenges in Section 9 and concluding reflections in Section 10.

2. Background: CALL and Intelligent Tutoring Systems

2.1. Evolution of Computer-Assisted Language Learning

The history of Computer-Assisted Language Learning mirrors broader shifts in cognitive psychology and educational computing, typically categorized into three distinct eras: behaviorist (1960s–1980s), communicative (1980s–2000s), and integrative (2000s–present) (Kannan & Munday, 2018). Behaviorist CALL, influenced by Skinner's operant conditioning, treated language learning as stimulus-response habit formation through repetitive drills. Communicative CALL shifted the pedagogical emphasis toward authentic meaning and contextual understanding, encouraging exploratory tasks. Integrative CALL subsequently leveraged multimedia networks and hypermedia to embed language practice into collaborative social environments.

The advent of Intelligent CALL (ICALL) marked a critical transition from deterministic software toward systems capable of parsing learner input and diagnosing underlying misconceptions through Natural Language Processing (Alkhatlan & Kalita, 2018). Rather than marking an answer merely as right or wrong, ICALL systems construct detailed learner models. This paradigm was empirically validated by Settles et al. (2018) through Duolingo's Second Language Acquisition Modeling (SLAM) benchmark, where tracking over a thousand learners demonstrated that modeling student-specific error profiles alongside grammatical features drastically improved predictions of learner mistakes compared to static baselines.

2.2. Intelligent Tutoring Systems for Language Learning

Classic Intelligent Tutoring System (ITS) architectures rely on four interconnected modules: the domain model (representing linguistic rules and target vocabulary), the student model (tracking cognitive mastery and historical errors), the tutoring model (selecting pedagogical interventions and scaffolding), and the interface model (managing user interaction) (Alkhatlan & Kalita, 2018). Translating these components into conversational interfaces has proven especially beneficial for oral practice; for instance, Pham et al. (2018) demonstrated that deploying an intelligent conversational agent on mobile devices improved learner task engagement by 23% over conventional static software.

2.3. Gamification and Engagement

Sustaining long-term user engagement remains one of the hardest problems in self-directed digital education. Modern platforms address this through behavioral mechanics—notably streak maintenance, dynamic leaderboards, and experience point tiers—that trigger intrinsic and extrinsic motivators (Settles & Meeder, 2016). When calibrated against empirical interaction logs, these gamified feedback loops significantly mitigate learner drop-off, ensuring students practice consistently enough for cognitive spacing mechanisms to take hold.

3. Speech Technologies for Language Learning

3.1. Automatic Speech Recognition (ASR)

Robust spoken dialogue practice depends fundamentally on the accuracy of the underlying speech recognizer. The introduction of OpenAI's Whisper (Radford et al., 2023), trained on 680,000 hours of weakly supervised multilingual audio, demonstrated that a unified encoder-decoder Transformer could achieve competitive transcription across 99 languages without domain-specific tuning. In Hindi, Whisper delivers word error rates (WER) around 12.4%, offering an operational baseline for conversational interfaces.

Within language tutoring, ASR fulfills two distinct pedagogical functions: decoding conversational intent for the dialogue engine, and providing granular phonetic feedback on learner pronunciation. Decades of CALL research demonstrate that acoustic confidence measures correlate closely with human-graded oral fluency (Cucchiarini et al., 2000). Furthermore, targeted empirical evaluations reveal that automated acoustic feedback can drive a 15–20% gain in phonological precision over an eight-week training curriculum (Kim, 2006).

Code-switching presents an acute challenge in multilingual environments, particularly across urban India where speakers routinely mix Hindi, Marathi, and English within a single utterance. Recent mixture-of-experts architectures demonstrate that multilingual models can route speech frames dynamically to preserve recognition accuracy during rapid language shifts (Gaur et al., 2021). From a deployment perspective, optimized runtimes such as Faster-Whisper (utilizing CTranslate2 quantization) reduce latency by a factor of four compared to standard PyTorch implementations, enabling low-latency transcription on modest hardware.

3.2. Text-to-Speech (TTS) Synthesis

Neural speech synthesis has transitioned from multi-stage cascading pipelines to fully end-to-end conditional models such as VITS (Kim, Kong, & Son, 2021). By integrating variational autoencoders with adversarial training, VITS produces expressive acoustic output directly from phoneme sequences. Open-source implementations like Piper (Hansen, 2023) execute with real-time factors (RTF) below 0.1 on standard consumer CPUs, making realistic, low-latency client-side synthesis practical without ongoing API subscription costs.

Speech pacing is a vital pedagogical parameter in second-language instruction. Novice learners rely on slowed articulation to distinguish phonetic boundaries, whereas advanced learners require natural conversational tempos to build real-world listening comprehension. Controlled studies confirm

that providing adjustable playback speeds improves novice listening comprehension by up to 32% (Cardoso, 2018). Modern synthesizers support this duration scaling without distorting fundamental frequency or introducing pitch artifacts, matching learner preferences for flexible speed controls during spoken drills (Bione et al., 2017).

3.3. Speech Pipeline for Conversational Practice

A functional spoken tutoring system links speech recognition, language reasoning, and speech synthesis into a continuous loop: audio from the learner is transcribed via ASR, evaluated and responded to by an LLM, converted to audio by TTS, and visually grounded via avatar lip-sync. Because natural conversational turn-taking requires round-trip latencies below two seconds, optimizing the interaction between Faster-Whisper, local LLM execution, and Piper TTS is crucial to prevent unnatural conversational delays.

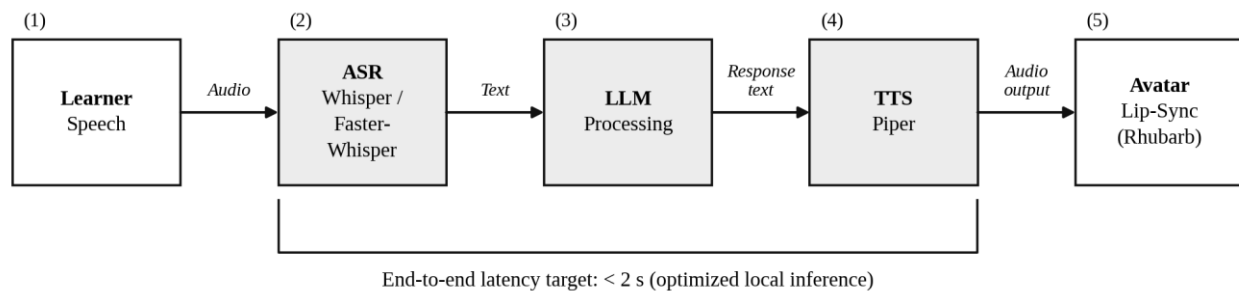


Fig. 1 – Voice & 3D Pipeline Architecture: Learner Speech → ASR → LLM → TTS → Avatar Lip-Sync.

4. Large Language Models for Conversational Language Tutoring

4.1. Foundations of LLM-Based Tutoring

The emergence of autoregressive Transformer models has rewritten the rulebook for conversational AI. Early dialogue agents depended on handcrafted finite-state graphs or brittle intent classifiers; in contrast, few-shot learners like GPT-3 and GPT-4 maintain conversational context across long turns and adapt to pedagogical instructions purely through natural language prompting (Brown et al., 2020). Open-weights models such as LLaMA (Touvron et al., 2023) have democratized this capability, allowing academic institutions and edge deployments to run locally hosted language tutors that protect student privacy and avoid cloud API fees.

4.2. LLMs in Educational Contexts

Early empirical evaluations of ChatGPT in educational settings show substantial promise for personalized error correction, contextual vocabulary explanation, and roleplay simulations (Kohnke et al., 2023). However, systematic literature reviews consistently emphasize that while generative models excel at generating motivating dialogue, their tendency to hallucinate grammatical explanations and inconsistently assess student proficiency remains a concern (Kasneci et al., 2023).

In a broad synthesis of 118 educational LLM interventions, Yan et al. (2024) noted that language acquisition generated the strongest positive learning outcomes, with 78% of examined studies reporting measurable performance improvements. To harness these capabilities responsibly, Han et al. (2024) established a pedagogical design framework emphasizing dynamic scaffolding, progressive error correction, and situated conversational tasks. Furthermore, research demonstrates that employing generative models to post-correct noisy non-native ASR transcripts significantly improves the semantic coherence of downstream dialogue (Mai & Carson-Berndsen, 2024).

4.3. Prompt Engineering for Language Tutoring

Getting reliable pedagogical behavior out of general-purpose LLMs demands systematic prompt design. Effective prompts must define four distinct elements: character persona (anchoring the agent in an authentic social context), proficiency calibration (restricting vocabulary and syntactic complexity to the learner's target CEFR tier), error handling protocols (instructing the model to offer gentle corrective recasts rather than pedantic interruptions), and environmental grounding. Moreover, prompting the LLM to return machine-readable structured JSON (containing the spoken reply, emotion tags, and viseme cues) enables straightforward synchronization with 3D avatar rendering engines.

5. 3D Virtual Characters and Lip Synchronization

5.1. Pedagogical Agents and Embodied Conversational Agents

Decades of research in human-computer interaction establish that embodied agents provide meaningful social scaffolding for learners. A meta-analysis by Johnson and Lester (2016) calculated an overall effect size of 0.36 for pedagogical agents on student learning outcomes. More specifically, in second-language speaking practice, interacting with an expressive 3D character significantly lowers Foreign Language Classroom Anxiety Scale (FLCAS) scores compared to audio-only bots, thereby increasing learners' Willingness to Communicate (WTC) (Ji & Park, 2023).

Synthesizing findings across 32 empirical evaluations, Heidig and Clarebout (2011) confirmed that visual persona cues foster learner motivation and persistence. However, agent nonverbal communication must be carefully tuned: Baylor and Kim (2009) demonstrated that while context-appropriate gestures build trust and perceived pedagogical authority, exaggerated or unmotivated animations elevate extraneous cognitive load. When designed thoughtfully, embodied agents provide social presence that bolsters long-term lexical recall and oral fluency (Huang et al., 2021).

These findings align with foundational principles of Embodied Conversational Agents (ECAs), which emphasize that speech, gaze, and facial expressions operate as an integrated communicative channel (Cassell et al., 2000). Furthermore, virtual agents capable of exhibiting basic rapport-building behaviors, such as rhythmic head nods and responsive gaze, reliably elicit longer, more sustained conversational responses from human interlocutors (Gratch et al., 2007).

5.2. 3D Avatar Technologies for Web-Based Deployment

Modern web standards allow high-fidelity 3D rendering directly within desktop and mobile browsers via WebGL, Three.js, and React Three Fiber. Production-ready pipelines typically leverage character authoring tools like Ready Player Me to deliver lightweight .glb avatar meshes, paired with skeletal animation libraries from Mixamo. As illustrated in Fig. 2, the client runtime orchestrates four key layers: importing the 3D geometry, mapping state-driven animation clips (idle, listening, speaking), blending facial morph targets to convey emotion, and compositing the character into a contextual 3D scene (e.g., a café, market, or office).

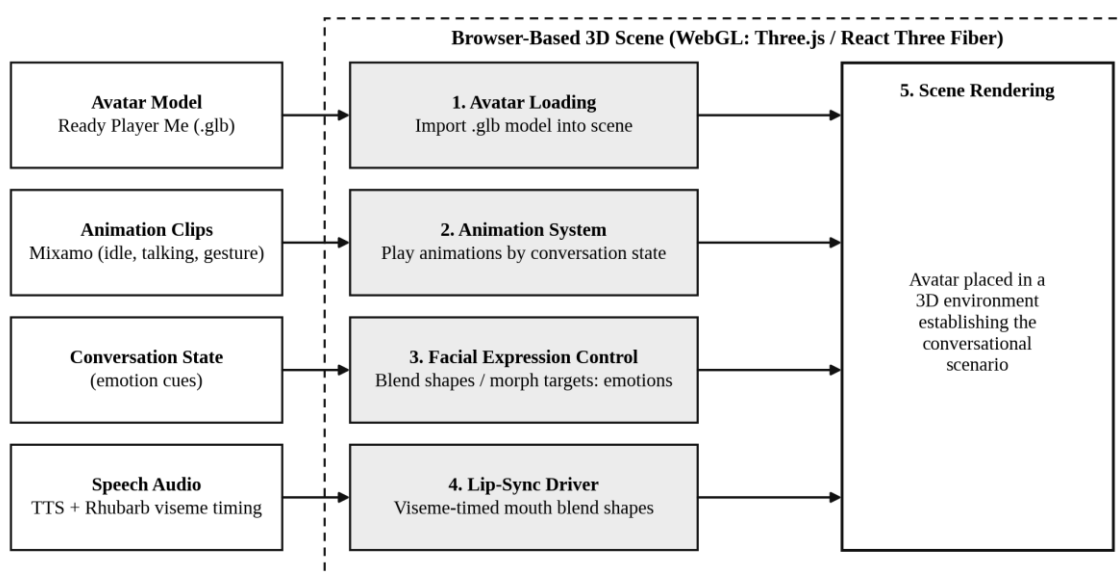


Fig. 2 – 3D Avatar Integration Architecture showing model loading, animation, lip-sync, and scene integration.

5.3. Real-Time Lip Synchronization

Convincing character interaction requires tight temporal alignment between generated speech audio and mouth geometry. Early systems mapped acoustic energy and formants directly to blendshape weights (Cao et al., 2005). Later deep learning models achieved robust 30-fps synthesis across diverse speaking styles (Taylor et al., 2017). To address complex co-articulation effects—where the mouth shape for a phoneme depends on preceding and following sounds—Edwards et al. (2016) introduced the JALI model, separating jaw and lip parameters for realistic viseme generation.

While deep end-to-end vertex prediction models like VOCA (Cudeiro et al., 2019) produce lifelike facial deformation, they require substantial GPU resources. For lightweight web deployment, phoneme-based acoustic analyzers such as Rhubarb Lip Sync (Wolf, 2023) offer a practical alternative: they parse synthesized audio into timestamped viseme cues that directly modulate standard morph targets on the avatar mesh. Because viseme timings are anchored to the generated audio waveform, lip movement remains synchronized even when the learner adjusts speech playback rates.

6. Adaptive Learning and Spaced Repetition

6.1. Spaced Repetition Algorithms

The psychological spacing effect—reviewing information at expanding intervals to consolidate memory traces—is among the most empirically validated concepts in educational science. The venerable SuperMemo SM-2 heuristic (Wozniak & Gorzelanczyk, 1994) introduced difficulty-weighted interval calculations that remain widely used in apps like Anki.

To move beyond heuristic rules, Settles and Meeder (2016) developed the Half-Life Regression (HLR) model at Duolingo, predicting recall probabilities by combining individual learner histories, word difficulty, and interaction latency. Their formulation improved recall prediction accuracy by 8.3% over SM-2. Similarly, Reddy et al. (2016) demonstrated that formulating interval scheduling as a machine learning optimization problem improved long-term retention by 17% compared to fixed-interval schedules.

More recently, the Free Spaced Repetition Scheduler (FSRS) formalized memory retention using the Difficulty, Stability, and Retrieval (DSR) framework, achieving target retention rates with 20–30% fewer reviews than SM-2 (Ye et al., 2022). Parallel theoretical work by Tabibian et al. (2019), modeling student memory via Marked Temporal Point Processes, confirms that data-driven adaptive scheduling significantly outperforms static heuristics by modeling individual forgetting curves.

6.2. Adaptive Difficulty and Personalization

Adaptive CALL systems rest upon three foundational pillars: learner modeling, intelligent content selection, and individualized feedback (Vandewaetere & Desmet, 2012). In a conversational AI context, these principles translate into dynamically adjusting conversation difficulty, introducing target vocabulary at measured intervals, and steering roleplay scenarios toward identified grammatical weaknesses based on continuous performance telemetry.

7. Natural Language Processing for Indian Languages

7.1. The Indian Language Technology Landscape

With 22 scheduled languages and hundreds of regional dialects, India represents one of the world's most complex linguistic ecosystems. In empirical surveys of NLP resource distribution, Hindi is categorized as a "rising star" with rapidly growing digital resources, whereas languages like Marathi remain classified as under-resourced despite tens of millions of native speakers (Joshi et al., 2020).

The AI4Bharat initiative has dramatically improved this landscape by releasing open foundational resources (Kakwani et al., 2020; Javed et al., 2023). Their ecosystem spans large-scale monolingual corpora (IndicNLP), multilingual masked language models (IndicBERT), acoustic models (IndicWav2Vec), expressive speech synthesizers (IndicTTS), and translation benchmarks (IndicTrans). Complementing these, the IndicNLP Library (Kunchukuttan, 2020) provides standardized text normalization, transliteration, and phonetic parsing essential for pre-processing Indic scripts.

7.2. Speech Technologies for Hindi and Marathi

Although ASR accuracy for Hindi has reached commercial viability on benchmark datasets (Radford et al., 2023), Marathi speech recognition continues to struggle with acoustic variability and limited training corpora. To mitigate these resource disparities, Javed et al. (2024) introduced IndicVoices, an open speech corpus spanning 7,300+ hours across 22 Indian languages, providing representative benchmarks for conversational Hindi and Marathi under diverse recording conditions.

On the synthesis side, neural models like Tacotron demonstrated early viability for Hindi with Mean Opinion Scores around 3.8/5 (Gupta & Gaur, 2019). Subsequent model releases, such as IndicBERT v2 (Doddapaneni et al., 2023), mitigated subword tokenization bloat in Devanagari script, enabling accurate grammatical error correction across both Hindi and Marathi.

7.3. Challenges Specific to Indian Languages

Applying NLP to Indian language education presents unique practical hurdles: complex Devanagari orthography and frequent transliteration into Roman script; rich agglutinative morphology in Marathi that causes vocabulary explosion; pervasive code-mixing (Hinglish/Maranglish) in daily discourse; and scarce annotated pedagogical datasets for intermediate grammar correction.

8. Comparative Analysis of Existing Systems

To evaluate where the field stands, we analyzed ten representative language platforms across seven critical dimensions: embodied 3D characters, adaptive AI dialogue, Hindi support, Marathi support, contextual 3D environments, and spaced repetition integration. Table 1 summarizes the comparison.

Table 1 – Comparative Analysis of Existing AI-Powered Language Learning Systems

System	3D Avatar	AI Conv.	Hindi	Marathi	3D Scenes	Spaced Rep.
Duolingo	X	✓*	✓	X	X	✓
Babbel	X	X	✓†	X	X	✓
Rosetta Stone	X	X	✓	X	X	X
Mondly	✓‡	X	✓	X	✓‡	X
ELSA Speak	X	X	X	X	X	X
ChatGPT	X	✓	✓†	✓†	X	X
Replika	✓	X	X	X	X	X
Immerse VR	✓‡	X	X	X	✓‡	X
Academic ECAs	✓	✓	X	X	✓	X
Luminya	✓	✓	✓	✓	✓	✓

* Premium/Max subscription only † Limited support ‡ Requires VR headset

8.1. Key Observations

1. Architectural Fragmentation: None of the evaluated systems bridges all six technological dimensions. Mainstream platforms typically optimize for a single modality: Duolingo emphasizes gamified spaced repetition drills but lacks 3D environments, whereas Mondly VR provides immersive scenes but relies on pre-scripted branching dialogues with zero dynamic AI adaptivity.
2. The Indic Language Deficit: While introductory Hindi lessons appear on several mainstream apps, they remain restricted to isolated vocabulary drills without spoken conversational practice. Marathi is virtually absent across dedicated learning applications; only general LLMs like ChatGPT support Marathi interaction, though without structured pedagogy, speech scaffolding, or pronunciation scoring.
3. Hardware and Accessibility Barriers: Existing virtual reality platforms (such as Mondly VR and Immerse VR) demand dedicated head-mounted displays, creating cost and accessibility barriers that prevent broad classroom adoption. No platform currently brings interactive 3D virtual tutors into lightweight, standard web browsers.
4. Constrained Conversational Intelligence: Dynamic, LLM-powered dialogue remains confined to expensive premium tiers (e.g., Duolingo Max) or unconstrained text bots (e.g., ChatGPT) that offer neither structured skill trees nor visual embodiment.

8.2. The Identified Gap

This comparative audit reveals an evident engineering gap: no existing platform simultaneously combines (1) browser-native 3D avatars with speech-driven lip synchronization, (2) LLM-powered dialogue adaptation with real-time scaffolding, (3) curriculum-aligned spaced repetition for systematic retention, (4) bidirectional learning support for under-resourced Indian languages like Hindi and Marathi, (5) situated 3D scenario immersion, and (6) an extensible, open-source technology stack.

9. Challenges and Future Directions

9.1. Technical Challenges

1. Turn-Taking Latency: Linking ASR transcription, LLM generation, TTS audio synthesis, and viseme calculation creates cumulative latency that easily disrupts conversational cadence. Maintaining sub-two-second response times requires local model serving, streaming token generation, and concurrent audio-viseme processing.
2. Acoustic Quality for Regional Languages: While Hindi models have matured, Marathi ASR and TTS struggle with dialectal diversity, acoustic noise, and morphological complexity (Joshi et al., 2020; Javed et al., 2023).
3. Client-Side Viseme Synchronization: Calculating accurate phoneme timings and driving avatar blend shapes in real time across low-power client devices without audio drift remains an ongoing performance bottleneck.
4. Hallucination and Pedagogical Soundness: LLMs can generate grammatically flawed corrections or fabricate linguistic rules, creating pedagogical risks. Strict prompt templates, grammar verification filters, and structured schema outputs are necessary to constrain model replies (Kasneci et al., 2023).

9.2. Research Challenges

1. Longitudinal Proficiency Measurement: Assessing embodied tutoring systems requires multi-month empirical interventions measuring actual communicative competence and oral fluency, rather than short-term user satisfaction metrics.
2. Multilingual Learner Modeling: Capturing cross-linguistic transfer effects—such as how a native Marathi speaker leverages cognates when learning Hindi—demands more sophisticated cognitive modeling.

3. Pragmatic and Sociolinguistic Nuance: Conversational competence requires learning formal versus informal registers, honorifics, and cultural norms—dimensions that generic language models frequently overlook.

9.3. Future Directions

1. Multimodal Learner Telemetry: Fusing acoustic prosody, speech latency, and webcam-based attention tracking to build multidimensional models of learner confusion and confidence.
2. Multi-User Immersive Environments: Enabling shared 3D virtual spaces where multiple learners and AI agents interact concurrently in collaborative roleplays.
3. Expanding Language Portfolios: Adapting the architecture to additional Indic language families (e.g., Tamil, Telugu, Bengali) and under-resourced global dialects.
4. Affective, Emotion-Aware Tutoring: Detecting acoustic indicators of frustration or anxiety to dynamically adjust tutoring tone, scaffolding, and avatar facial demeanor.
5. Edge-Optimized Deployment: Compressing ASR, LLM, and TTS models for completely offline execution on entry-level mobile devices, ensuring accessibility in low-connectivity regions.

10. Conclusion

This survey reviewed the intersection of conversational AI, 3D graphics, speech processing, and adaptive pedagogy for digital language learning. A systematic evaluation of ten contemporary platforms underscores a stark technological divide: current systems separate visual immersion from intelligent dialogue and largely ignore Indic languages.

The building blocks required to bridge this divide are now commercially and technically mature. Whisper provides dependable multilingual speech transcription; neural synthesizers like Piper generate natural audio with low computational overhead; WebGL frameworks animate 3D avatars directly in consumer browsers; and open LLMs provide adaptive conversational reasoning. The critical opportunity lies in orchestrating these components into a unified, responsive learning platform.

Systems like Luminya illustrate how combining browser-based 3D avatars, adaptive dialogue orchestration, and native support for Hindi and Marathi can make immersive language education broadly accessible. As these foundational technologies continue to evolve, multimodal conversational tutors will play a central role in breaking down accessibility and language barriers worldwide.

Acknowledgements

The authors gratefully acknowledge the open-source communities behind Whisper, Piper TTS, Rhubarb Lip Sync, React Three Fiber, Ready Player Me, and the AI4Bharat initiative, whose foundational contributions have made the development of systems like Luminya possible.

REFERENCES

- Alkhatlan, A., & Kalita, J. (2018). Intelligent Tutoring Systems: A Comprehensive Historical Survey with Recent Developments. *International Journal of Computer Applications*, 181(43), 1–20.
- Baylor, A. L., & Kim, S. (2009). Designing Nonverbal Communication for Pedagogical Agents: When Less Is More. *Computers in Human Behavior*, 25(2), 450–457.
- Bione, T., Grimshaw, J., & Cardoso, W. (2017). An Evaluation of TTS as a Pedagogical Tool for Pronunciation Instruction: The 'Foreign' Language Context. In K. Borthwick, L. Bradley, & S. Thoušny (Eds.), *CALL in a Climate of Change: Adapting to Turbulent Global Conditions – Short Papers from EUROCALL 2017* (pp. 56–61). Research-publishing.net.
- Brown, T., et al. (2020). Language Models Are Few-Shot Learners. *Advances in Neural Information Processing Systems (NeurIPS)*, 33, 1877–1901.
- Cao, Y., Tien, W. C., Faloutsos, P., & Pighin, F. (2005). Expressive Speech-Driven Facial Animation. *ACM Transactions on Graphics*, 24(4), 1283–1302.
- Cardoso, W. (2018). Learning L2 Pronunciation with a Text-to-Speech Synthesizer. In P. Taalas, J. Jalkanen, L. Bradley, & S. Thoušny (Eds.), *Future-Proof CALL: Language Learning as Exploration and Encounters – Short Papers from EUROCALL 2018* (pp. 43–48). Research-publishing.net.
- Cassell, J., Sullivan, J., Prevost, S., & Churchill, E. F. (Eds.). (2000). *Embodied Conversational Agents*. MIT Press.
- Cucchiaroni, C., Strik, H., & Boves, L. (2000). Quantitative Assessment of Second Language Learners' Fluency by Means of ASR Technology. *Journal of the Acoustical Society of America*, 107(2), 989–999.
- Cudeiro, D., Bolkart, T., Laidlaw, C., Ranjan, A., & Black, M. J. (2019). Capture, Learning, and Synthesis of 3D Speaking Styles. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10101–10111.
- Doddapaneni, S., et al. (2023). Towards Leaving No Indic Language Behind: Building Monolingual Corpora, Benchmark and Models for Indic Languages. *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*, 12402–12426.
- Edwards, P., Landreth, C., Fiume, E., & Singh, K. (2016). JALI: An Animator-Centric Viseme Model for Expressive Lip Synchronization. *ACM Transactions on Graphics*, 35(4), 127:1–127:11.

- Gaur, N., Farris, B., Haghani, P., Leal, I., Moreno, P. J., Prasad, M., Ramabhadran, B., & Zhu, Y. (2021). Mixture of Informed Experts for Multilingual Speech Recognition. *IEEE ICASSP*, 6234–6238.
- Gratch, J., Wang, N., Gerten, J., Fast, E., & Duffy, R. (2007). Creating Rapport with Virtual Agents. In C. Pelachaud et al. (Eds.), *Intelligent Virtual Agents (IVA 2007)*, Lecture Notes in Computer Science (Vol. 4722, pp. 125–138). Springer, Berlin, Heidelberg.
- Gupta, A., & Gaur, S. (2019). Hindi Text-to-Speech System Using Deep Learning. *International Conference on Communication and Electronics Systems (ICCES)*, 1297–1301.
- Han, J., Zhao, W., & Chen, L. (2024). LLM-Based Conversational AI Tutors for Language Learning: A Design Framework. *Proc. ACM Learning at Scale (L@S)*, 45–56.
- Hansen, M. (2023). Piper: A Fast, Local Neural Text to Speech System [Computer software]. GitHub. <https://github.com/rhasspy/piper>
- Heidig, S., & Clarebout, G. (2011). Do Pedagogical Agents Make a Difference to Student Motivation and Learning? *Educational Research Review*, 6(1), 27–54.
- Huang, X., Zou, D., Cheng, G., & Xie, H. (2021). A Systematic Review of AR and VR Enhanced Language Learning. *Sustainability*, 13(9), 4639.
- Javed, T., Bhogale, K., Raman, A., Kumar, P., Kunchukuttan, A., & Khapra, M. M. (2023). IndicSUPERB: A Speech Processing Universal Performance Benchmark for Indian Languages. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(11), 12942–12950.
- Javed, T., et al. (2024). IndicVoices: Towards Building an Inclusive Multilingual Speech Dataset for Indian Languages. *Findings of the Association for Computational Linguistics: ACL 2024*, 10740–10782.
- Ji, H., & Park, M. (2023). The Effects of Embodied Conversational Agents on Foreign Language Anxiety and Speaking Competence. *Computers & Education*, 194, 104698.
- Johnson, W. L., & Lester, J. C. (2016). Face-to-Face Interaction with Pedagogical Agents, Twenty Years Later. *International Journal of AI in Education*, 26(1), 25–36.
- Joshi, P., Santy, S., Budhiraja, A., Bali, K., & Choudhury, M. (2020). The State and Fate of Linguistic Diversity and Inclusion in the NLP World. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, 6282–6293.
- Kakwani, D., et al. (2020). IndicNLP Suite: Monolingual Corpora, Evaluation Benchmarks and Pre-trained Multilingual Language Models for Indian Languages. *Findings of EMNLP 2020*, 4948–4961.
- Kannan, J., & Munday, P. (2018). New Trends in Second Language Learning and Teaching through the Lens of ICT, Networked Learning, and Artificial Intelligence. *Círculo de Lingüística Aplicada a la Comunicación*, 76, 13–30.
- Kasneci, E., et al. (2023). ChatGPT for Good? On Opportunities and Challenges of Large Language Models for Education. *Learning and Individual Differences*, 103, 102274.
- Kim, I. S. (2006). Automatic Speech Recognition: Reliability and Pedagogical Implications for Teaching Pronunciation. *Educational Technology & Society*, 9(1), 322–334.
- Kim, J., Kong, J., & Son, J. (2021). Conditional Variational Autoencoder with Adversarial Learning for End-to-End Text-to-Speech. *Proceedings of the 38th International Conference on Machine Learning (ICML)*, PMLR 139, 5530–5540.
- Kohnke, L., Moorhouse, B. L., & Zou, D. (2023). ChatGPT for Language Teaching and Learning. *RELC Journal*, 54(2), 537–550.
- Kunchukuttan, A. (2020). The IndicNLP Library. AI4Bharat Open Research Initiative. https://github.com/anoopkunchukuttan/indic_nlp_library
- Mai, Z., & Carson-Berndsen, J. (2024). Enhancing Conversational Quality in Language Learning Chatbots: An Evaluation of GPT4 for ASR Error Correction. *IEEE ICASSP*, 11001–11005.
- Pham, X. L., Pham, T., Nguyen, Q. M., Nguyen, T. H., & Cao, T. H. (2018). Chatbot as an Intelligent Personal Assistant for Mobile Language Learning. *Proceedings of the 2018 2nd International Conference on Education and E-Learning (ICEEL 2018)*, 16–21.
- Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2023). Robust Speech Recognition via Large-Scale Weak Supervision. *Proceedings of ICML*, 28492–28518.
- Reddy, S., Labutov, I., Banerjee, S., & Joachims, T. (2016). Unbounded Human Learning: Optimal Scheduling for Spaced Repetition. *Proceedings of ACM SIGKDD*, 1815–1824.
- Settles, B., & Meeder, B. (2016). A Trainable Spaced Repetition Model for Language Learning. *Proceedings of ACL*, 1848–1858.
- Settles, B., Brust, C., Gustafson, E., Hagiwara, T., & Madnani, N. (2018). Second Language Acquisition Modeling. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2018, NAACL-HLT)*, 56–65.
- Tabibian, B., Upadhyay, U., De, A., Zareade, A., Schölkopf, B., & Gomez-Rodriguez, M. (2019). Enhancing Human Learning via Spaced Repetition Optimization. *PNAS*, 116(10), 3988–3993.
- Tan, X., Qin, T., Soong, F., & Liu, T. Y. (2021). A Survey on Neural Speech Synthesis. *arXiv preprint arXiv:2106.15561*.
- Taylor, S., et al. (2017). A Deep Learning Approach for Generalized Speech Animation. *ACM Transactions on Graphics (SIGGRAPH)*, 36(4), 93:1–93:11.
- Touvron, H., et al. (2023). LLaMA: Open and Efficient Foundation Language Models. *arXiv preprint arXiv:2302.13971*.
- Vandewaetere, M., & Desmet, P. (2012). Introduction to the Special Issue on Adapting in CALL. *Computer Assisted Language Learning*, 25(5), 399–404.
- Wolf, D. S. (2023). Rhubarb Lip Sync: Automated Lip-Sync Tool for 2D and 3D Characters [Computer software]. GitHub. <https://github.com/DanielSWolf/rhubarb-lip-sync>
- Wozniak, P. A., & Gorzelanczyk, E. J. (1994). Optimization of Repetition Spacing in the Practice of Learning. *Acta Neurobiologiae Experimentalis*, 54, 59–62.

-
- Yan, L., et al. (2024). Practical and Ethical Challenges of Large Language Models in Education: A Systematic Scoping Review. *British Journal of Educational Technology*, 55(1), 90–112.
 - Ye, J., Su, J., & Cao, Y. (2022). A Stochastic Shortest Path Algorithm for Optimizing Spaced Repetition Scheduling. *Proceedings of ACM SIGKDD*, 4381–4390.