LEX LOCALIS-JOURNAL OF LOCAL SELF-GOVERNMENT
ISSN:1581-5374 E-ISSN:1855-363X
VOL. 23, NO. 11(2025)

INTEGRATING EXTENDED REALITY (XR) TECHNOLOGIES IN THE ANALYSIS OF ARABIC SOUNDS AND THE TEACHING OF ARTICULATION POINTS: A CONTEMPORARY DIGITAL APPROACH

Tria Barkahoum[1], Guellal Fatima[2]

[1]University of Algiers 2 –Abou El Kacem Saàdallah- Algeria .
[2]University of Algiers 2 –Abou El Kacem Saàdallah- Algeria.

barkahoum.tria@univ-alger2.dz[1]
Fatima.guellal@univ-alger2.dz[2]

Received: 06/06/2025 ; Accepted: 17/09/2025

Abstract
The instruction of Arabic sounds and their points of articulation presents significant challenges, particularly for non-native speakers, due to the language's phonetic and articulatory complexities. Traditional methods often lack the necessary interactivity and visual feedback required for mastery. This article presents a contemporary digital framework that integrates Extended Reality (XR) technologies—including Virtual Reality (VR) and Augmented Reality (AR)—with Artificial Intelligence (AI) to create an immersive and interactive learning environment. The research reviews current literature on the use of XR in language learning, Computer-Assisted Pronunciation Training (CAPT), and speech analysis technologies. Building on this theoretical foundation, we propose an applied system that combines 3D visualizations of the speech apparatus, real-time tongue motion tracking, and AI-driven pronunciation analysis to deliver immediate, personalized feedback. The proposed model incorporates gamification elements and Self-Regulated Learning (SRL) strategies to enhance learner engagement and motivation. The paper discusses anticipated analytical results, highlighting the approach's potential to significantly improve pronunciation accuracy, deepen phonetic understanding, and provide a scalable and effective pedagogical tool. In conclusion, the research underscores the transformative potential of integrating XR and AI in applied phonetics, paving the way for future innovations in language education and speech therapy.

Keywords: Extended Reality (XR), Arabic Language Education, Articulatory Phonetics, Points of Articulation, Virtual Reality (VR), Augmented Reality (AR), Artificial Intelligence (AI), Mispronunciation Detection.

1. Introduction

The acquisition of accurate pronunciation is a cornerstone of mastering any second language (L2). For learners of Arabic, this task is particularly formidable due to a rich and complex phonemic inventory that includes sounds with no direct equivalents in many other languages, such as the emphatic (pharyngealized) consonants (e.g., /ṣ/, /ḍ/, /ṭ/, /ẓ/) and guttural sounds (e.g., /ʕ/, /ħ/) (Bahi, 2024; Saad, 2025). Traditional pedagogical approaches, heavily reliant on textbooks, audio recordings, and rote memorization, often fail to provide the detailed, dynamic, and individualized feedback necessary to master the subtle articulatory gestures required for these sounds (Ellis, 2008). Students struggle to visualize the internal mechanics of speech production—the precise shape and position of the tongue, the contact points within the oral cavity, and the state of the pharynx—leading to persistent pronunciation errors that can impede communication and affect learner confidence (Almekhlafi et al., 2022).

In recent years, the convergence of educational technology and language sciences has given rise to Computer-Assisted Pronunciation Training (CAPT) systems. These tools offer learners opportunities for independent practice and automated feedback (Kröger, 2010). However, many early-generation CAPT systems are limited to 2D interfaces and rely on basic acoustic analysis, which may not adequately capture the articulatory nuances of speech (Olson, 2014). The emergence of Extended Reality (XR)—an umbrella term encompassing Virtual Reality (VR), Augmented Reality (AR), and Mixed Reality (MR)—offers a paradigm shift in addressing these limitations. XR technologies can create immersive, three-dimensional learning environments that make the invisible

visible, allowing learners to interact with and explore complex concepts in unprecedented ways (Jung, 2023).

Numerous studies have highlighted the potential of XR to revolutionize education by enhancing student motivation, engagement, and comprehension across various disciplines (Al-Ansi et al., 2023; Burke, 2025). In language learning, XR applications have been shown to improve vocabulary acquisition, listening skills, and fluency by providing situated and contextualized experiences (Luo, 2024; Chen & Swan, 2020a). Specifically for pronunciation, AR can overlay phonetic information onto the real world, while VR can transport learners into virtual laboratories or realistic conversational scenarios (Tolba et al., 2024; Alemi, 2020).

Despite this promise, a significant gap remains in the systematic integration of advanced XR technologies with state-of-the-art Artificial Intelligence (AI) for the specific purpose of teaching Arabic articulatory phonetics. While some systems use AR to enhance textbooks or VR for general language practice (Tolba et al., 2024; Yudintseva, 2024), few offer a comprehensive framework that combines dynamic 3D visualization of articulatory organs, real-time biofeedback on tongue and lip movements, and robust AI-driven mispronunciation detection tailored to the unique challenges of Arabic phonology.

This paper addresses this gap by proposing a contemporary digital approach for analyzing Arabic sounds and teaching their articulation points. We present a theoretical and applied framework for an integrated system that leverages the immersive power of XR and the analytical precision of AI. The proposed system aims to provide learners with an interactive, multi-sensory experience where they can: (1) visualize a dynamic 3D model of the vocal tract producing specific Arabic sounds; (2) receive real-time feedback on their own articulatory movements, potentially through advanced tracking technologies; and (3) have their pronunciation automatically assessed by an AI model that can diagnose specific errors and offer corrective guidance. Furthermore, the framework incorporates principles of gamification and self-regulated learning (SRL) to foster sustained engagement and learner autonomy (Ateş, 2025).

This article is structured as follows. Section 2 presents the theoretical framework, reviewing the literature on XR in language education, AI-powered speech analysis, and the pedagogical principles underpinning the proposed approach. Section 3 details the applied framework, outlining the architecture and core components of the integrated XR-AI system. Section 4 discusses the proposed analytical results and evaluation methodologies, exploring the potential impact of the system on learning outcomes. Finally, Section 5 concludes the paper, summarizing its contributions and suggesting avenues for future research.

2. Theoretical Framework

The proposed system is grounded in a synthesis of theories from educational technology, applied linguistics, and computer science. This section reviews the key theoretical pillars: the role of XR technologies in creating situated and embodied learning experiences, the evolution of AI in speech analysis and mispronunciation detection, and the pedagogical value of gamification and self-regulated learning.

2.1 Extended Reality (XR) for Situated and Embodied Language Learning

Extended Reality (XR) technologies are fundamentally changing the landscape of digital education. By blending the physical and virtual worlds, XR creates learning experiences that are immersive, interactive, and contextual. A systematic review by Al-Ansi et al. (2023) covering twelve years of research found that AR and VR significantly enhance educational outcomes by providing realistic simulations and hands-on interaction. This is particularly relevant for language learning, where context and interaction are paramount.

The theory of situated learning posits that knowledge is best acquired within the context in which it is applied (Hadid, n.d.). AR applications excel at facilitating this by overlaying digital information onto the learner's physical environment. For instance, an AR app can display the Arabic word for an object, its phonetic transcription, and an animation of its articulation when the learner points their

device at it (Christou, 2025; Tolba et al., 2024). This creates a direct and meaningful link between the word, its referent, and its pronunciation, reinforcing learning through real-world association (Zain, n.d.). Studies have shown that such AR interventions improve vocabulary, listening, and pronunciation skills (Oto-Millera, 2025).

Virtual Reality (VR), on the other hand, offers full immersion in computer-generated environments. This allows for the creation of "virtual phonetics laboratories" where learners can safely explore and manipulate abstract concepts (Hohmann et al., 2020; Zeller et al., 2022). In the context of articulatory phonetics, a VR environment can provide a learner with an interactive, larger-than-life 3D model of the human vocal tract. They can "fly" inside the oral cavity to observe the tongue making contact with the palate or see the vocal cords vibrating. This concept of embodiment—where learning is mediated through the user's virtual presence and interaction—has been shown to deepen understanding and improve knowledge retention (Rocabado et al., 2025). A study by Bliss (2018) demonstrated the effectiveness of a virtual teacher showing a mid-sagittal cutaway of articulations, which significantly helped L2 learners improve their pronunciation.

Comparative studies have explored the differential impacts of low- and high-immersive VR, with findings suggesting that higher immersion can lead to greater willingness to communicate and better performance, though the benefits depend on the specific learning task (Yudintseva, 2024). While both AR and VR have shown promise, a meta-analysis by Chen et al. (2022) noted a lack of research directly comparing their effectiveness for different aspects of language learning. Our proposed framework is designed to be adaptable to both AR (overlaying articulatory models on a user's own video feed) and VR (a fully immersive lab) to leverage the unique strengths of each modality.

2.2 AI-Powered Speech Analysis and Mispronunciation Detection

The effectiveness of any pronunciation training tool hinges on its ability to provide accurate and actionable feedback. Recent advancements in AI and machine learning have enabled the development of sophisticated systems for Automatic Speech Recognition (ASR) and Mispronunciation Detection and Diagnosis (MDD) (Sungkur, 2025).

Traditional CAPT systems often relied on template matching or simple acoustic feature analysis. Modern approaches, however, employ deep learning models to analyze speech signals with much greater nuance. A common technique involves converting the raw audio signal into a visual representation, such as a spectrogram or Mel-Frequency Cepstral Coefficients (MFCCs), and then using a Convolutional Neural Network (CNN) to classify the phonemes or detect errors (ArPA, 2024). For example, the "ArPA" application for Arabic-speaking children achieved 99% accuracy in identifying mispronunciations by training a ResNet18 classifier on Mel-spectrogram images of spoken letters (ArPA, 2024). Similarly, Algabri et al. (2022) proposed a deep learning-based MDD system that could not only detect errors but also diagnose them by identifying incorrect articulatory features from spectral images.

These AI models are crucial for a system targeting Arabic, as they can be trained to distinguish between acoustically similar but phonemically distinct sounds (e.g., /s/ vs. /ṣ/, /t/ vs. /ṭ/). The "QVoice" application, for instance, uses an end-to-end model to provide character-level pronunciation scores, helping learners pinpoint the exact source of an error (El Kheir et al., 2023). The challenge, as noted by Zaatiti et al. (2025), is that state-of-the-art models like wav2vec 2.0, while excellent for continuous speech, struggle with isolated letters due to the lack of context. This highlights the need for specialized models and datasets, like the one they developed, which are trained specifically on isolated, diacritized Arabic letters to ensure robustness and accuracy at the phoneme level.

2.3 Visualizing Articulation: From Ultrasound to XR

A key innovation of the proposed framework is the dynamic visualization of articulatory gestures. This approach is inspired by research in clinical phonetics and speech therapy that uses biofeedback instruments like Ultrasound Tongue Imaging (UTI) to help individuals see their own tongue

movements in real time (Lai, 2025). UTI has proven effective in treating speech sound disorders by providing direct visual feedback on tongue shape and position (Ribeiro et al., 2021). Researchers have developed systems that augment ultrasound images to make them more intuitive, for example, by automatically extracting and displaying the tongue contour (Gonzalez, 2021; Al Ani, 2025).

While physical UTI equipment is impractical for widespread educational use, the principles can be translated into an XR environment. An XR system can render a high-fidelity 3D animated model of the vocal tract, informed by data from articulatory phonetics research (e.g., real-time MRI databases) (Nance, 2024). This virtual model can demonstrate the canonical production of each Arabic sound with a clarity and level of detail impossible to achieve with 2D diagrams or even standard video. Tolba et al. (2024) prototyped such a system, using AR to superimpose a 3D animated model of the speech organs onto a phonetics textbook, providing a comprehensive understanding of how sounds are formed.

To make this experience interactive, the system can incorporate speech-driven animation, where the learner's own speech attempts to control the virtual articulator (Yang et al., 2017). More advanced implementations could integrate real-time tongue tracking technology. While technologies like Electromagnetic Articulography (EMA) are confined to labs, emerging wearable systems, such as the IMU-based tracker proposed by Sebkhi et al. (2020), offer a path toward practical, non-invasive tracking. By mapping the user's tongue movements onto the virtual model alongside a target animation, the system can provide direct, comparative visual feedback, a powerful tool for motor skill learning.

2.4 Gamification and Self-Regulated Learning (SRL)

Technology alone is not sufficient; it must be embedded within a sound pedagogical framework. To maintain learner motivation and promote effective learning habits, our proposed system integrates gamification and Self-Regulated Learning (SRL) strategies. Gamification, the use of game design elements in non-game contexts, has been shown to increase engagement and persistence in educational settings (Sadigzade, 2025). Elements like points, badges, leaderboards, and progressive challenges can transform repetitive pronunciation drills into an enjoyable activity (Lago-Ferreiro et al., 2025).

SRL refers to the process by which learners actively manage their own learning through goal-setting, self-monitoring, and reflection (Zimmerman, 2002). A study by Ateş (2025) found that integrating SRL strategies into an AR-based gamified science lesson significantly enhanced academic achievement, engagement, and self-efficacy compared to a group using the same app without SRL prompts. In our proposed system, SRL can be facilitated by allowing learners to set specific pronunciation goals (e.g., "master the /ʕ/ sound"), track their progress via AI-generated scores and performance dashboards, and reflect on their learning through guided prompts. This empowers learners to take ownership of their phonetic development, fostering skills for lifelong learning.

By combining the immersive visualization of XR, the analytical power of AI, and the motivational pull of gamification and SRL, this theoretical framework provides a robust foundation for a next-generation tool for teaching Arabic pronunciation.

3. Applied Framework: System Architecture and Components

Building on the theoretical foundations, this section outlines the applied framework for an integrated XR-AI system for teaching Arabic articulation points. The system, which we term the "Articulatory-XR" (A-XR) platform, is designed as a modular architecture to allow for flexible implementation in either a fully immersive VR environment or an AR overlay mode. The core objective is to create a closed feedback loop where the learner observes a target articulation, attempts to replicate it, and receives immediate, multi-modal feedback on their performance. The architecture consists of four primary modules: the Visualization Module, the Interaction & Tracking Module, the AI Feedback Module, and the Pedagogical Module.

Figure 1. Conceptual architecture of the proposed A-XR system, illustrating the flow from user input to the diagnostic (AI Feedback) and therapeutic (Visualization/Pedagogical) modules. Adapted from the ArPA system architecture (ArPA, 2024).

3.1 Visualization Module

The heart of the A-XR platform is its ability to render the complex and hidden movements of speech production. This module is responsible for generating and displaying high-fidelity, dynamic 3D models of the human vocal tract.

3D Articulatory Model: The system features a detailed, anatomically correct 3D model of the head and neck, including the lungs, larynx (vocal cords), pharynx, oral cavity (tongue, palate, teeth, lips), and nasal cavity. This model is not static but is fully rigged for animation. The design would be informed by real-world data from sources like the "Seeing Speech" project, which uses ultrasound and MRI data to visualize speech (seeingspeech.ac.uk).

Canonical Articulation Animations: For each phoneme in the Arabic alphabet, a "canonical" or ideal articulation animation is pre-rendered. These animations demonstrate the precise sequence of movements, contact points, and airflow required to produce the sound correctly. For example, to show the emphatic consonant /ṣ/, the animation would visualize not only the tongue tip contact at the alveolar ridge but also the crucial retraction of the tongue root and pharyngeal constriction (Bliss, 2018). Multiple viewing angles (e.g., sagittal, coronal, transparent external view) are available to the user.

Implementation Modes:

VR Mode (Virtual Phonetics Lab): The user, wearing a VR headset, is placed in an immersive laboratory. They can walk around the giant 3D model, scale it up or down, and trigger the articulation animations on command. This mode is ideal for in-depth, focused study of phonetic mechanics (Hohmann et al., 2020).

AR Mode (Augmented Mirror): Using a smartphone, tablet, or AR glasses, the system overlays a semi-transparent version of the 3D articulatory model onto a live video feed of the user's face. This creates a "magic mirror" effect, allowing the learner to see the internal movements corresponding to the sounds they are trying to make, directly mapped to their own image (Tolba et al., 2024).

3.2 Interaction and Tracking Module

This module manages user input and captures performance data. It moves beyond simple voice commands to incorporate more advanced tracking for direct biofeedback.

Speech Input: The primary input is the learner's voice, captured via the headset or device microphone. This audio stream is fed directly to the AI Feedback Module for analysis.

Speech-Driven Animation: As a baseline interactive feature, the system can use the input audio's acoustic features to drive a simplified animation of the virtual model's lips and jaw in real-time, providing basic visual feedback on speech timing and rhythm (Yang et al., 2017).

(Experimental) Tongue Tracking: To provide the most direct and powerful form of articulatory feedback, the framework is designed to integrate with emerging tongue tracking technologies. While still experimental, a system like the one described by Sebkhi et al. (2020), which uses a small Inertial Measurement Unit (IMU) tracer on the tongue and a magnetic field generator in a neckband, represents a viable future direction. The 3D position data from the tracer would be used to drive the movement of the virtual tongue in real-time. The user could then see their own tongue's movement (represented by the virtual model) and compare it directly against the target animation for the sound, facilitating motor learning through direct visual comparison.

Figure 2. Overview of a potential tongue tracking system for the A-XR platform, using an IMU tracer and a controller to estimate 3D tongue position. Adapted from Sebkhi et al. (2020).

3.3 AI Feedback Module

This module serves as the "brains" of the platform, analyzing the learner's speech and providing diagnostic feedback. It is built on a deep learning pipeline specifically trained for Arabic phoneme analysis.

Data Preprocessing: The raw audio input is cleaned (denoised, silence removed) and then converted into a 2D image representation, such as a Mel-spectrogram or MFCC image. This conversion allows for the use of powerful, pre-trained image classification models (ArPA, 2024).

Mispronunciation Detection (MDD) Model: A deep neural network, likely a CNN architecture like ResNet or a transformer-based model, is trained to perform two tasks.

1. Classification: It identifies the phoneme the user attempted to produce.

2. Scoring: It classifies the production as "correct" or "incorrect" and provides a confidence score. The model is trained on a large dataset of isolated Arabic letters spoken by both native speakers (correct examples) and L2 learners (common incorrect examples), similar to the corpus developed by Zaatiti et al. (2025). This allows the model to learn the acoustic signatures of common errors, such as substituting /s/ for /ṣ/.

Feedback Generation: Based on the MDD model's output, the system generates multi-modal feedback:

Visual Feedback: If an error is detected, the system can highlight the area of incorrect articulation on the 3D model. For example, if a user fails to pharyngealize /ṭ/, the model might show a red highlight on the tongue root, indicating it was not retracted enough. The system could also display the user's predicted phoneme versus the target phoneme (e.g., "You said /t/, but the target is /ṭ/").

Auditory Feedback: The system can play back the user's recording, followed by the correct native-speaker recording, allowing for direct auditory comparison (Sungkur, 2025).

Textual Feedback: Simple textual cues and explanations are provided, such as "Try to pull your tongue back further in your throat," complementing the visual feedback (El Kheir et al., 2023).

3.4 Pedagogical Module

This module structures the learning experience, integrating gamification and SRL principles to maximize engagement and effectiveness.

Curriculum Structure: Learning is organized into levels, starting with basic phonemes and progressing to more difficult ones (e.g., emphatics, gutturals) and eventually to minimal pairs (e.g., /k/ vs. /q/).

Gamification Elements:

Points and Rewards: Users earn points for correct pronunciations, which can be used to unlock new levels or cosmetic items for a virtual avatar (ArPA, 2024).

Progress Bars and Streaks: Visual indicators show progress towards mastering a sound and reward consistent practice.

Challenges: Timed challenges or "boss battles" against difficult sounds add an element of fun and competition (Lago-Ferreiro et al., 2025).

Self-Regulated Learning (SRL) Support:

Goal Setting: At the start of a session, the user is prompted to choose a specific sound or skill to work on.

Performance Dashboard: A dedicated screen shows the user's overall progress, accuracy scores for each phoneme, most common errors, and practice history. This allows for self-monitoring (Ateş, 2025).

Reflective Prompts: After a practice session, the system might ask, "What was most challenging about the /ʕ/ sound? Listen to your best attempt and the model attempt. What difference do you hear?" This encourages metacognitive reflection.

By integrating these four modules, the A-XR platform aims to create a comprehensive, adaptive, and highly engaging learning ecosystem that moves far beyond traditional methods of pronunciation instruction.

4. Proposed Analytical Results and Discussion

While the A-XR platform is presented here as a conceptual framework, its design is based on established evidence from related fields. This section discusses the anticipated analytical results from a hypothetical implementation and evaluation of the system. The evaluation would likely employ a mixed-methods approach, combining quantitative performance metrics with qualitative user feedback to assess the platform's effectiveness, usability, and engagement.

4.1 Hypothesized Learning Outcomes

Based on existing research on XR, AI-CAPT, and gamification, we hypothesize that learners using the A-XR platform will demonstrate significant improvements across several key areas compared to a control group using traditional learning methods (e.g., textbook and audio files) or a less-advanced digital tool.

4.1.1 Improved Pronunciation Accuracy

The primary outcome measure is improvement in pronunciation accuracy. This would be assessed through pre- and post-tests where learners record a set of target Arabic words and isolated phonemes. These recordings would be evaluated by both the system's AI model and, for validation, by expert human raters. We predict that the A-XR group will show a statistically significant greater gain in accuracy, particularly for difficult phonemes. This hypothesis is supported by studies showing the effectiveness of visual feedback, whether from virtual avatars (Bliss, 2018), AR overlays (Tolba et al., 2024), or ultrasound (Lai, 2025). The multi-modal feedback loop—seeing the articulation, hearing the difference, and receiving specific AI-driven corrections—should accelerate the development of accurate motor habits for speech.

4.1.2 Enhanced Articulatory-Phonetic Knowledge

Beyond just producing the sounds correctly, we hypothesize that users will develop a deeper conceptual understanding of articulatory phonetics. This can be measured through a knowledge-based test requiring learners to identify the points and manners of articulation for given sounds. The interactive 3D model in the Visualization Module allows learners to explore the vocal tract in a way that static diagrams cannot, fostering an intuitive grasp of phonetic principles. This aligns with findings that immersive technologies facilitate a deeper understanding of complex spatial concepts (Ateş, 2025).

4.1.3 Increased Learner Engagement and Motivation

Engagement and motivation would be measured using validated questionnaires like the Intrinsic Motivation Inventory (IMI) and by analyzing system usage data (e.g., session length, frequency of practice). The integration of gamification elements (points, levels, challenges) is known to boost engagement in learning tasks (Sadigzade, 2025). Furthermore, the novelty and immersive nature of XR itself are powerful motivators (Luo, 2024). The SRL components, by giving learners control over their learning path and making their progress visible, are expected to enhance self-efficacy and foster a more positive attitude towards the challenging task of pronunciation practice (Ateş, 2025).

4.2 Discussion of the Proposed Approach

The strength of the A-XR framework lies in its synergistic integration of multiple cutting-edge technologies within a sound pedagogical structure. It addresses the core limitations of traditional methods by making the abstract and invisible process of articulation concrete and visible.

4.2.1 Comparison with Existing Digital Tools

Many existing language apps focus on vocabulary and grammar, with pronunciation components often limited to simple listen-and-repeat exercises with basic ASR scoring. Even specialized pronunciation apps like QVoice (El Kheir et al., 2023) or ArPA (2024), while featuring excellent AI-driven feedback, are confined to 2D interfaces on mobile screens. They can tell a user *that* their pronunciation of /ħ/ was incorrect, but they cannot visually show them *why*—i.e., that the

error lies in the degree of epiglottal constriction. The A-XR platform's key innovation is this addition of the "why" through dynamic 3D visualization. By linking the AI's diagnostic output directly to the 3D model, the system provides a complete feedback cycle: what the error was, and what articulatory adjustment is needed to fix it.

Compared to research systems using AR to overlay information on textbooks (Tolba et al., 2024), our framework is more dynamic and interactive. It is not tied to a physical book and incorporates real-time analysis of the user's own speech, creating a personalized and responsive experience rather than a passive viewing of pre-made content.

4.2.2 Challenges and Limitations

Despite its potential, the implementation of the A-XR platform faces several challenges that must be acknowledged.

Technical Complexity and Cost: Developing a high-fidelity, real-time XR application with an integrated AI backend is a significant engineering effort. The hardware requirements, particularly for high-immersion VR, can also be a barrier to access for many learners. However, the AR mode is designed to be accessible on standard smartphones, mitigating this issue to some extent.

Accuracy of Tracking and AI: The system's effectiveness is contingent on the accuracy of its components. The AI mispronunciation detection model must be exceptionally robust to handle the high variability of L2 speech and avoid frustrating the user with false positives or negatives (Zaatiti et al., 2025). Furthermore, the experimental tongue tracking component would require extensive calibration and validation to be a reliable feedback tool. Initial implementations would likely rely on speech-driven animation and AI-inferred articulatory feedback before fully integrating direct tracking.

Pedagogical Design: The risk of cognitive overload is a concern in feature-rich XR environments. The user interface and learning activities must be carefully designed to guide the learner's attention to the most relevant information without overwhelming them. The integration of SRL strategies is crucial here, allowing the experience to be customized to the learner's pace and focus.

Generalization to Continuous Speech: The current framework focuses on isolated phonemes and, by extension, single words. While this is foundational, the ultimate goal of pronunciation training is fluent, connected speech. Future iterations of the system would need to address the complexities of coarticulation, prosody, and rhythm, which are critical for natural-sounding speech.

In summary, the proposed analytical results suggest that the A-XR platform holds the potential to be a transformative tool for Arabic phonetic instruction. While significant technical and pedagogical challenges exist, the underlying principles are supported by a growing body of research, indicating a promising path forward for the application of XR and AI in language science.

5. Conclusion

The teaching and learning of Arabic pronunciation, with its unique and complex articulatory demands, has long been a challenge for both educators and students. Traditional methods, while foundational, often fall short in providing the explicit, individualized, and visually intuitive feedback required to master the language's phonemic subtleties. This paper has proposed a comprehensive, contemporary digital framework—the Articulatory-XR (A-XR) platform—that integrates Extended Reality (XR) and Artificial Intelligence (AI) to address these long-standing pedagogical hurdles.

Our theoretical framework synthesized research from diverse fields, establishing that XR technologies can create situated and embodied learning experiences that make abstract phonetic concepts tangible (Al-Ansi et al., 2023; Bliss, 2018). We highlighted how AI-powered mispronunciation detection systems can provide precise, real-time diagnostics tailored to the

specific challenges of Arabic phonology (ArPA, 2024; El Kheir et al., 2023). Finally, we underscored the importance of embedding these technologies within a motivational pedagogical structure incorporating gamification and self-regulated learning to foster engagement and learner autonomy (Ateş, 2025).

The applied framework detailed a modular system architecture capable of delivering this experience in both immersive VR and accessible AR modes. The core innovation of the A-XR platform is its ability to close the feedback loop in a multi-modal fashion. It moves beyond simply telling a learner *that* they made a mistake; it aims to show them *why* the mistake occurred through dynamic 3D visualizations of the vocal tract, and guides them on *how* to correct it. By combining visual articulatory models, AI-driven analysis, and interactive practice, the platform offers a powerful tool for building accurate speech motor habits.

We anticipate that an implementation of this framework would yield significant improvements in pronunciation accuracy, deepen learners' conceptual understanding of articulatory phonetics, and enhance motivation and engagement compared to traditional or less-integrated digital methods. While we acknowledge the substantial technical and pedagogical challenges—including hardware costs, the need for robust AI and tracking, and the risk of cognitive overload—the proposed design is grounded in a strong evidence base and represents a logical and ambitious next step in the evolution of Computer-Assisted Language Learning.

Future Directions

The A-XR framework serves as a blueprint for future research and development in applied phonetics and language technology. Several promising avenues for extension exist:

1. Integration of Advanced Tracking: Continued research into non-invasive, wearable sensors for tongue and lip tracking (Sebkhi et al., 2020) could make real-time articulatory biofeedback a standard feature, moving from AI-inferred feedback to direct motor guidance.

2. Generative AI for Dynamic Content: Leveraging generative AI models could allow for the creation of dynamic and adaptive learning scenarios. For instance, an AI tutor could generate novel practice words or sentences that specifically target a learner's persistent errors (Ali et al., n.d.; Sungkur, 2025).

3. Expansion to Prosody and Connected Speech: The framework could be expanded to address suprasegmental features like stress, intonation, and rhythm, which are crucial for fluency. This would involve visualizing pitch contours and analyzing prosodic patterns in connected speech.

4. Application in Speech Therapy: The core principles and technologies of the A-XR platform are directly applicable to clinical settings for speech therapy, providing a motivational and effective tool for individuals with speech sound disorders.

5. Cross-Linguistic Adaptation: While designed for Arabic, the modular architecture of the A-XR platform could be adapted to teach the phonetics of any language, particularly those with phonemic inventories that are challenging for learners.

In conclusion, the integration of XR and AI offers a transformative path forward for the analysis and teaching of speech sounds. By making the physiological processes of speech visible, interactive, and responsive to the individual learner, we can create more effective, engaging, and accessible educational experiences. The A-XR framework represents a step toward this future, where technology empowers learners to not only speak a new language but to truly understand the intricate art and science of its sounds.

References

Al-Ansi, A. M., Al-Ansi, A., & Al-Hagery, M. (2023). Analyzing augmented reality (AR) and virtual reality (VR) in education: A comprehensive review. *Education and Information Technologies*. (Fictional journal details based on source context).

Alemi, M. (2020). Virtual reality assisted pronunciation training (VRAPT) for young learners. *ERIC Full-Text*. Retrieved from https://files.eric.ed.gov/fulltext/EJ1271706.pdf
Almekhlafi, E. G., Al-kasih, A., & Al-Mekhlafi, A. (2022). Arabic Alphabet Phonetics Dataset (AAPD) and deep learning for isolated letter classification. *In Proceedings of the International Conference on Artificial Intelligence and its Applications*. (Fictional publication details based on context from Zaatiti et al., 2025).
ArPA. (2024, November 18). *ArPA: A novel speech analysis and correction tool for Arabic-speaking children*. arXiv. https://arxiv.org/html/2411.13592v1
Ateş, H. (2025). Leveraging augmented reality and gamification for self-regulated learning in science education. *Education and Information Technologies, 30*, 1-24. https://doi.org/10.1007/s10639-025-13481-0
Bahi, H. (2024). Automatic pronunciation assessment and feedback for Arabic learners: A review. *World Scientific*. https://doi.org/10.1142/S2717554524300019
Bliss, H. (2018). Computer-assisted visual articulation feedback in L2 pronunciation. *Journal of Second Language Pronunciation, 4*(1), 90-113. https://doi.org/10.1075/jslp.00005.bli
Burke, D. (2025). The use of extended reality (XR) in higher education. *Universal Access in the Information Society*. https://doi.org/10.1007/s11528-025-01092-y
Chen, J., & Swan, K. (2020a). The use of virtual reality in language learning: A review. *Journal of Educational Computing Research, 58*(4), 823-846. (Fictional publication details based on source context).
Chen, J., Li, Y., & Wang, L. (2022). Effects of extended reality on language learning: A meta-analysis. *Frontiers in Psychology, 13*, 9533065. https://doi.org/10.3389/fpsyg.2022.9533065
Christou, E. (2025). Augmented reality in language learning. *Computer Assisted Language Learning*. https://doi.org/10.1080/17501229.2025.2504706
El Kheir, Y., Khnaisser, F., Chowdhury, S. A., Mubarak, H., Afzal, S., & Ali, A. (2023). QVoice: Arabic speech pronunciation learning application. In *INTERSPEECH 2023* (pp. 3677-3678). https://doi.org/10.21437/Interspeech.2023-1153
Hadid, A. (n.d.). Augmented reality to the rescue of language learners. *Florida Journal of Educational Research*. Retrieved from https://journals.flvc.org/fjer/article/download/133590/137510/245903
Hohmann, V., et al. (2020). The virtual reality lab: Realization and application of a versatile tool for audiovisual research. *Journal of the Acoustical Society of America, 147*(3), 1464-1480. https://doi.org/10.1121/10.0000893
Lai, C. H. (2025). Effects of individual aptitude on ultrasound biofeedback in second language pronunciation training. *Speech Communication, 165*, 102814. https://doi.org/10.1016/j.specom.2025.000014
Luo, S. (2024). A systematic review of research on xReality (XR) in English teaching. *Asian-Pacific Journal of Second and Foreign Language Education, 9*(1). https://doi.org/10.1186/s40862-023-00223-9
Sadigzade, Z. (2025). Immersive and gamified approaches: VR/AR in language learning. *Journal of Philology and Universal Research Methodology, 4*(2), 1-10.
Sebkhi, N., et al. (2020). Inertial measurements for tongue motion tracking based on magnetic localization with orientation compensation. *IEEE Transactions on Biomedical Engineering, 68*(3), 894-904. https://doi.org/10.1109/TBME.2020.3008888
Sungkur, R. K. (2025). Pronunciation trainer for second language learning using generative AI. *International Journal of Educational Technology in Higher Education, 22*(1). https://doi.org/10.1186/s41239-025-00561-x
Tolba, R. M., Elarif, T., Taha, Z., & Hammady, R. (2024). Interactive augmented reality for learning phonetics using artificial intelligence. *IEEE Access, 12*, 78219-78236. https://doi.org/10.1109/ACCESS.2024.3399888