SCIENTIFIC CULTURE

www.sci-cult.com

SCIENTIFIC CULTURE, Vol. 12, No. 2.1, (2026), pp. 10852-10860
Open Access. Online & Print

DOI: 10.5281/zenodo.20088684

ROBOT-ASSISTED PRONUNCIATION LEARNING (RAPL): A SYSTEMATIC REVIEW AND FRAMEWORK FOR ADULT L2 LEARNERS

Mohammad Abdullah Alhammad

Imam Mohammad Ibn Saud Islamic University (IMSIU)
ORCID https://orcid.org/0009-0007-0203-9346

Received: 06/04/2026
Accepted: 29/04/2026

Corresponding Author: Mohammad Abdullah Alhammad (Maalhammad@imamu.edu.sa)

ABSTRACT

Advancements in artificial intelligence and social robotics have begun to reshape second language (L2) instruction, giving rise to Robot-Assisted Language Learning (RALL). Even so, while RALL has been widely explored in relation to vocabulary development and conversational skills, its application to pronunciation remains fragmented and under-theorized. This systematic review synthesizes empirical findings on Robot-Assisted Pronunciation Learning (RAPL) for adult L2 learners and seeks to establish both a conceptual and a practical framework for future research. Drawing on 21 studies published between 2015 and 2025, the review identifies two key instructional paradigms: explicit modeling (robot-led articulation training) and implicit interaction (dialogue-based phonetic exposure), and evaluates their comparative effects on segmental accuracy, prosodic control, learner motivation, and affective engagement. The evidence suggests that humanoid robots equipped with multimodal feedback systems (auditory, visual, and gestural) can improve pronunciation accuracy and reduce speech anxiety by offering nonjudgmental, adaptive feedback. At the same time, persistent methodological limitations, including small sample sizes, short intervention durations, and inconsistent outcome measures, constrain the generalizability of existing findings. Challenges related to automatic speech recognition accuracy, cultural mismatches in prosody modeling, and insufficient personalization further restrict pedagogical impact. The review concludes by proposing a RAPL Framework that integrates embodied interaction, adaptive feedback loops, and affect-sensitive computing to optimize pronunciation learning in adult education, while also identifying critical directions for longitudinal, cross-linguistic validation.

KEYWORDS: Adaptive Feedback; Affective Engagement; Artificial Intelligence; Embodiment; Humanoid Robots; Language Education; Multimodal Feedback; Pronunciation Learning; Robot-Assisted Language Learning (RALL); Robot-Assisted Pronunciation Learning (RAPL).

1. INTRODUCTION

Pronunciation is central to intelligibility, communicative success, and learner confidence in second-language acquisition, yet classroom instruction still tends to treat it as an occasional rather than sustained concern. Over the last decade, social robots and embodied conversational agents have been tested as pronunciation tutors capable of modeling articulations, delivering immediate feedback, and creating low-stakes practice environments that ease performance anxiety. Taken together, this growing body of Robot-Assisted Language Learning (RALL) research offers promising indications for pronunciation teaching, but it remains fragmented, methodologically heterogeneous, and only intermittently focused on phonetic outcomes. A focused, critical synthesis is therefore overdue.

Controlled trials using humanoid platforms (e.g., NAO, Furhat) show that embodied robots can support articulation practice and prosodic training, with some experiments reporting measurable gains in segmental accuracy and prosodic control (Krisdityawan et al., 2022; Iio et al., 2019, 2024). Other studies, however, find no statistically significant advantage over screen-based or computer-mediated training (Amioka et al., 2023; Riedmann et al., 2024). Several studies also point to affective and motivational benefits, including reduced speaking anxiety, stronger engagement, and greater willingness to practice, even when acoustic gains are small or inconsistent. These affective outcomes should not be treated as secondary. Reduced anxiety and greater willingness to practice are direct mediators of long-term pronunciation improvement.

Three enabling features recur where robot-mediated pronunciation succeeds. First, multimodal feedback combining auditory models, visual articulatory cues, and gestures helps learners notice and practice phonetic contrasts. Second, adaptive tutoring informed by learner state supports sustained practice. Third, embodiment, meaning the physical presence of a robot versus a screen agent, can boost motivation and perceived realism. These benefits are real but bounded. Automatic speech recognition (ASR) tuned poorly to L2 phonetic variability remains a pervasive constraint, and novelty effects can inflate short-term engagement measures. Many RALL studies also rely on small samples, brief interventions, and outcome measures emphasizing immediate gains rather than durable transfer.

Beyond technical limitations, deeper conceptual gaps persist. Most RALL work frames pronunciation as a narrow, surface-level skill without engaging phonetic theory or sociophonetic variability, and few studies examine how robot feedback should map onto perceptual stages of phonological acquisition. This paper responds to those gaps by focusing explicitly on Robot-Assisted Pronunciation Learning (RAPL) for adult L2 learners, moving beyond "does it work?" toward "how and for whom does it work?" The review synthesizes empirical findings to compare explicit modeling and implicit interaction paradigms, evaluates how multimodal feedback, embodiment, and affect-sensitive adaptation mediate outcomes, and proposes a practical RAPL Framework linking robotic affordances with prioritized research outcomes: segmental accuracy, suprasegmental control, automaticity, and learner willingness to communicate.

2. LITERATURE REVIEW

2.1. Conceptual Evolution of Robot-Assisted Language Learning

Robot-Assisted Language Learning (RALL) has developed rapidly, driven by advances in artificial intelligence, natural language processing, and social robotics (Belpaeme et al., 2018; Deng et al., 2024). Early studies positioned robots as pedagogical assistants or conversational partners capable of delivering repetitive, affect-sensitive interaction (Belpaeme et al., 2018; Van den Berghe et al., 2018; Kanero et al., 2022; Rohlfing et al., 2022), simulating forms of embodied communicative practice that computer-based systems simply cannot replicate (Belpaeme et al., 2018). Through gaze, gesture, and turn-taking, robots created environments that approximate the social reciprocity of human conversation (Engwall & Lopes, 2022; Skantze, 2021; Jouen et al., 2025).

Yet these contributions came with a notable blind spot. Most studies concentrated on vocabulary acquisition and general communication, treating pronunciation as an incidental variable rather than a central target (Van den Berghe et al., 2018; Zinina et al., 2023; Deng et al., 2024). The large-scale systematic review by Deng et al. (2024) confirmed this pattern, finding that only a small subset of RALL research focused directly on phonetic or prosodic learning outcomes, a striking imbalance given the recognized importance of pronunciation in communicative competence. What this left behind is a conceptual and empirical gap of real consequence: the embodied affordances of robots, particularly their capacity to model articulatory gestures and deliver real-time, multimodal feedback, have never been systematically theorized as tools for pronunciation

learning (Alimardani et al., 2022; Hong et al., 2024; Deng et al., 2024).

2.2. Robots as Pronunciation Tutors: Empirical Findings

The subset of RALL studies that explicitly address pronunciation presents a varied picture, both methodologically and pedagogically. Experimental designs range from explicit articulation training to implicit conversational interaction, with mixed evidence regarding effectiveness.

In explicit paradigms, robots function as articulation tutors offering direct modeling and feedback on specific phonetic targets. Krisdityawan et al. (2022) compared Japanese learners of English who trained with a NAO robot to those using traditional e-learning software. Although no statistically significant group difference emerged, a critical finding that cautions against overstating embodiment effects, the robot group showed improved segmental accuracy and higher engagement, suggesting motivational benefits even in the absence of large acoustic gains. Amioka et al. (2023) found that congruent robot lip movements did not enhance pronunciation accuracy; mismatched or computer-generated visuals produced better outcomes. This counterintuitive finding highlights a key tension in RAPL: embodiment and visual realism can enhance social engagement but may also interfere with perceptual processing when learners cannot interpret visual articulatory cues accurately.

Implicit approaches leverage interactional practice rather than direct phonetic correction. Iio et al. (2019, 2024) demonstrated that extended conversational sessions with humanoid robots improved fluency, rhythm, and pronunciation naturalness, with low-anxiety repetition appearing as the key driver. Khalifa et al, (2016-2019) extended this work through multi-robot dialogue systems, showing that pronunciation could be internalized through simulated multiparty exchanges. Across both paradigms, multimodal feedback emerged as consequential: Zinina et al. (2023) found that combining speech and gesture cues improved learners' tonal accuracy in Mandarin more effectively than voice-only systems, though gestural cues appeared to enhance emotional engagement rather than serving as precise phonetic models.

2.3. Benefits, Limitations, and Theoretical Gaps

The literature converges on several pedagogical advantages. Robots deliver consistent modeling and individualized feedback conducive to self-paced practice (Donnermann et al. (2022)), and their nonjudgmental nature reduces pronunciation anxiety, a well-documented barrier in adult L2 learning (Iio et al., 2024). However, these advantages must be weighed against substantial methodological and theoretical limitations that are often underacknowledged in the literature.

Critically, the evidence base suffers from persistent weaknesses. Most studies rely on small samples, frequently fewer than 30 participants, and short interventions ranging from single sessions to a few weeks, making it difficult to distinguish genuine acquisition from transient performance effects. Outcome measures vary widely: some studies use acoustic analysis, others perceptual ratings, and still others self-report, rendering cross-study comparison unreliable. Novelty effects represent a particular threat to validity, with engagement peaking early in robot interaction and declining once robot responses become predictable (Riedmann et al., 2024), meaning that positive short-term affective scores may not reflect sustained learning conditions.

ASR systems continue to struggle with accented input, frequently misjudging learner pronunciation and providing misleading feedback (Belpaeme et al., 2021). Current systems also overemphasize segmental correction while neglecting suprasegmental features, namely stress, rhythm, and intonation, that contribute most to intelligibility (Derwing & Munro, 2023). At a theoretical level, few RALL studies ground their feedback mechanisms in established models of speech perception, such as Flege's Speech Learning Model or Best's Perceptual Assimilation Model. The field therefore risks remaining empirically rich but theoretically thin, driven by technological novelty rather than linguistic insight.

2.4. Linking the Gaps to the Present Study

This systematic review responds to those shortcomings in three ways. First, it narrows the analytical lens to adult learners, whose cognitive and affective profiles differ markedly from those of children typically studied in RALL. Second, it centers pronunciation, the domain where robotic embodiment should in principle offer the greatest advantage, yet where evidence remains least consolidated. A recent meta-analysis of RALL studies found a medium overall effect size (Lee & Lee, 2022), but effects for pronunciation specifically, and for adult learners in particular, remain poorly characterized. Third, the review moves beyond binary effectiveness questions to ask why certain configurations of instruction and feedback produce better outcomes than others, synthesizing three

design variables, embodiment, feedback modality, and affective adaptation, into a framework that integrates phonetic theory with robotic pedagogy.

3. METHODS

3.1. Research Design

This study adopts a systematic review design guided by the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA 2020) framework. Unlike broader RALL reviews that aggregate vocabulary, grammar, and speaking skills (e.g., Deng et al., 2024), this review isolates studies reporting phonetic, phonological, or prosodic outcomes to develop an evidence-based conceptual framework linking robotic design features with pronunciation performance.

The systematic review was chosen as the most appropriate methodology for two reasons. First, the empirical base for robot-assisted pronunciation learning is characterized by heterogeneous designs, varied populations, and inconsistent outcome measures, conditions under which narrative synthesis is better suited than meta-analysis to preserving the nuance of individual findings. Second, the field currently lacks a consolidated conceptual framework, and the goal here is not only to catalogue findings but to identify structural patterns across studies that can inform future research design and robotic system development. A purely quantitative synthesis would risk imposing false precision on a literature not yet sufficiently standardized to support reliable pooling of effect sizes.

The PRISMA 2020 framework was selected to ensure transparency and methodological rigor. It provides a structured checklist and flow diagram governing each stage of the review process, from database search and duplicate removal through eligibility screening, full-text assessment, and final inclusion. Adherence to this framework allows readers to evaluate the completeness of the search, the consistency of eligibility decisions, and the risk of selection bias, all of which matter considerably given the narrow thematic scope of this review. Because pronunciation-specific RALL studies are relatively rare within the broader literature, precise and reproducible eligibility criteria are essential to avoid both over-inclusion of tangentially related work and under-inclusion through overly restrictive filtering.

The review's theoretical contribution lies in its analytical focus. Where prior systematic reviews of RALL (e.g., Van den Berghe et al., 2018; Deng et al., 2024; Lee & Lee, 2022) have addressed language learning broadly, this review treats pronunciation as a distinct target domain with its own evidential and conceptual logic. Pronunciation learning involves phonetic perception, motor articulation, prosodic sensitivity, and affective risk-taking in ways that do not fully overlap with vocabulary retention or conversational fluency. By isolating studies that operationalize these phonetic dimensions, the review creates conditions for a more precise analysis of which robotic affordances, specifically embodiment level, feedback modality, and affective responsiveness, matter most for pronunciation outcomes and why.

3.2. Search Strategy

A comprehensive search was conducted in Scopus, Web of Science, ERIC, Google Scholar, and SpringerLink covering publications from January 2015 to March 2025. Search strings combined controlled descriptors and free-text terms:

("robot-assisted language learning" OR "RALL" OR "social robot" OR "humanoid robot") AND ("pronunciation" OR "articulation" OR "prosody" OR "speech training" OR "accent improvement") AND ("adult" OR "university students" OR "second-language learners")

Additional backward and forward citation searches were performed using key seed articles (e.g., Krisdityawan et al., 2022; Amioka et al., 2023; Zinina et al., 2023; Iio et al., 2019, 2024) to capture studies not indexed under standardized keywords.

3.3. Inclusion and Exclusion Criteria

Studies were included if they focused on adult or university-aged L2 learners aged 18 years or older, involved the use of a physical or virtual robot for pronunciation, phonetic, or prosodic training, employed an empirical quantitative, qualitative, or mixed-methods design, were published in English in peer-reviewed sources, and reported measurable learning or affective outcomes, such as changes in pronunciation accuracy, intelligibility, prosody, motivation, anxiety, or engagement. Studies were excluded if they dealt solely with vocabulary, grammar, or comprehension, involved child or K–12 participants, were purely theoretical and did not present data, or duplicated the same findings across conference and journal versions.

3.4. Study Selection and Data Extraction

The initial search retrieved 742 records. After removing duplicates (n = 187), 555 titles and abstracts were screened for relevance. Of these, 68 full-text articles were examined, and 21 studies met all inclusion criteria for final synthesis. Screening and

selection were conducted independently by two reviewers, with discrepancies resolved through discussion.

Each study was coded for: (a) publication details; (b) participant demographics; (c) robot type and embodiment level; (d) instructional strategy (explicit vs. implicit); (e) feedback modality; (f) pronunciation focus (segmental, suprasegmental, or combined); and (g) learning outcomes. Effect sizes and significance levels were recorded where reported. It is important to note that many studies did not report standardized effect sizes, precluding formal meta-analysis and limiting the precision with which comparative claims can be made.

3.5. Analytical Approach

Given the heterogeneity of study designs and outcome metrics, a narrative synthesis was employed following Popay et al. (2006). Results were integrated through: (1) descriptive mapping of study characteristics and methodological quality; (2) thematic synthesis of pedagogical and technological patterns; and (3) framework development connecting convergent findings. Risk of bias was assessed qualitatively, with explicit attention to threats including small sample sizes, short intervention durations, novelty effects, and non-standardized outcome measures.

3.6. Ethical Considerations

As a secondary synthesis of previously published data, this review required no human-subject approval. All included studies were screened for evidence of ethical compliance in their original design.

4. RESULTS

The 21 studies included in this review were published between 2015 and 2025 and varied in design, duration, and technological sophistication. Despite these differences, three consistent themes emerged: (1) instructional approach, (2) feedback modality, and (3) affective outcomes. Importantly, the findings must be interpreted with caution given the methodological heterogeneity and limitations of the included studies.

4.1. Explicit vs. Implicit Instructional Approaches

A primary distinction across studies concerns the degree of instructional explicitness, specifically whether the robot served as a direct pronunciation tutor or an interactional partner facilitating spontaneous practice.

Explicit instruction generally produced faster short-term gains in segmental accuracy, but the evidence is qualified by important caveats. Krisdityawan et al. (2022) and Zinina et al. (2023) both employed humanoid robots providing explicit modeling and immediate corrective feedback, with learners showing measurable improvement in targeted phonemes such as /th/, /r/, and /l/ and lexical stress patterns. However, Krisdityawan et al. (2022) found no statistically significant group differences between the robot and e-learning conditions, and follow-up data across studies showed partial regression once robot interaction ceased. This suggests that apparent gains may reflect practiced performance rather than durable acquisition.

Implicit interactional approaches, where learners engaged in task-based conversation with robots, yielded slower but potentially more durable gains in fluency and naturalness. Iio et al. (2019, 2024) demonstrated improvements in rhythm, stress-timing, and connected speech through conversational repetition, and Amioka et al. (2023) reported improved comprehensibility following multi-week interaction despite minimal segmental feedback. These findings are encouraging, but interpreting them is complicated by the absence of control groups in several studies and the reliance on perceptual ratings rather than acoustic measurement.

The contrast between paradigms suggests a trade-off between precision and automatization. Robots functioning as explicit tutors help learners refine phonetic detail but may encourage overreliance on correction; those operating as conversational partners support procedural fluency but may not target persistent articulatory errors. A mature RAPL model therefore requires a hybrid pedagogy, though empirical evidence for this specific integration remains limited.

4.2. Feedback Modalities: Audio, Visual, Gestural, and Multimodal Integration

Feedback modality proved consequential for learner engagement and accuracy outcomes, though the findings resist a simple "more modalities are better" conclusion.

Auditory feedback is the most direct and cognitively accessible channel, with learners consistently rating robotic speech models as useful for identifying phoneme contrasts and prosodic contours (Donnermann et al., 2022). Visual feedback produced more mixed results. Amioka et al. (2023) found that congruent lip synchronization sometimes reduced comprehension, likely because learners

struggled to interpret visual articulatory cues in an unfamiliar language. Engwall and Lopes (2022), by contrast, found that simplified displays showing tongue or lip movement trajectories enhanced articulatory awareness, suggesting that the pedagogical design of visuals, not simply their presence, determines their value.

Gestural feedback appears to operate through a different mechanism, enhancing motivation and attentional focus rather than directly improving accuracy. Zinina et al. (2023) reported that Mandarin learners exposed to a gesturing robot produced more native-like tonal contours, yet perception scores did not significantly change, suggesting that gestures acted as emotional and attentional cues rather than phonetic models. Multimodal systems combining all three channels (e.g., Khalifa et al., 2019; Belpaeme et al., 2021) yielded the most balanced outcomes, but the optimal combination and timing of modalities remain empirically unresolved.

4.3. Affective and Motivational Outcomes

A consistent pattern across studies is that robots exert meaningful affective influence on learners. Adults approaching pronunciation practice with anxiety perceived robots as less threatening than human evaluators, allowing more experimental oral production (Donnermann et al, 2022; Iio et al., 2024). Multiple studies reported reductions in communication apprehension and increases in willingness to speak. Importantly, affective engagement appeared to mediate, rather than substitute for, acoustic learning: learners reporting higher enjoyment and empathy toward the robot achieved stronger gains in several studies.

However, affective advantages are time-limited and methodologically fragile. Riedmann et al. (2024) observed significant engagement decline after several sessions once the robot's dialogue became predictable, a novelty effect that is a fundamental threat to validity in short-term RALL studies. Because most reviewed studies lasted fewer than eight weeks, it is difficult to determine whether reported affective benefits reflect durable shifts in learner attitude or temporary responses to novel technology. Self-report instruments, which dominated affective measurement across studies, are also susceptible to social desirability bias when learners evaluate a robot-instructor they have interacted with repeatedly.

4.4. Integrated Synthesis

Three major insights emerge from these findings, though each carries important qualifications. First, explicit correction accelerates phonetic refinement but lacks demonstrated long-term durability; implicit interaction promotes sustained fluency but may overlook accuracy errors. Second, multimodal feedback supports deeper perceptual-motor learning, provided modalities are pedagogically aligned. Third, learner motivation and comfort moderate the impact of robot interaction on pronunciation gains (Shen et al, 2019). The studies collectively remain methodologically fragmented, with short durations, small samples, and limited theoretical grounding, making it premature to draw strong causal conclusions. The RAPL Framework proposed below is therefore best understood as a theoretically motivated research agenda rather than a validated instructional model.

4.5. Discussion

4.5.1. From Findings to Framework

The synthesis demonstrates that effective robot-assisted pronunciation learning depends on three interdependent dimensions: (1) embodiment and presence, (2) feedback modality and adaptivity, and (3) affective alignment and learner engagement. Existing studies tend to isolate these dimensions rather than analyzing how they reinforce one another. The proposed RAPL Framework integrates them into a single model, positioning pronunciation learning as an embodied, feedback-driven, and affectively mediated process. It must be acknowledged, however, that empirical validation of this integration has yet to occur.

4.5.2 Dimension 1: Embodiment and Presence

Embodiment, comprising the robot's physical or simulated human-like features such as gaze, gesture, and proxemics, shapes learner perception and interaction. Evidence shows that physical presence enhances motivation and encourages practice, but embodiment alone does not guarantee learning (Krisdityawan et al., 2022; Engwall & Lopes, 2022). The framework distinguishes between expressive embodiment (gestures and facial cues that sustain attention) and articulatory embodiment (visible modeling of speech gestures), proposing that effective RAPL design requires aligning embodiment level with learning stage.

It bears emphasis that the assumption of embodiment superiority over screen-based alternatives is not consistently supported. Amioka et al. (2023) found that robot audiovisual speech sometimes underperformed computer-generated alternatives, and Krisdityawan et al. (2022) found no significant advantage for the robot condition overall.

The embodiment benefit may be primarily motivational and context-dependent rather than inherently phonetic.

4.5.3 Dimension 2: Feedback Modality and Adaptivity

Pronunciation learning depends on timely, specific, and multimodal feedback. Within the RAPL Framework, this process is organized through a Feedback Adaptivity Loop (FAL) that operates in three connected stages: detection, diagnosis, and delivery. In the detection stage, the robot's speech-recognition and acoustic-analysis modules identify deviations from target phonemes or prosodic patterns. In the diagnosis stage, the system determines whether the problem arises from perception, such as mishearing, or production, such as misarticulation. In the delivery stage, the robot selects the most appropriate feedback channel, whether auditory modeling, visual articulatory cues, or corrective gesture, based on the learner's performance history and engagement level. Through this adaptive cycle, feedback becomes more than a static correction tool and instead functions as a context-sensitive mechanism.

At present, however, RALL systems only approximate this kind of adaptivity, since few can genuinely integrate acoustic analytics with embodied response in a robust way. The technical viability of real-time and accurate diagnosis of L2 pronunciation errors, especially for accented speech or low-resource language pairs, remains an open challenge. For this reason, the FAL should be understood not as a fully achievable current specification, but as a design target that can guide future development in robot-assisted pronunciation learning.

4.5.4 Dimension 3: Affective Alignment and Engagement

Robots influence not only how learners speak but how they feel about speaking. Within the RAPL Framework, Affective Alignment refers to the robot's capacity to sense and respond to emotional cues such as facial expression, vocal tone, and engagement duration. When the system detects fatigue or frustration, it can modulate feedback tone, simplify tasks, or inject encouragement. This responsiveness is designed to sustain long-term engagement and minimize the novelty drop-off documented in RALL trials.

Affective design must also incorporate sociophonetic awareness: robots using culturally or prosodically inappropriate speech models risk misalignment with learner expectations. The challenge of building culturally calibrated affective systems is substantial and remains largely unaddressed in the literature. Most existing systems use anglocentric prosodic norms, limiting their relevance for learners of other language varieties.

4.5.5 The RAPL Framework in Practice

The three dimensions yield a cyclical learning model: (1) Embodied Interaction, in which the learner engages with a robot capable of both expressive and articulatory embodiment; (2) Adaptive Feedback, in which the system monitors speech output and provides multimodal corrective input through the FAL; and (3) Affective Calibration, in which the robot adjusts behavior to sustain motivation and confidence. This cycle repeats iteratively across two complementary instructional modes: Explicit-Focused Sessions emphasizing phoneme or prosody drills with direct multimodal correction, and Implicit-Interactive Sessions promoting communicative fluency and self-monitoring through dialogue tasks.

4.6. Theoretical and Pedagogical Implications

The RAPL Framework contributes theoretically by linking robotic embodiment with established models of speech learning. It aligns with Flege's Speech Learning Model (SLM-r; Flege & Bohn, 2021) in emphasizing perceptual attunement through multimodal input, and with the Motor Theory of Speech Perception, which connects observed articulatory gestures to phonetic encoding. By embedding linguistic theories into robotic design, RAPL aims to shift the research agenda from demonstration to explanation.

Pedagogically, the framework suggests that robots should be treated as scalable pronunciation partners capable of extending practice beyond class time. Integration with cloud-based ASR and learner analytics could provide longitudinal tracking, enabling instructors to interpret learning trajectories rather than only point-in-time scores. However, this depends on advances in ASR for non-native speech that have not yet been realized at scale.

4.7. Limitations and Future Directions

The RAPL framework faces three challenges. The first concerns technical limitations of current ASR systems, which are not reliable with speakers with accents of their native languages. The second challenge involves the over-automation of teaching by robots; instead, teacher and robot co-teaching models should be employed. Finally, the framework needs to be validated through research studies.

While the framework originated from a small literature, it must be tested.

Future research using this framework can investigate the impact of pronunciation robot assistants on populations with languages other than English and East Asian languages, create pronunciation datasets in different varieties of languages, develop affective sensing algorithms that account for diverse cultural backgrounds, and track language learning over months instead of weeks.

5. CONCLUSION

This systematic review demonstrates that robot-assisted pronunciation learning has matured from an experimental curiosity into a field with genuine pedagogical potential, but that potential remains constrained by the methodological limitations of the current evidence base. Across 21 studies, robots have shown capacity to improve pronunciation accuracy, prosodic awareness, and learner motivation, though these effects are contingent on the orchestration of embodiment, feedback, and affect, and are frequently confounded by novelty effects, small samples, and short interventions that limit causal inference.

Three key insights emerge. First, hybrid instruction combining explicit articulatory guidance with implicit conversational practice addresses both the precision and fluency dimensions of adult pronunciation learning, though this combination has not yet been rigorously tested as a designed curriculum. Second, multimodal feedback works best when delivered adaptively in response to real-time learner performance and affective state. However, the technical requirements for such adaptivity substantially exceed current system capabilities. Third, affective alignment functions as a learning multiplier, though its benefits appear time-limited without ongoing adaptation to counteract novelty decay.

The RAPL Framework proposed here integrates these dimensions into a conceptual model that treats pronunciation as an embodied, adaptive, and affectively mediated process, grounded in phonetic and psycholinguistic theory rather than technological novelty. It should be understood as a theoretically motivated roadmap for future research rather than a validated instructional model: one that requires longitudinal, cross-linguistic, and multi-modal experiments to test its proposed mechanisms. Collaboration between linguists, engineers, and learning scientists remains essential to ensure that robotic systems embody linguistically valid articulatory and prosodic models. If implemented thoughtfully and evaluated rigorously, robot-assisted pronunciation learning offers a pathway toward continuous, data-driven improvement: human in purpose, technological in precision, and phonetically grounded.

Acknowledgement: This work was supported and funded by the Deanship of Scientific Research at Imam Mohammad Ibn Saud Islamic University (IMSIU) (grant number IMSIU-DDRSP2602).

REFERENCES

Alimardani, M., Harinandansingh, J., Ravin, L., & De Haas, M. (2022). Motivational gestures in robot-assisted language learning: A study of cognitive engagement using EEG brain activity. In 2022 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN) (pp. 1393–1398). IEEE. https://doi.org/10.1109/RO-MAN53752.2022.9900508

Amioka, S., Janssens, R., Wolfert, P., Ren, Q., Pinto Bernal, M. J., & Belpaeme, T. (2023). Limitations of audiovisual speech on robots for second language pronunciation learning. In Proceedings of the 2023 ACM/IEEE International Conference on Human-Robot Interaction (HRI) (pp. 359–367). ACM/IEEE. https://doi.org/10.1145/3568294.3580057

Belpaeme, T., Kennedy, J., Ramachandran, A., Scassellati, B., & Tanaka, F. (2018). Social robots for education: A review. Science Robotics, 3(21), eaat5954. https://doi.org/10.1126/scirobotics.aat5954

Belpaeme, T., Vogt, P., Van den Berghe, R., Bergmann, K., Goksun, T., De Haas, M., et al. (2018). Guidelines for designing social robots as second language tutors. International Journal of Social Robotics, 10(3), 325–341. https://doi.org/10.1007/s12369-018-0467-6

Deng, Q., Fu, C., Ban, M., & Iio, T. (2024). A systematic review on robot-assisted language learning for adults. Frontiers in Psychology, 15, 1471370. https://doi.org/10.3389/fpsyg.2024.1471370

Derwing, T. M., & Munro, M. J. (2023). Pronunciation and the second language speaker: Intelligibility in research and practice (2nd ed.). Routledge.

Donnermann, M., Schaper, P., & Lugrin, B. (2022). Social robots in applied settings: A long-term study on adaptive robotic tutors in higher education. Frontiers in Robotics and AI, 9, 831633. https://doi.org/10.3389/frobt.2022.831633

Engwall, O., & Lopes, J. (2022). Interaction and collaboration in robot-assisted language learning for adults. Computer Assisted Language Learning, 35(9), 1273–1309. https://doi.org/10.1080/09588221.2020.1799821

Flege, J. E., & Bohn, O.-S. (2021). The revised Speech Learning Model (SLM-r). In R. Wayland (Ed.), Second language speech learning (pp. 3–83). Cambridge University Press.

Hong, Z. W., Tsai, M. H. M., Ku, C. S., Cheng, W. K., Chen, J. T., & Lin, J. M. (2024). Utilizing robot-tutoring approach in oral reading to improve Taiwanese EFL students' English pronunciation. Cogent Education, 11(1), 2342660. https://doi.org/10.1080/2331186X.2024.2342660

Iio, T., Yoshikawa, Y., Ogawa, K., & Ishiguro, H. (2019). Effects of communication with a robot on second language learning. International Journal of Social Robotics, 11(3), 409–420.

Iio, T., Yoshikawa, Y., Ogawa, K., & Ishiguro, H. (2024). Comparison of outcomes between robot-assisted language learning system and human tutors: Focusing on speaking ability. International Journal of Social Robotics, 16(3), 743–761. https://doi.org/10.1007/s12369-024-01134-0

Jouen, A. L., Clerfeuille, F., Pointeau, G., Le Lec, F., Teissier, V., Jamet, F., & Sugimoto, M. (2025). Once upon a time… Acquisition of second language vocabulary through robotic storytelling in classroom settings. International Journal of Social Robotics. Advance online publication. https://doi.org/10.1007/s12369-025-01253-2

Kanero, J., Oranç, C., Koşkulu, S., Kumkale, T., Göksun, T., & Küntay, A. C. (2022). Are tutor robots for everyone? The influence of attitudes, anxiety, and personality on robot-led language learning. International Journal of Social Robotics, 14(2), 297–312. https://doi.org/10.1007/s12369-021-00789-3

Khalifa, A., Kato, T., & Yamamoto, S. (2016). Joining-in-type humanoid robot assisted language learning system. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016) (pp. 245–249). European Language Resources Association (ELRA).

Khalifa, A., Kato, T., & Yamamoto, S. (2019). Learning effect of implicit learning in joining-in-type robot-assisted language learning system. International Journal of Emerging Technologies in Learning, 14(2), 92–102. https://doi.org/10.3991/ijet.v14i02.9212

Krisdityawan, E., Yokota, S., Matsumoto, A., Chugo, D., Muramatsu, S., & Hashimoto, H. (2022). Effect of embodiment on improving Japanese students' English pronunciation and prosody with a humanoid robot. In 2022 15th International Conference on Human System Interaction (HSI) (pp. 1–6). IEEE. https://doi.org/10.1109/HSI55341.2022.9869469

Lee, H., & Lee, J. H. (2022). The effects of robot-assisted language learning: A meta-analysis. Educational Research Review, 35, 100425.

Popay, J., Roberts, H., Sowden, A., Petticrew, M., Arai, L., Rodgers, M., & Duffy, S. (2006). Guidance on the conduct of narrative synthesis in systematic reviews. ESRC Methods Programme.

Riedmann, P., Gervits, F., De Haas, M., & Belpaeme, T. (2024). Beyond novelty: Long-term engagement in robot-assisted language learning. Frontiers in Robotics and AI, 11, 1345621. https://doi.org/10.3389/frobt.2024.1345621

Rohlfing, K. J., Altvater-Mackensen, N., Caruana, N., van den Berghe, R., Bruno, B., Tolksdorf, N. F., & Hanulíková, A. (2022). Social/dialogical roles of social robots in supporting children's learning of language and literacy: A review and analysis of innovative roles. Frontiers in Robotics and AI, 9, 971749. https://doi.org/10.3389/frobt.2022.971749

Shen, W. W., Tsai, M. H. M., Wei, G. C., Lin, C. Y., & Lin, J. M. (2019). ETAR: An English teaching assistant robot and its effects on college freshmen's in-class learning motivation. In L. Ronningsbakk, T. T. Wu, F. Sandnes, & Y. M. Huang (Eds.), Innovative technologies and learning: ICITL 2019 (Lecture Notes in Computer Science, vol. 11937, pp. 77–86). Springer. https://doi.org/10.1007/978-3-030-35343-8_9

Skantze, G. (2021). Turn-taking in conversational systems and human-robot interaction: A review. Computer Speech & Language, 67, 101178. https://doi.org/10.1016/j.csl.2020.101178

Van den Berghe, R., Verhagen, J., Oudgenoeg-Paz, O., van der Ven, S., & Leseman, P. (2018). Social robots for language learning: A review. Review of Educational Research, 89(2), 259–295. https://doi.org/10.3102/0034654318821286

Zinina, A., Kotov, A., Arinkin, N., & Gureyeva, A. (2023). Can a robot companion help students learn Chinese tones? The role of speech and gesture cues. In A. G. Kravets, M. V. Shcherbakov, & P. P. Groumpos (Eds.), Creativity in intelligent technologies and data science: CIT&DS 2023 (Communications in Computer and Information Science, vol. 1909). Springer. https://doi.org/10.1007/978-3-031-44615-3_29.