Beyond the AI Conversation Partner: A Critical Integrative Review of Generative AI-Supported L2 Speaking and the SPEAK Implementation Framework

DMPPK Dasanayake*

Assistant Lecturer in English, Sri Lanka Institute of Advanced Technological Education (SLIATE), Sri Lanka

***Corresponding author**

DOI: https://doi.org/10.47772/IJRISS.2026.100700044

Received: 10 July 2026; Accepted: 15 July 2026; Published: 23 July 2026

ABSTRACT

Generative artificial intelligence (GenAI) has rapidly expanded from text-based assistance to voice-enabled dialogue, automated oral feedback, and multimodal speaking assessment. These developments appear to address a persistent problem in second-language (L2) education: learners need frequent, low-risk opportunities to speak, but teachers cannot always provide individual interaction and feedback at scale. However, the availability of an apparently fluent conversation partner does not establish that durable speaking development, fair assessment, or transfer to human communication will follow. This critical integrative review examines three questions: what learning and affective outcomes are associated with GenAI-supported L2 speaking practice; what limitations weaken its pedagogical value; and what implementation conditions support responsible use. Targeted searches of Google Scholar, publisher platforms, open research repositories, and citation chains produced an analytical corpus of 11 core publications, supported by 14 theoretical, methodological, assessment, feedback-literacy, and ethics sources published or retained for interpretation through July 2026. Evidence was appraised for contextual clarity, task alignment, outcome validity, feedback transparency, and the strength of claims. The synthesis indicates that GenAI can increase practice volume, support rehearsal, provide scenario-based interaction, and reduce fear of immediate human judgement. Adaptive prompts and rapid feedback may also support vocabulary retrieval, discourse organisation, pronunciation awareness, and willingness to communicate. Yet the evidence remains dominated by small samples, short interventions, self-report measures, prototype studies, and technical benchmarks. Recurring risks include inaccurate or overconfident feedback, accent and speech-recognition bias, unnatural interaction, dependency, privacy concerns, unequal access, and weak transfer evidence. The review proposes the SPEAK framework: Structured tasks, Progressive intelligibility-focused feedback, Equity and ethics, Agency and anxiety-sensitive practice, and Knowledgeable teacher oversight. GenAI should therefore be used as a structured rehearsal and feedback resource within teacher-designed speaking pedagogy, not as an autonomous replacement for human interaction or professional judgement.

Keywords: generative artificial intelligence; second-language speaking; conversational agents; pronunciation feedback; speaking anxiety

INTRODUCTION

Speaking is one of the most visible and consequential outcomes of second-language learning. It is required for classroom participation, oral assessment, interviews, professional collaboration, service encounters, and social interaction. Unlike written production, speaking requires learners to conceptualise a message, retrieve linguistic resources, articulate it, monitor comprehensibility, and respond to an interlocutor under immediate time pressure [3, 14]. The difficulty is not simply a lack of grammatical knowledge. Learners may understand vocabulary and structures but still hesitate, abandon turns, depend on memorised expressions, or remain silent because they fear negative evaluation.

A central pedagogical problem is therefore one of opportunity. Speaking develops through repeated, purposeful production and interaction, yet individual learners often receive little sustained speaking time in teacher-led classes. Large enrolments, limited lesson hours, examination-oriented curricula, mixed proficiency, and teacher workload restrict the amount of personalised practice and feedback available. These constraints are especially important in many Asian and resource-constrained English as a Foreign Language (EFL) settings, where learners may have limited access to English outside the classroom and may perceive public error as socially costly.

Conversational artificial intelligence appears to offer a practical response. Voice-enabled large language models, adaptive chatbots, automated speech-recognition systems, and multimodal assessment tools can now conduct scenario-based dialogue, ask follow-up questions, transcribe learner speech, explain vocabulary, reformulate utterances, and generate immediate feedback. Unlike fixed dialogue trees, generative systems can respond to unexpected content and sustain interaction across topics. Learners can repeat a task privately, request simpler language, rehearse an interview, or compare alternative formulations without waiting for a teacher or partner.

The educational promise is substantial, but it must be interpreted carefully. A system may produce fluent language while misjudging pronunciation, accepting an unintelligible utterance, correcting a legitimate regional variety, or giving advice that is linguistically plausible but pragmatically unsuitable. A pleasant interaction can increase confidence without improving independent performance. Technical accuracy on a benchmark does not demonstrate that learners understand or use feedback, and a higher amount of practice does not guarantee transfer to spontaneous human interaction. Moreover, voice data create privacy risks, subscription costs create inequity, and model behaviour can reflect accent, cultural, and training-data bias.

Research is developing rapidly but remains fragmented across applied linguistics, computer-assisted language learning, speech technology, educational technology, and human-computer interaction. Some studies evaluate learner perceptions or willingness to communicate; others test prototype conversation systems or automatic scoring models; still others focus on pronunciation, oral presentations, or virtual agents. The field needs a synthesis that distinguishes conversational availability from pedagogical quality and that connects learner outcomes with task design, feedback, teacher mediation, and ethical safeguards.

The present review responds to that need. It focuses on GenAI-supported L2 speaking, defined here as the use of generative or adaptive conversational and speech technologies to create, guide, evaluate, or provide feedback on oral language practice. The review includes the immediate GenAI period as well as closely related conversational-agent and speech-assessment research that clarifies the mechanisms and limitations of current systems. The purpose is not to determine whether one commercial application is superior. It is to identify what the evidence currently supports, where claims exceed evidence, and how educators can integrate these tools without weakening communicative, human, and ethical dimensions of speaking development.

Research Questions

1. What speaking, interactional, motivational, and affective outcomes are associated with GenAI-supported L2 speaking practice?

2. What technological, pedagogical, methodological, and ethical limitations reduce the educational value of GenAI-supported speaking tools?

3. What design and implementation conditions support effective, equitable, and responsible use in L2 education?

Contribution of the Review

The review makes three contributions. First, it separates five functions that are often treated as interchangeable: conversation practice, task rehearsal, language support, corrective feedback, and oral assessment. Second, it evaluates positive findings alongside the methodological strength and ecological validity of the evidence. Third, it translates the synthesis into the SPEAK implementation framework, intended for teachers, curriculum designers, researchers, and institutions. The framework emphasises that educational value is produced by the relationship among task, learner, feedback, teacher, and context rather than by the sophistication of the model alone.

Conceptual Foundations

Speaking as Interactional and Multidimensional Performance

Speaking proficiency includes fluency, grammatical and lexical control, intelligibility, discourse organisation, interactional responsiveness, and strategic competence. Communicative success depends not only on producing accurate sentences but also on managing turns, interpreting another speaker, signalling uncertainty, repairing misunderstanding, and adapting language to audience and purpose [4, 7]. Consequently, a speaking tool that improves isolated pronunciation or produces longer responses may still leave important aspects of interaction undeveloped.

This distinction is particularly important for AI-mediated practice. Automatic systems can count pauses, compare acoustic patterns, estimate grammatical accuracy, and generate reformulations. They are less reliable at judging whether a learner has built rapport, interpreted implicit meaning, responded sensitively, or negotiated disagreement in a culturally appropriate way. Evaluation should therefore distinguish measurable speech features from broader communicative competence.

Interaction, Output, and Noticing

Interactionist accounts of second-language acquisition explain why conversational technology may be useful. Meaning-focused exchange can create pressure to produce output, notice gaps, request clarification, and modify language after feedback [13, 20]. A conversational system can increase the frequency of these opportunities by remaining available beyond class and by allowing repetition without exhausting a human partner.

However, interaction becomes developmentally useful only when it generates cognitively meaningful work. If the system supplies complete answers, automatically repairs every message, or accepts minimal responses, the learner may avoid retrieval, formulation, and monitoring. The pedagogical issue is therefore not whether dialogue occurs, but whether the design preserves a productive gap between what the learner can currently do and what the task requires.

Anxiety, Willingness to Communicate, and Agency

Speaking is strongly shaped by anxiety and willingness to communicate. Fear of negative evaluation can reduce participation even when learners possess sufficient linguistic knowledge [9, 16]. Private interaction with an AI agent can lower the social cost of error, give learners time to restart, and permit rehearsal before public performance. This low-risk environment is one of the most consistent perceived benefits in emerging studies.

Yet reduced anxiety should not become permanent avoidance of human communication. Speaking confidence is educationally valuable when it supports movement towards more demanding interaction, not when it confines learners to a predictable system that is always patient, always available, and unlikely to challenge face-threatening behaviour. Learner agency also requires the ability to question feedback, choose whether to adopt a reformulation, and maintain ownership of ideas and voice.

From Accuracy to Intelligibility

Pronunciation support is a major attraction of voice-enabled systems. Nevertheless, contemporary pronunciation pedagogy prioritises intelligibility and comprehensibility rather than imitation of a single native-speaker norm [7, 11]. Automatic feedback that penalises accent features which do not obstruct understanding may reinforce linguistic inequality. Conversely, speech recognition that silently converts an unclear utterance into the intended text may give learners false confidence.

A responsible system should therefore explain what it is evaluating, distinguish confidence scores from human understanding, tolerate legitimate accent variation, and direct attention to features that affect meaning. Automated pronunciation scores are best treated as formative signals to be interpreted alongside human listening, not as unquestionable judgements of speaking ability.

MATERIALS AND METHODS

Review Design

A critical integrative review design was adopted. Integrative reviews are appropriate when a developing field contains experimental, qualitative, mixed-methods, design-based, and technical evaluation studies and when the purpose is to generate an explanatory synthesis rather than calculate a single pooled effect [21, 23]. The evidence on AI-supported speaking is too heterogeneous in intervention, learner population, outcome measure, and publication status for a defensible meta-analysis. The review combined learner-facing evidence with system and assessment studies because the educational claims of conversational AI depend partly on technological capabilities. However, technical benchmark performance was not treated as proof of learning. Reviews and conceptual papers were used to contextualise findings, while empirical and design-oriented studies formed the primary analytical corpus.

Search Strategy and Scope

Searches were updated in July 2026 and covered publications from January 2020 to July 2026. This period captures the transition from rule-based and neural conversational agents to large language model and multimodal voice systems. Searches used Google Scholar, publisher platforms, open research repositories, and backward and forward citation chaining from major reviews and relevant empirical papers. Search combinations included (generative AI OR ChatGPT OR large language model OR conversational agent OR chatbot OR speech recognition) AND (second language OR foreign language OR EFL OR ESL) AND (speaking OR oral proficiency OR conversation OR pronunciation OR oral presentation OR speaking anxiety OR willingness to communicate). Additional searches used adaptive dialogue, automated speaking assessment, intelligibility, corrective feedback, voice mode, and embodied agent. Because the newest work often appears first as conference manuscripts or public preprints, the review did not exclude a study solely because peer review was incomplete. Instead, publication status and evidential limitations were recorded and weighted in interpretation. The inclusion of emerging evidence is intended to map current developments, not to imply that all sources have equivalent authority.

Eligibility Criteria

Criterion	Included	Excluded
Population/context	Learners using English or another target language as an L2/EFL/ESL in an educational or structured practice context.	Studies unrelated to language learning or without a speaking-related purpose.
Technology	Generative, adaptive, conversational, speech-recognition, or multimodal AI used for oral practice, feedback, support, or assessment.	Static audio/video materials without adaptive or automated interaction.
Evidence	Learner outcomes, interaction data, perceptions, feedback quality, system evaluation, or oral-assessment validity.	Promotional claims, product reviews, opinion pieces, or papers without usable evidence.
Time/language	English-language publications from 2020 to July 2026; seminal SLA and assessment sources retained for theory.	Earlier technology studies unless needed to explain a core mechanism.
Publication status	Peer-reviewed work prioritised; high-relevance full conference manuscripts and preprints included with status acknowledged.	**ABSTRACT**-only records or unverifiable citations.

Table 1: Eligibility criteria for the critical integrative review.

Search and Selection Transparency

The searches were iterative, targeted, and supplemented by backward and forward citation chaining. A complete de-duplicated log of every raw search result was not retained; therefore, this critical integrative review does not claim PRISMA-style exhaustiveness. The final analytical corpus comprised 11 core publications that met the criteria in Table 1. Fourteen additional theoretical, methodological, assessment, feedback-literacy, and ethics sources were retained for contextual interpretation, giving 25 references overall. Appendix A identifies every core publication and explains its role in the synthesis.

Stage	Transparent report
Search approach	Targeted searches of Google Scholar, publisher platforms, open repositories, and citation chains; updated in July 2026.
Raw search records	A complete de-duplicated total was not retained because the search was iterative; no PRISMA flow count is claimed.
Core analytical corpus	11 publications meeting the eligibility criteria and directly informing the results.
Contextual sources	14 theoretical, methodological, assessment, feedback-literacy, and ethics sources.
Total references	25 sources listed in the reference section.

Table 2: Search and selection transparency summary.

Appraisal and Synthesis

Each included publication was examined for the clarity of its educational context, description of the AI system, alignment between task and outcome, duration and intensity of exposure, appropriateness of comparison, validity of the speaking measure, transparency of feedback, and correspondence between results and conclusions. Particular attention was given to whether a study measured actual speaking performance, learner perception, system accuracy, or short-term task success, because these outcomes cannot be interpreted as equivalent.

The evidence was coded into five connected domains: access to speaking practice; task and interaction quality; feedback and assessment; affect, motivation, and agency; and implementation risks. Themes were compared across learner-facing and technical studies. Contradictory or negative findings were retained. The final synthesis generated the SPEAK framework by identifying conditions that recurred across benefits, limitations, and practical recommendations. The coding and theme-development process was informed by established thematic-analysis principles [2], while retaining an integrative-review focus rather than treating the publications as interview data.

RESULTS

Evidence Base and General Pattern

The analytical corpus showed rapid technological development but a comparatively immature educational evidence base. Learner-facing studies commonly involved prototypes, small convenience samples, short practice periods, or perception surveys. Technical studies reported increasingly strong performance in dialogue generation, speech recognition, or automated scoring, but rarely examined whether learners understood feedback or transferred improvement to a new human interaction. Reviews consistently identified personalisation, availability, and immediate response as advantages while warning about reliability, teacher readiness, privacy, and over-reliance [18, 24].

The overall pattern was promising but conditional. AI-supported speaking created more opportunities to produce language, rehearse tasks, and receive rapid support. Positive outcomes were strongest when the system was

connected to a defined communicative goal and when learners were required to formulate their own responses. Evidence was weaker when studies treated satisfaction, engagement, or technical score agreement as direct evidence of durable speaking development.

Expanded Practice and Rehearsal

The clearest affordance was increased access to spoken practice. Conversational systems can be used outside class, repeated without scheduling a partner, and adjusted to learner pace. Li et al. [12] demonstrated a language-learning dialogue system designed to adapt to proficiency and provide grammar-related feedback. Qian et al. [19] further showed the value of curriculum-constrained dialogue, in which generated interaction incorporated target vocabulary rather than drifting into unrestricted conversation. These designs illustrate an important principle: open-endedness is educationally useful only when it remains connected to curricular purpose.

Scenario-based systems also supported rehearsal. Xu et al. [25] developed LLM-based situational dialogues intended to combine open conversation with focused practice across familiar and unseen topics. Cha et al. [6] evaluated a ChatGPT-based platform for oral-presentation practice with 13 EFL learners and identified both the perceived value of personalised feedback and design weaknesses requiring improvement. Such systems can allow learners to rehearse interviews, presentations, customer-service exchanges, and academic discussions before performing for a human audience.

Practice quantity alone, however, is an incomplete outcome. A learner may repeat a familiar exchange while relying on narrow vocabulary, accepting AI-generated completions, or avoiding unpredictable interaction. Stronger designs included an outcome, changing information, follow-up questions, or post-task reflection. The evidence therefore supports GenAI as an opportunity multiplier, but not as a guarantee of communicative complexity.

Interactional Support and Adaptive Scaffolding

Generative systems can alter vocabulary, speech rate, prompt difficulty, and feedback according to learner response. Qian et al. [19] reported that curriculum-infused dialogue improved understanding of target words and increased interest in English practice. The capacity to request clarification, examples, slower speech, or reformulation can support learners who would otherwise withdraw from a demanding task.

Adaptive support must nevertheless be calibrated. Excessive assistance reduces the need for retrieval and message construction. When the system supplies complete sentences before the learner attempts a response, interaction becomes assisted performance rather than learning. A productive sequence begins with learner production, offers a limited cue or clarification, and only then provides more explicit support. This preserves cognitive effort while preventing repeated failure.

Natural interaction is another unresolved issue. Iizuka and Mori [10] found that conversational agents using speech based on spontaneous rather than read-speech models elicited shorter response times, more backchannels, and perceptions closer to human conversation. The finding suggests that voice quality and turn timing influence user behaviour. Yet even advanced systems can produce long, overly polished, socially formulaic responses that do not resemble ordinary conversation. Learners therefore need experience with interruption, hesitation, ambiguity, disagreement, and varied human accents in addition to AI dialogue.

Anxiety, Confidence, and Willingness to Communicate

Learners frequently described AI interaction as less threatening than speaking before teachers or peers. The absence of visible impatience and public judgement creates space for repetition, self-correction, and experimentation. Embodied and augmented-reality systems have also reported perceived reductions in speaking anxiety and increased autonomy. For example, Bendarkawi et al. [1] evaluated an LLM-powered augmented-reality environment for group conversation with 10 participants; users reported lower anxiety and greater autonomy than they associated with in-person practice.

This affective benefit is educationally important because anxiety can reduce output and limit the very practice needed for development. AI can function as a rehearsal stage between private planning and public performance. However, self-reported comfort should not be interpreted as a proficiency gain. A learner may feel confident because the system is unusually tolerant, predicts intended words, or avoids direct correction. Confidence becomes robust when it is tested through progressively less supported interaction with peers, teachers, and unfamiliar listeners.

The most defensible use is therefore transitional. Learners may begin with private AI rehearsal, move to paired practice using the same communicative goal, and complete the cycle through a human report, role-play, presentation, or discussion. This sequence converts reduced anxiety into participation rather than avoidance.

Feedback on Language, Pronunciation, and Oral Performance

GenAI can provide immediate reformulations, vocabulary alternatives, grammar explanations, discourse suggestions, and comments on presentation structure. Cha et al. [6] found that personalised oral-presentation feedback was valued but also identified weaknesses in feedback quality and platform design. Meng [17] reported that customised ChatGPT voice configurations produced more balanced feedback and emotional support than uncustomised versions, although cultural responsiveness did not improve consistently. These studies indicate that prompt and system design influence the type of feedback learners receive.

Pronunciation and fluency assessment present greater technical and pedagogical difficulty. Fu et al. [8] demonstrated competitive scoring of accuracy and fluency using a multimodal LLM architecture on a pronunciation dataset. Ma et al. [15] reported that speech large language models outperformed previous baselines on two oral-proficiency datasets and generalised across tasks. These are important technical advances, but they do not establish that automated scores are fair across accents or that feedback improves learning.

Speech recognition may both reveal and conceal problems. A transcript can help learners notice when pronunciation prevents recognition, but a powerful model may infer intended meaning despite unclear articulation. Conversely, a legitimate accent may be misrecognised because it is underrepresented in training data. Educators should therefore avoid presenting a single automated score as an objective measure. Feedback should prioritise intelligibility, meaning, and listener effort, and learners should compare AI output with human judgement and self-reflection.

Feedback timing also matters. Continuous interruption can fragment fluency and increase cognitive load, whereas delayed feedback may allow errors to recur without attention. A useful pattern is to complete a short communicative task, select two or three high-value issues, practise revised language, and repeat the task. GenAI is particularly suitable for this cycle because it can preserve a transcript and generate targeted examples, but the selection of priorities should remain pedagogically controlled.

Transfer, Dependence, and the Limits of Simulated Communication

The most important unanswered question is transfer. Many studies demonstrate successful interaction with a system or improvement on a closely aligned task, but few test whether gains remain after support is removed or whether learners communicate more effectively with unfamiliar humans. AI partners are unusually cooperative: they can infer incomplete language, wait indefinitely, ignore pragmatic awkwardness, and adapt continuously to one user. Human interaction is less predictable and includes identity, power, emotion, humour, disagreement, and shared social history.

Dependence can arise when the system becomes a source of ready-made wording. Learners may ask the model to generate a complete answer, memorise it, and obtain a polished performance without developing flexible resources. The risk is especially high in oral presentations and interviews where scripted language can temporarily improve delivery. To protect learning, tasks should require learners to explain choices, respond to unanticipated questions, paraphrase without the tool, and perform a comparable transfer task.

The review therefore distinguishes performance support from development. Performance support helps a learner complete the present task; development increases the capacity to perform future tasks independently. Both can be valuable, but they should not be reported as the same outcome.

Equity, Privacy, and Cultural-Linguistic Bias

Voice-enabled GenAI can widen access for learners who lack conversation partners, but it can also create new inequalities. Advanced voice modes, stable connectivity, modern devices, and sufficient data may require paid access. Learners with older phones, limited bandwidth, speech disabilities, or accents poorly represented in training data may receive a weaker experience. A programme that makes a commercial tool compulsory can therefore reproduce socioeconomic and linguistic disadvantage.

Voice interaction also involves sensitive data. Recordings may reveal identity, accent, location, health information, or personal experiences. Learners need clear information about what is recorded, where it is processed, whether it is retained, and how it may be used. Institutions should minimise personal disclosure, avoid uploading identifiable assessment data without authority, provide non-AI alternatives, and follow age-appropriate consent procedures. These safeguards align with UNESCO guidance on human-centred, transparent, and privacy-conscious use of generative AI in education [22].

Cultural responsiveness remains uncertain. Meng [17] found that targeted customisation did not consistently improve cultural responsiveness. A model may use culturally inappropriate examples, treat one pragmatic norm as universal, or reproduce stereotypes. Teacher review and locally relevant scenario design are therefore essential, particularly when tasks involve workplace hierarchy, gender, religion, politeness, or sensitive social issues.

Integrated Synthesis

Evidence domain	Promising contribution	Main limitation	Educational inference
Practice access	Frequent, repeatable, private oral interaction beyond class.	Quantity may remain repetitive or poorly aligned.	Use defined outcomes, target language, and transfer tasks.
Scaffolding	Adaptive prompts, clarification, vocabulary, and reformulation.	Too much support can remove productive struggle.	Require learner production before graduated assistance.
Affect	Lower perceived judgement; greater willingness to rehearse.	Comfort may not transfer to human interaction.	Use AI as a bridge to peer and public speaking.
Feedback	Immediate comments on language, discourse, and delivery.	Accuracy, prioritisation, and cultural fit are inconsistent.	Limit feedback, verify it, and combine it with human judgement.
Pronunciation/assessment	Rapid speech analysis and increasingly strong benchmark performance.	Accent bias and construct validity remain concerns.	Prioritise intelligibility; do not use one score as a high-stakes decision.
Equity/ethics	Potential access for learners without partners.	Cost, data use, disability access, and bias may exclude learners.	Provide alternatives, minimise data, and audit access and fairness.

Table 3: Integrated synthesis of the evidence.

DISCUSSION

What the Evidence Supports

The current evidence supports a bounded conclusion. GenAI can expand access to oral rehearsal, create scenario-based dialogue, offer rapid language support, and reduce the immediate social threat of speaking. It is particularly useful where learners have limited opportunities outside class or where teachers cannot provide individual practice at the desired frequency. Its value is strongest for short, defined tasks and iterative rehearsal, such as preparing an interview response, practising a service encounter, organising an oral presentation, or repeating a problem-solving report.

The evidence does not yet support treating GenAI as an autonomous speaking tutor capable of replacing human interaction, diagnostic judgement, or assessment. Most learner-facing studies remain too small or short to establish durable proficiency gains. Technical studies show what systems can score or generate, not necessarily what learners acquire. Transfer, long-term retention, interactional competence, and fairness across accents remain underexamined.

These findings are consistent with a broader principle in technology-enhanced language learning: a tool's affordances become educational only through task design, learner activity, feedback use, and teacher mediation. The same voice model can support meaningful rehearsal or encourage passive imitation depending on how it is integrated.

The Continuing Role of the Teacher

GenAI changes rather than removes the teacher's role. Teachers select communicative purposes, calibrate difficulty, model interaction strategies, decide which feedback deserves attention, and create movement from private practice to social communication. They also recognise affective and cultural factors that an automated system may miss. In large classes, AI can absorb some routine rehearsal and preliminary feedback while teachers focus on interactional quality, individual needs, and higher-value feedback.

Teacher oversight should not mean monitoring every private conversation. It can be implemented through shared prompt templates, brief learner reflection logs, sample transcript analysis, peer comparison, and classroom transfer tasks. Learners can be asked to identify one useful suggestion, one questionable response, and one change they independently chose. This develops AI literacy and feedback judgement rather than tool obedience.

Relevance to Resource-Constrained Contexts

In resource-constrained EFL contexts, implementation should begin with feasibility rather than novelty. Mobile-first access, short audio turns, low-bandwidth options, shared institutional accounts, and non-AI alternatives are more important than immersive avatars. The teacher can provide a small bank of locally relevant scenarios and prompts that learners use on available devices in rotating groups or outside class. Classroom time can then be used for human interaction and feedback.

A modest implementation may be more educationally sustainable than a technologically complex one. For example, learners can conduct a five-minute AI interview rehearsal, note two language problems, revise one answer, and perform a peer interview in class. This design uses the system for what it does well-availability and repetition-while preserving human unpredictability, accountability, and social meaning.

The Speak Implementation Framework

The synthesis produced the SPEAK framework as a practical planning and evaluation model. The five components are mutually dependent. A programme cannot be considered strong because it provides structured tasks while ignoring privacy, or because it reduces anxiety while giving unreliable feedback.

Component	Core principle	Teacher actions	Indicators of quality
S - Structured tasks	AI dialogue serves a clear communicative outcome rather than unrestricted chatting.	Define role, audience, information gap, time, target functions, and a human transfer task.	Learners produce extended responses, make decisions, and complete an identifiable outcome.
P - Progressive feedback	Feedback is selective, intelligibility-focused, and followed by revised performance.	Delay most correction until task completion; prioritise two or three issues; compare AI and human feedback.	Learners explain, apply, and transfer feedback instead of copying reformulations.
E - Equity and ethics	Access, privacy, disability, linguistic fairness, and transparency are designed into use.	Provide alternatives; minimise personal data; disclose limitations; audit device, cost, and accent effects.	No learner is penalised for lacking paid access; data practices are understood and consented to.
A - Agency and anxiety-sensitive practice	AI lowers unnecessary threat while preserving learner ownership and movement toward human interaction.	Allow private rehearsal and restarting; require learner choices, reflection, and peer/public progression.	Confidence is accompanied by increased human participation, not dependence on the tool.
K - Knowledgeable teacher oversight	Professional judgement governs task design, feedback priorities, and assessment.	Create prompt templates; sample outputs; correct misleading advice; use multiple sources of evidence.	AI informs learning but does not independently determine grades or replace teacher diagnosis.

Table 4: The SPEAK framework for responsible GenAI-supported L2 speaking.

Structured Tasks

Unrestricted conversation may be engaging but produces uneven language and weak evidence of progress. Structured tasks specify who is speaking, why, what information or decision is needed, and what counts as completion. They can still allow creativity. A hotel complaint, job interview, project meeting, academic consultation, or problem-solving task gives the system a role while preserving learner choice.

Tasks should include a transfer stage. After AI rehearsal, learners perform a changed version with a peer, respond to an unanticipated follow-up question, or summarise the exchange without support. This prevents close task familiarity from being mistaken for general speaking development.

Progressive Feedback

Feedback should move from meaning to form and from less to more explicit support. During the first attempt, the system may ask for clarification rather than immediately rewriting the utterance. After completion, it can identify a small number of high-value issues involving intelligibility, vocabulary precision, grammar that affects meaning, discourse organisation, or interaction strategy.

The learner should decide what to change and explain why. A revised attempt provides evidence of uptake, while a later transfer task provides stronger evidence of learning. Automated scores should be accompanied by examples and confidence statements where possible. Requiring learners to interpret, evaluate, and apply comments also develops feedback literacy rather than passive dependence on corrections [5].

Equity and Ethics

Institutions should identify the minimum technical requirement, cost, language availability, accessibility features, data-retention policy, and age restrictions before adoption. Learners should not be required to disclose personal experiences to create realistic dialogue. Fictionalised roles and institutionally approved scenarios can achieve authenticity without unnecessary exposure.

Fairness audits should include local accents and proficiency levels. When a system repeatedly misrecognises a group, the response should be to question the system, not to label the learners deficient. Alternative human or non-AI routes must remain available.

Agency and Anxiety-Sensitive Practice

Anxiety-sensitive design offers planning time, private rehearsal, restart options, and graduated difficulty. It avoids public display of low scores and separates formative practice from high-stakes judgement. At the same time, learner agency requires active evaluation of system output. Students should be encouraged to reject unsuitable wording, retain their own accent and identity, and ask the system to explain rather than simply replace.

Progression towards human communication is essential. A practical sequence is private rehearsal, AI-supported repetition, peer interaction, teacher feedback, and independent performance.

Knowledgeable Teacher Oversight

Teachers need sufficient AI literacy to understand that fluent output may be wrong, speech recognition may be biased, and commercial systems can change without notice. Oversight includes selecting tasks,

reviewing sample output, establishing acceptable use, and interpreting automated feedback in relation to course outcomes.

For assessment, AI should provide one source of formative evidence rather than the sole basis for high-stakes decisions. Human raters remain necessary for pragmatic appropriateness, interactional responsiveness, and contextual judgement.

Pedagogical Implications

A Classroom Integration Cycle

A five-stage cycle can operationalise the framework. First, the teacher introduces the communicative context, success criteria, and responsible-use rules. Second, learners complete a brief AI rehearsal using a structured prompt. Third, they inspect the transcript or feedback and select a limited number of changes. Fourth, they repeat or adapt the task with a peer or small group. Fifth, they reflect on what transferred, what the AI handled poorly, and what requires teacher support.

This cycle protects oral production as the learner's work. The system provides interaction and feedback, but the learner must formulate, evaluate, revise, and transfer. Teachers can assess the final human performance and use AI interaction only as formative evidence.

Example Prompt Template

A teacher-approved prompt can state: 'Act as a customer who has received the wrong hotel booking. Conduct a six-turn conversation with an intermediate English learner. Ask one question at a time. Do not write the learner's answer. If meaning is unclear, request clarification. After the conversation, give feedback on one strength, one intelligibility issue, one useful expression, and one interaction strategy. Do not score accent similarity.' This type of prompt defines role, turn structure, support, feedback limits, and a communicative rather than native-like target.

Curriculum and Institutional Policy

Curricula should identify where AI-supported rehearsal adds value and where human interaction is indispensable. Suitable uses include low-stakes practice, vocabulary activation, task repetition, self-review, and preliminary feedback. Unsuitable uses include sole high-stakes scoring, unsupervised collection of sensitive voice data, and substitution for all peer or teacher interaction.

Institutional policy should state approved platforms, data requirements, acceptable assistance, disclosure expectations, and alternative arrangements. Professional development should combine technical familiarity with task design, pronunciation pedagogy, feedback literacy, and ethics.

Research Agenda

Future research should move beyond short-term satisfaction and tool comparison. Randomised or carefully controlled quasi-experimental studies should compare GenAI-supported practice with equally intensive human, non-AI, and blended alternatives. Outcomes should include independent speaking tasks, delayed post-tests, human listener judgements, interactional competence, and willingness to communicate.

Transfer requires particular attention. Studies should test whether learners can use language with unfamiliar human partners, respond to interruption and misunderstanding, and adapt to different accents and pragmatic expectations. Process data should examine how learners request, evaluate, reject, and apply feedback rather than merely recording how often they use a tool.

Fairness research should include underrepresented accents, lower-proficiency learners, speech and hearing disabilities, multilingual practices, and low-bandwidth contexts. Model and platform versions should be reported because system behaviour changes over time. Cost, device access, and data policy should be treated as research variables rather than background details. Finally, researchers should distinguish performance support, task learning, and general speaking development. Claims should match the outcome measured. A better presentation after AI rehearsal is evidence of supported performance; improvement on a new unaided task and delayed assessment is stronger evidence of development.

Limitations of the Review

The review has five limitations. First, the field is developing rapidly, and new systems and studies may appear after the July 2026 search update. Second, the evidence includes peer-reviewed papers, conference manuscripts, and preprints of unequal certainty; publication status was considered in interpretation, but emerging studies may change after review. Third, terminology is inconsistent across conversational agents, chatbots, GenAI, speech LLMs, and automated assessment, which complicates retrieval and comparison. Fourth, the heterogeneity and limited number of robust learner-outcome studies prevented meta-analysis. Fifth, the iterative search and citation-chaining process did not preserve a complete de-duplicated log of raw records, so a PRISMA-style flow count cannot be reported.

The SPEAK framework is an evidence-informed synthesis rather than a validated measurement scale. It should be tested through implementation studies in varied educational and linguistic contexts.

CONCLUSION

Generative AI has created a new form of speaking resource: an available, adaptable, and increasingly voice-capable interaction partner. Current evidence indicates meaningful potential for increasing rehearsal, supporting scenario-based dialogue, reducing immediate fear of judgement, and providing rapid language and pronunciation-related feedback. These benefits are particularly relevant where learners receive limited individual speaking time.

The same evidence shows why technological fluency should not be confused with pedagogical reliability. Small samples, short interventions, prototype designs, uncertain transfer, feedback inaccuracies, accent bias, privacy

risks, and unequal access limit strong claims. Human interaction remains essential for pragmatic judgement, genuine social meaning, cultural responsiveness, and assessment validity.

The SPEAK framework positions GenAI as one component of a teacher-designed progression: Structured tasks, Progressive intelligibility-focused feedback, Equity and ethics, Agency and anxiety-sensitive practice, and Knowledgeable teacher oversight. Used in this way, GenAI can extend speaking pedagogy without replacing the human relationships through which communicative competence ultimately develops.

Declarations

Funding: The author received no specific funding for this study.

Conflict of interest: The author declares no conflict of interest.

Ethical statement: This review is based on published and publicly available literature and did not involve human participants, personal data collection, or an intervention conducted by the author.

Data availability: The publications forming the analytical corpus are identified in the reference list and Appendix A. No new participant dataset was generated.

Author contribution: DMPPK Dasanayake conceptualised the review, conducted the search and synthesis, developed the SPEAK framework, and prepared the manuscript.

Use of generative AI in manuscript preparation: Generative artificial intelligence was used for language refinement, structural support, reference checking, and manuscript formatting. The author independently reviewed and verified the arguments, citations, interpretations, and final wording and accepts full responsibility for the submitted manuscript.

REFERENCES

1. Bendarkawi, J., Ponce, A., Mata, S., Aliu, A., Liu, Y., Zhang, L., Liaqat, A., Rao, V. N., & Monroy-Hernández, A. (2025). ConversAR: Exploring embodied LLM-powered group conversations in augmented reality for second language learners. arXiv. https://arxiv.org/abs/2505.24000
2. Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. Qualitative Research in Psychology, 3(2), 77-101. https://doi.org/10.1191/1478088706qp063oa
3. Bygate, M. (1987). Speaking. Oxford University Press.
4. Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. Applied Linguistics, 1(1), 1-47. https://doi.org/10.1093/applin/I.1.1
5. Carless, D., & Boud, D. (2018). The development of student feedback literacy: Enabling uptake of feedback. Assessment & Evaluation in Higher Education, 43(8), 1315-1325. https://doi.org/10.1080/02602938.2018.1463354
6. Cha, J., Han, J., Yoo, H., & Oh, A. (2024). CHOP: Integrating ChatGPT into EFL oral presentation practice. arXiv. https://arxiv.org/abs/2407.07393
7. Council of Europe. (2020). Common European Framework of Reference for Languages: Learning, teaching, assessment - Companion volume. Council of Europe Publishing. https://www.coe.int/en/web/common-european-framework-reference-languages/cefr-companion-volume-and-its-language-versions
8. Fu, K., Peng, L., Yang, N., & Zhou, S. (2024). Pronunciation assessment with multi-modal large language models. arXiv. https://arxiv.org/abs/2407.09209
9. Horwitz, E. K., Horwitz, M. B., & Cope, J. (1986). Foreign language classroom anxiety. The Modern Language Journal, 70(2), 125-132. https://doi.org/10.1111/j.1540-4781.1986.tb05256.x
10. Iizuka, T., & Mori, H. (2022). How does a spontaneously speaking conversational agent affect user behavior? arXiv. https://arxiv.org/abs/2205.00755
11. Levis, J. M. (2005). Changing contexts and shifting paradigms in pronunciation teaching. TESOL Quarterly, 39(3), 369-377. https://doi.org/10.2307/3588485

12. Li, Y., Chen, C.-Y., Yu, D., Davidson, S., Hou, R., Yuan, X., Tan, Y., Pham, D., & Yu, Z. (2022). Using chatbots to teach languages. arXiv. https://arxiv.org/abs/2208.00376
13. Long, M. H. (1996). The role of the linguistic environment in second language acquisition. In W. C. Ritchie & T. K. Bhatia (Eds.), Handbook of second language acquisition (pp. 413-468). Academic Press.
14. Luoma, S. (2004). Assessing speaking. Cambridge University Press.
15. Ma, R., Qian, M., Tang, S., Bannò, S., Knill, K. M., & Gales, M. J. F. (2025). Assessment of L2 oral proficiency using speech large language models. arXiv. https://arxiv.org/abs/2505.21148
16. MacIntyre, P. D., Clément, R., Dörnyei, Z., & Noels, K. A. (1998). Conceptualizing willingness to communicate in a L2: A situational model of L2 confidence and affiliation. The Modern Language Journal, 82(4), 545-562. https://doi.org/10.1111/j.1540-4781.1998.tb05543.x
17. Meng, F. (2026). Customizing ChatGPT for second language speaking practice: Genuine support or just a marketing gimmick? arXiv. https://arxiv.org/abs/2603.14884
18. Petrovic, J., & Jovanovic, M. (2020). Conversational agents for learning foreign languages: A survey. arXiv. https://arxiv.org/abs/2011.07901
19. Qian, K., Shea, R., Li, Y., Fryer, L. K., & Yu, Z. (2023). User adaptive language learning chatbots with a curriculum. arXiv. https://arxiv.org/abs/2304.05489
20. Swain, M. (1985). Communicative competence: Some roles of comprehensible input and comprehensible output in its development. In S. M. Gass & T. K. Madden (Eds.), Input in second language acquisition (pp. 235-253). Newbury House.
21. Torraco, R. J. (2016). Writing integrative literature reviews: Using the past and present to explore the future. Human Resource Development Review, 15(4), 404-428. https://doi.org/10.1177/1534484316671606
22. UNESCO. (2023). Guidance for generative AI in education and research. UNESCO. https://unesdoc.unesco.org/ark:/48223/pf0000386693
23. Whittemore, R., & Knafl, K. (2005). The integrative review: Updated methodology. Journal of Advanced Nursing, 52(5), 546-553. https://doi.org/10.1111/j.1365-2648.2005.03621.x
24. Woo, J. H., & Choi, H. (2021). Systematic review for AI-based language learning tools. arXiv. https://arxiv.org/abs/2111.04455
25. Xu, S., Qin, L., Chen, T., Zha, Z., Qiu, B., & Wang, W. (2024). Large language model based situational dialogues for second language learning. arXiv. https://arxiv.org/abs/2403.20005

APPENDIX A: ANALYTICAL CORPUS AND ROLE IN THE SYNTHESIS

Source	Design/type	Primary focus	Use in this review
Petrovic & Jovanovic [18]	Survey	Foreign-language conversational agents	Mapped affordances and longstanding limitations of chatbot practice.
Woo & Choi [24]	Systematic review	AI-based language-learning tools	Identified feedback, assessment, adaptation, and teacher-preparation themes.
Li et al. [12]	System/design study	Adaptive language-learning chatbot	Illustrated proficiency adaptation and automated grammar feedback.
Iizuka & Mori [10]	User experiment	Spontaneous vs read-speech agent voice	Showed effects of voice naturalness on response timing and backchannels.
Qian et al. [19]	Design evaluation	Curriculum-adaptive chatbot	Supported curriculum alignment and target-vocabulary practice.
Xu et al. [25]	System study	LLM situational dialogue	Demonstrated scenario control and generalisation to unseen topics.
Cha et al. [6]	Learner/platform evaluation	ChatGPT oral-presentation practice	Provided evidence on personalised feedback, perceptions, and design weaknesses.
Fu et al. [8]	Technical benchmark	Multimodal pronunciation scoring	Demonstrated competitive accuracy/fluency scoring, not learning effects.
Bendarkawi et al. [1]	Prototype user evaluation	LLM-powered AR group conversation	Reported perceived anxiety reduction and autonomy with a small sample.
Ma et al. [15]	Technical benchmark	Speech-LLM oral-proficiency scoring	Showed strong benchmark and cross-task performance; fairness remains open.
Meng [17]	Comparative system analysis	Customised ChatGPT voice modes	Linked customisation with feedback balance and emotional support; cultural gains uncertain.

Table 5: Core analytical corpus and its role in the synthesis.

Note. Seminal SLA, assessment, feedback-literacy, ethics, and review-method sources were used for theoretical and methodological interpretation but are not repeated in this corpus table.