frontiers | Frontiers in Education

TYPE Mini Review
PUBLISHED 15 April 2025
DOI 10.3389/feduc.2025.1552470

OPEN ACCESS

EDITED BY
Zhiru Sun,
University of Southern Denmark, Denmark

REVIEWED BY
Carmen Aguirre,
Complutense University of Madrid, Spain
Pedro Luchini,
National University of Mar del Plata, Argentina

*CORRESPONDENCE
Michiko Toyama
✉ toyama3@bunkyo.ac.jp

RECEIVED 28 December 2024
ACCEPTED 26 March 2025
PUBLISHED 15 April 2025

CITATION
Toyama M and Hori T (2025) Technology-enhanced multimodal approaches in classroom L2 pronunciation training.
Front. Educ. 10:1552470.
doi: 10.3389/feduc.2025.1552470

Technology-enhanced multimodal approaches in classroom L2 pronunciation training

Michiko Toyama[1]* and Tomoko Hori[2]

[1]Faculty of Business Administration, Bunkyo University, Tokyo, Japan, [2]Faculty of Health and Sports Science, Juntendo University, Chiba, Japan

Advances in L2 pronunciation instruction have embraced technology-enhanced multimodal approaches, engaging auditory, visual, and kinesthetic modalities to create interactive, immersive learning environments. This review examines three key methods: gesture-based techniques, speech visualization tools, computer-assisted pronunciation training. These approaches visualize auditory and articulatory features, reinforce prosody, or offer real-time feedback, enhancing learner engagement and retention. Despite their potential, challenges such as accessibility, technical limitations, and pedagogical integration remain. This review highlights the transformative potential of multimodal methods in L2 pronunciation training and outlines directions for future research and innovation.

KEYWORDS

multimodal teaching, multimodal learning, L2 phonology, gestures, computer-assisted pronunciation training (CAPT)

1 Introduction

Recent research increasingly highlights the potential of combining sensory modalities—auditory, visual, and kinesthetic—to enhance L2 pronunciation by reinforcing articulatory features and prosodic structures in diverse and complementary ways. These multimodal approaches align closely with interdisciplinary theories such as embodied cognition (Barsalou, 2020) and experiential learning (Kolb, 1984). Embodied cognition emphasizes the effectiveness of engaging multiple sensory pathways in learning, supporting the use of gestures and kinesthetic methods in pronunciation instruction. Experiential learning theory further guides differentiated teaching by highlighting diverse learner preferences. Systematic embodied pronunciation training thus offers considerable potential for clearer speech and effective communication (Chan, 2018). Embodied multisensory teaching techniques highlighting prosodic (e.g., Gluhareva and Prieto, 2017) and segmental (Xi et al., 2020) features thus provide robust support for phonological learning.

Technological advancements have extended these principles into the classroom, providing innovative tools to address traditional challenges in pronunciation instruction. For example, speech visualization software, audiovisual feedback, and computer-assisted pronunciation training (CAPT) platforms offer learners real-time, individualized guidance to refine their pronunciation. Similarly, immersive virtual environments (IVEs) and gesture-based methods integrate auditory and visual stimuli to deepen engagement and support long-term retention.

This review explores three key multimodal approaches—gesture-based techniques, speech visualization, CAPT—and examines their applications, benefits, and challenges in L2 pronunciation training. By synthesizing recent developments, it highlights the transformative potential of these methods and identifies opportunities for future research and innovation.

2 Multimodal learning in L2 pronunciation training

Multimodal learning has become central to L2 pronunciation training. This section explores key multimodal strategies, beginning with visual and kinesthetic cues, speech visualization tools, and computer-assisted pronunciation training (CAPT).

2.1 Visual and kinesthetic cues

2.1.1 Co-speech gestures

Research on using gestures in L2 pronunciation instruction is expanding (Thompson and Renandya, 2020). Gesture-based approach leverages the body's role in language processing, aligning with embodied cognition by linking speech production to physical actions. While promising, research in this area remains marked by contradictions and open questions. These inconsistencies often stem from the types of gestures used and the varied L2 pronunciation targets. The findings on gesture studies are summarized in Table 1 which shows how different gesture types affect various L2 targets.

Suprasegmental studies show relatively consistent results. Observing beat gestures, which mimic rhythmic patterns, has been shown to improve L2 accentedness (Gluhareva and Prieto, 2017). Similarly, pitch gestures, which represent pitch movement, have been found to aid learners in producing L2 Spanish intonation (Yuan et al., 2019) and in perceiving L2 English intonation (Hori et al., 2025). Morett (2023) further supports this finding, showing that pitch gestures conveying lexical tones enhance L2 Mandarin learners' phonological processing, as evidenced by N400 event-related potential (ERP) responses, which suggest improved lexical tone differentiation. These findings suggest that hand gestures are valuable resources for representing prosodic features, particularly in contexts where consistent co-speech gestures can be repeatedly demonstrated, such as in video clips. While these are the results of experimental research, Smotrova (2017)'s observational study showed how reiterative gestures help visualize and embody abstract pronunciation phenomena, with students adopting and adapting these gestures to improve control over suprasegmental features. Regarding body gestures, she showed that the teacher's consistent use of upward body movement to visualize L2 English word stress improved the learner's word stress placement.

Segmental studies show mixed results. Hoetjes and van Maastricht (2020) found that an iconic gesture for lip-rounding improved L2 Spanish learners' production of/u/but another iconic gesture, illustrating the action of pushing the tongue between the teeth, failed to facilitate the production of more challenging/θ/, which requires the creation of a new phonemic category. In Kelly and Lee (2012)'s study, L1 English speakers learned Japanese word pairs differing by only a gemminate (hard:/ite/vs./itte/) or by a geminate and segmental composition (easy:/tate/vs./butta/). Iconic gestures for word meanings facilitated identification of the easier pairs but disrupted learning of the harder pairs. Kelly et al. (2014) and Hirata et al. (2014) found beat gestures ineffective for improving Japanese vowel length contrasts.

Although often overlooked in L2 pedagogy, research highlights the value of co-speech gestures in language teaching (Hardison and Pennington, 2021). By engaging students through visual and physical representations, gestures foster a deeper connection to L2 sounds and boost retention (Smotrova, 2017), creating a more engaging and accessible pronunciation learning environment.

2.1.2 Facial cues and articulatory visualization

Displaying a speaker's face, including lip and jaw movements, on-screen can support L2 segmental learning. Hardison (2003) found that facial cues complement auditory input, helping Japanese and Korean L2 English learners better articulate/r/and/l/. Her multimodal approach in the high-variability perception training (HVPT), which exposes learners to multiple speakers and diverse phonetic content, significantly improved both auditory and articulatory skills. Similarly, Hazan et al. (2005) showed that combining auditory and visual cues improved learners' perception and production of L2 English contrasts, such as/b/-/v/and/r/-/l/. In line with these findings, Hirata and Kelly (2010) found that audiovisual training with a video of lip movements was more effective than audio-only training in helping learners perceive L2 Japanese vowel length contrasts (e.g., /rika/-/rika:/"science" vs. "liquor").

Sueyoshi and Hardison (2005) found that audiovisual input combining gestures and facial cues significantly improved comprehension for low-intermediate and advanced English learners, with gestures being particularly beneficial for less proficient learners. These findings underscore the importance of multimodal approaches in L2 instruction to support diverse learner needs.

2.2 Speech visualization tools

Speech visualization tools enhance L2 pronunciation training by offering visual feedback (e.g., spectrograms, pitch tracings) to clarify abstract pronunciation concepts. Indirect feedback, like acoustic displays in Praat (Boersma and Weenink, 2001), are beneficial but typically require specialized training (Levis, 2007), yielding variable results (Bliss et al., 2018; Olson, 2022). Direct feedback tools [e.g., ultrasound, Electromagnetic Articulography (EMA)], provide real-time visualization of articulator movements, offering more intuitive guidance for learners (Bliss et al., 2018). Ongoing technological advancements continue to make these multimodal feedback tools more accessible, allowing teachers to effectively integrate them into interactive classroom pronunciation activities (Levis and Pickering, 2004).

2.3 Computer-assisted pronunciation training (CAPT)

The field of computer-assisted language learning has expanded significantly, with computer-assisted pronunciation training (CAPT) gaining momentum through web-based and mobile applications (Rogerson-Revell, 2021). CAPT integrates automated speech recognition (ASR) and artificial intelligence (AI), to offer

TABLE 1 Findings on gesture studies in L2 pronunciation.

	Study	Gesture type	Participants	Effect observed
Suprasegmental studies	Gluhareva and Prieto, 2017	Beat gestures	20 L1 Spanish-L2 English	Improved accentedness
	Yuan et al., 2019	Pitch gestures	64 L1 Mandarin-L2 Spanish	Improved intonation production
	Morett, 2023	Pitch gestures	44 L1 English- L2 Mandarin	Enhanced lexical tone discrimination (via N400 ERP)
	Hori et al., 2025	Pitch gestures	80 L1 Japanese- L2 English	Improved intonation perception
	Smotrova, 2017	Upward body movement	12 Mixed L1s- L2 English	Improved word stress placement
Segmental studies	Hoetjes and van Maastricht, 2020	Iconic gestures	50 L1 Dutch- L2 Spanish	Improved production of /u/, but not /θ/
	Kelly and Lee, 2012	Iconic gestures for word meanings	42 L1 English-L2 Japanese	Improved identification of easy geminate/singleton pairs
	Kelly et al., 2014; Hirata et al., 2014	Beat gestures for rhythm	88 L1 English-L2 Japanese	Ineffective in improving vowel length contrasts

personalized learning experiences through interactive tools that enhance L2 perception, production, and feedback, making them invaluable for both independent practice and teacher-led lessons.

2.3.1 Computer-animated agents and visual feedback

A prominent feature of CAPT is the use of computer-animated agents or avatars, which visually model articulation to assist learners in internalizing challenging pronunciation. Tools like Baldi® (Massaro and Light, 2003) and ARTUR (Engwall, 2008) use animated characters with realistic articulators to demonstrate lip rounding, tongue placement, and palate positioning. Baldi allows learners to adjust playback speed and viewing angles of articulatory movements, providing versatile visual cues that complement auditory input. ARTUR focuses on precise visualizations of internal articulatory processes, helping learners analyze tongue and palate configurations in detail.

2.3.2 Immersive virtual environments (IVEs) as advanced CAPT

Immersive Virtual Environments (IVEs), including virtual and augmented reality platforms, represent an extension of CAPT by incorporating interactive avatars and contextualized learning scenarios. Unlike traditional CAPT tools, IVEs provide learners with immersive, real-world practice opportunities that promote both fluency and accuracy (Legault et al., 2019). Online platforms such as *ImmerseMe* allow learners to engage in lifelike conversations with virtual agents, offering self-conscious learners who are hesitant to engage in conversations with native speakers a safe space to use the target language until they have built up their confidence (He and Smith, 2019).

Recent studies have explored the potential of IVEs in enhancing language learning through interactive and engaging platforms. For example, Berns et al. (2013) found that game-like 3D virtual environments significantly improved learners' motivation and skills in L2 German vocabulary, pronunciation, listening, and writing. The study suggested that combining VR with traditional classroom methods may be the most effective approach. Similarly, O'Brien and Levy (2008) reported that students who learned L2 German commands through a VR game found it was especially helpful for improving their listening skills. Over half of the participants felt the activity enhanced their German knowledge, appreciating its practicality and finding it more engaging than traditional coursework.

These findings underscore the potential of IVEs to enhance language acquisition by integrating skill development with cultural immersion and learner engagement, offering opportunities beyond those of traditional media-supported classrooms.

3 Discussion

Technology-enhanced multimodal approaches in L2 pronunciation training engage learners' sensory modalities and improve phonological awareness. This section discusses findings, implications, challenges, and future directions for integrating multimodal methods in classrooms.

3.1 Benefits of multimodal pronunciation training

Multimodal approaches effectively enhance segmental and prosodic features. Visual and kinesthetic cues, such as co-speech gestures (2.1), link physical movements with phonetic outcomes, improving articulatory and auditory skills (Chan, 2018) and increasing intelligibility for L2 learners (Wheeler and Saito, 2022). Speech visualization tools (2.2), including spectrograms and articulatory simulations, provide immediate feedback for refining pronunciation (Bliss et al., 2018; Engwall, 2008). CAPT tools (2.3) support pronunciation learning with real-time feedback and scalability. Particularly IVEs, offer immersive, real-world experiences that boost fluency and accuracy while supporting classroom instruction and independent practice.

3.2 Challenges and limitations

Despite their benefits, multimodal techniques face several implementation challenges. Advanced tools like ultrasound, EMA, and IVEs often require substantial investment, technical expertise, and adequate infrastructure (Bliss et al., 2018). Similarly, CAPT tools rely heavily on high-quality speech recognition to be effective.

Institutional constraints, such as uneven access to technological resources, can further limit the adoption of these tools. For instance, CAPT platforms dependent on consistent internet connectivity and advanced computing infrastructure may be inaccessible in resource-limited institutions, leading to equity gaps in pronunciation training.

Learner variability also presents pedagogical challenges. Some learners prefer concrete, gesture-based methods, while others favor analytical speech visualization. Effectiveness may also vary with individual proficiency levels and language development stages. Teachers must therefore carefully balance technological resources with personalized instruction to optimize learning outcomes.

3.3 Pedagogical implications

A blended approach combining technology with teacher-led instruction appears highly effective for pronunciation training. Technology offers consistency, scalability, and real-time feedback, while human teachers provide essential emotional engagement, adaptability, and holistic guidance. Co-speech gestures, guided by teacher explanations, help learners identify and internalize auditory and articulatory features. Structured practice with PC-recorded gestures allows students to mimic movements, reinforcing phonological learning. For ease of adoption, especially for teachers new to this approach, research-validated gestures are recommended. Simulation-based tools like virtual tutors (e.g., ARTUR) further scaffold learners' progress in controlled, interactive environments (Engwall, 2008) but they lack the motivational support unique to human teachers. Immersive platforms like ImmerseMe offer interactive dialogue scenarios filmed in 360-degree video, enabling learners to practice pronunciation in realistic conversational contexts. Dictation and repetition tasks with virtual interlocutors further encourage students to actively practice speaking, receive immediate automated feedback, and refine their pronunciation. These virtual interactions also provide self-conscious learners with a safe, anxiety-reducing environment for speaking practice (He and Smith, 2019). Teachers complement technology by offering personalized encouragement, addressing learner variability, and ensuring contextual appropriateness in practice.

To cater to diverse learner needs, theories such as the Experiential Learning Theory (Kolb, 1984) offer a valuable framework. His theory highlights four distinct learning styles —diverging, assimilating, converging, and accommodating —which are based on a four-stage learning cycle, i.e., concrete experience, reflective observation, abstract conceptualization, and active experimentation. Diverging learners who prefer observing and reflecting may benefit from watching co-speech gestures to study pronunciation from different perspectives to gather information. Assimilating learners who are analytical and like to work alone may benefit from spectrograms, articulatory simulations, and visual acoustic feedback. Converging learners who enjoy solving practical problems may benefit from virtual tutors and real-world scenario simulations in IVEs. Accommodating learners who favor hands-on, intuitive experiences may benefit from practicing rhythm by physically mimicking beat gestures.

3.4 Future directions

Future research should examine the long-term impacts of multimodal approaches on L2 pronunciation, especially in diverse classroom contexts. Exploring learner engagement with CAPT tools can clarify their broader applicability, and developing affordable, user-friendly technologies is essential for wider accessibility. Additionally, researchers could explore potential ethical issues, including biases in AI-driven pronunciation feedback systems. Such biases may disproportionately impact learners from diverse linguistic or cultural backgrounds, highlighting the importance of ethically informed development and rigorous evaluation of CAPT tools to ensure equitable and fair learning outcomes. Personalizing multimodal feedback through AI and machine learning also offers promising opportunities to deliver precise, adaptable support for differentiated instruction.

4 Conclusion

In summary, technology-enhanced multimodal approaches hold great promise for L2 pronunciation training, as they engage learners holistically through auditory, visual, and kinesthetic modalities. While challenges such as accessibility and learner variability persist, a blended instructional model that leverages teacher expertise alongside advanced tools can optimize learning outcomes. Future research and technological advancements will be key to expanding the potential of these approaches in L2 pronunciation classrooms.

Author contributions

MT: Writing – original draft, Writing – review and editing. TH: Funding acquisition, Writing – review and editing.

Funding

The author(s) declare that financial support was received for the research and/or publication of this article. This study was supported by JSPS KAKENHI (grant number JP21K00741).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The authors declare that Gen AI was used in the creation of this manuscript. Generative AI tools were used solely to assist with language proofreading during the preparation of this manuscript. The AI was employed to help identify and correct grammatical errors and improve sentence clarity. All content was written, reviewed, and verified by the authors to ensure academic integrity and accuracy.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Barsalou, L. (2020). Challenges and opportunities for grounding cognition. *J. Cogn.* 3:31. doi: 10.5334/joc.116

Berns, A., Gonzalez-Pardo, A., and Camacho, D. (2013). Game-like language learning in 3-D virtual environments. *Comput. Educ.* 60, 210–220. doi: 10.1016/j.compedu.2012.07.001

Bliss, H., Abel, J., and Gick, B. (2018). Computer-assisted visual articulation feedback in L2 pronunciation instruction: A review. *J. Sec. Lang. Pronunciation* 4, 129–153. doi: 10.1075/jslp.00006.bli

Boersma, P., and Weenink, D. (2001). Praat, a system for doing phonetics by computer. *Glot Int.* 5, 341–345.

Chan, M. J. (2018). Embodied pronunciation learning: Research and practice. *Catesol. J.* 30, 47–68. doi: 10.5070/B5.35964

Engwall, O. (2008). "Can audio-visual instructions help learners improve their articulation? An ultrasound study of short-term changes," in *Proceedings of Interspeech*, 2631–2634.

Gluhareva, D., and Prieto, P. (2017). Training with rhythmic beat gestures benefits L2 pronunciation in discourse-demanding situations. *Lang. Teach. Res.* 21, 609–631. doi: 10.1177/1362168816651463

Hardison, D. M. (2003). Acquisition of second-language speech: Effects of visual cues, context, and talker variability. *Appl. Psycholinguist.* 24, 495–522. doi: 10.1017/S0142716403000250

Hardison, D. M., and Pennington, M. C. (2021). Multimodal second-language communication: Research findings and pedagogical implications. *RELC J.* 52, 62–76. doi: 10.1177/0033688220966635

Hazan, V., Sennema, A., Iba, M., and Faulkner, A. (2005). Effect of audiovisual perceptual training on the perception and production of consonants by Japanese learners of English. *Speech Commun.* 47, 360–378. doi: 1016/j.specom.2005.04.007

He, L., and Smith, J. (2019). "ImmerseMe," in *Proceedings of the 10th Pronunciation in Second Language Learning and Teaching Conference*, (Ames, IA: Iowa State University), 461–466.

Hirata, Y., and Kelly, S. D. (2010). Effects of lips and hands on auditory learning of second-language speech sounds. *J. Speech Lang. Hear. Res.* 53, 298–310. doi: 10.1044/1092-4388(2009/08-0243)

Hirata, Y., Kelly, S. D., Huang, J., and Manansala, M. (2014). Effects of hand gestures on auditory learning of second-language vowel length contrasts. *J. Speech Lang. Hear. Res.* 57, 2090–2101. doi: 10.1044/2014_JSLHR-S-14-0049

Hoetjes, M., and van Maastricht, L. (2020). Using gesture to facilitate L2 phoneme acquisition: The importance of gesture and phoneme complexity. *Front. Psychol.* 11:575032. doi: 10.3389/fpsyg.2020.575032

Hori, T., Akatsuka, M., and Toyama, M. (2025). Effects of observing pitch gestures on the perception of English intonation by Japanese learners of English. *J. Sec. Lang. Pronunciation.* doi: 10.1075/jslp.24015.hor [Epub ahead of print].

Kelly, S. D., and Lee, A. L. (2012). When actions speak too much louder than words: Hand gestures disrupt word learning when phonetic demands are high. *Lang. Cogn. Neurosci.* 27, 793–807. doi: 10.1080/01690965.2011.581125

Kelly, S. D., Hirata, Y., Manansala, M., and Huang, J. (2014). Exploring the role of hand gestures in learning novel phoneme contrasts and vocabulary in a second language. *Front. Psychol.* 5:673. doi: 10.3389/fpsyg.2014.00673

Kolb, D. A. (1984). *Experiential Learning: Experience as the Source of Learning and Development*. Englewood Cliffs, NJ: Prentice-Hall.

Legault, J., Zhao, J., Chi, Y. A., Chen, W., Klippel, A., and Li, P. (2019). Immersive virtual reality as an effective tool for second language vocabulary learning. *Languages* 4:13. doi: 10.3390/languages4010013

Levis, J. (2007). Computer technology in teaching and researching pronunciation. *Annu. Rev. Appl. Linguistics* 27, 184–202. doi: 10.1017/S0267190508070098

Levis, J., and Pickering, L. (2004). Teaching intonation in discourse using speech visualization technology. *System* 32, 505–524. doi: 10.1016/j.system.2004.09.009

Massaro, D. W., and Light, J. (2003). "Read my tongue movements: Bimodal learning to perceive and produce non-native speech /r/ and /l/," in *Proceedings of 8th European Conference on Speech Communication and Technology (Eurospeech 2003)*, 2249–2252. doi: 10.21437/Eurospeech.2003-629

Morett, L. M. (2023). "Observing gestures during L2 word learning facilitates differentiation between unfamiliar speech sounds and word meanings," in *Proceedings of the Annual Meeting of the Cognitive Science Society*, (Psychology Press).

O'Brien, M. G., and Levy, R. M. (2008). Exploration through virtual reality: Encounters with the target culture. *Can. Modern Lang. Rev.* 64, 663–691. doi: 10.3138/cmlr.64.4.663

Olson, D. J. (2022). Phonetic feature size in second language acquisition: Examining VOT in voiceless and voiced stops. *Sec. Lang. Res.* 38, 913–940. doi: 10.1177/02676583211008951

Ramírez Verdugo, D. (2006). A study of intonation awareness and learning in non-native speakers of english. *Lang. Awareness* 15, 141–159. doi: 10.2167/la404.0

Rogerson-Revell, P. M. (2021). Computer-assisted pronunciation training (CAPT): Current issues and future directions. *RELC J.* 52, 189–205. doi: 10.1177/0033688220977406

Smotrova, T. (2017). Making pronunciation visible: Gesture in teaching pronunciation. *TESOL Quart.* 51, 59–89. doi: 10.1002/tesq.276

Sueyoshi, A., and Hardison, D. M. (2005). The role of gestures and facial cues in second language listening comprehension. *Lang. Learn.* 55, 661–699. doi: 10.1111/j.0023-8333.2005.00320.x

Thompson, A. A., and Renandya, W. A. (2020). Use of gesture for correcting pronunciation errors. *TEFLIN J.* 31, 342–359. doi: 10.15639/teflinjournal.v31i2/342-359

Wheeler, P., and Saito, K. (2022). Second language speech intelligibility revisited: Differential roles of phonological accuracy, visual speech, and iconic gesture. *Modern Lang. J.* 106, 429–448. doi: 10.1111/modl.12779

Xi, X., Li, P., Baills, F., and Prieto, P. (2020). Hand gestures facilitate the acquisition of novel phonemic contrasts when they appropriately mimic target phonetic features. *J. Speech Lang. Hear. Res.* 63, 3571–3585. doi: 10.23641/asha

Yuan, C., González-Fuente, S., Baills, F., and Prieto, P. (2019). Observing pitch gestures favors the learning of Spanish intonation by Mandarin speakers. *Stud. Sec. Lang. Acquisition* 41, 5–32. doi: 10.1017/S0272263117000316