frontiers | Frontiers in Communication

TYPE Original Research
PUBLISHED 05 January 2026
DOI 10.3389/fcomm.2025.1620465

OPEN ACCESS

EDITED BY
Loulou Kosmala,
Université Paris-Est Créteil Val de Marne, France

REVIEWED BY
Plinio Almeida Barbosa,
State University of Campinas, Brazil
Xiaotong Xi,
Shandong University of Finance and Economics, China

*CORRESPONDENCE
Noriko Yamane
✉ yamanen@hiroshima-u.ac.jp
Masahiro Shinya
✉ mshinya@hiroshima-u.ac.jp

RECEIVED 29 April 2025
ACCEPTED 13 October 2025
PUBLISHED 05 January 2026

CITATION
Yamane N, Shinya M, Tan X and Chiya A (2026) Differential effects of hand and mouth gesture training on L2 English pronunciation: targeting suprasegmental and segmental features.
Front. Commun. 10:1620465.
doi: 10.3389/fcomm.2025.1620465

Differential effects of hand and mouth gesture training on L2 English pronunciation: targeting suprasegmental and segmental features

Noriko Yamane[1]*, Masahiro Shinya[1]*, Xiaofeng Tan[1] and Amos Chiya[1,2]

[1]Graduate School of Humanities and Social Sciences, Hiroshima University, Higashihiroshima, Japan, [2]Nagoya University of Commerce and Business, Aichi, Japan

Human communication inherently integrates speech and gesture. Acquiring second language (L2) pronunciation, encompassing both segmental (e.g., vowels) and suprasegmental features (e.g., rhythm, fluency), remains a major challenge. This study investigated how two types of gesture training—manual (hand gesture training) versus articulatory (mouth gesture training)—influence these features in Japanese EFL learners. Forty university students participated in a four-week counterbalanced design, receiving hand gesture training (rhythmic circular motions) and mouth gesture training (bio-visual feedback for /æ/ vs. /ʌ/ distinction). Speech rate (as a suprasegmental proxy) and second formant (F2) values of target vowels (as a segmental proxy) were measured at pre, mid-, and post-training. Results revealed distinct effects: hand gesture training significantly improved speech rate across both groups, enhancing suprasegmental fluency, while mouth gesture training significantly improved F2 distinction for /æ/. These findings suggest that hand and mouth gestures target complementary aspects of L2 pronunciation. Taken together, the results support an embodied, multimodal approach to pronunciation instruction, highlighting the pedagogical value of integrating suprasegmental fluency practice with segmental refinement.

KEYWORDS

multimodal communication, embodied cognition, speech–gesture integration, suprasegmental fluency, segmental accuracy, vowel production, biovisual feedback, Japanese EFL learners

Introduction

Human communication is inherently multimodal, with speech and gesture tightly intertwined in the construction of meaning. Gestures are not merely ancillary to speech; rather, they are deeply integrated with cognitive and interactive processes, shaping and being shaped by the dynamics of real-time interaction. Research on gesture-speech coupling has highlighted how gestures facilitate comprehension, structure discourse, and serve cognitive functions such as disambiguation and conceptual organization.

Theoretical background

Research increasingly shows that speech and gesture form an integrated cognitive and interactional system, rather than parallel channels. Within embodied cognition, gestures

ground linguistic meaning in sensorimotor experience (Johnson and Lakoff, 2002; Barsalou, 2008) with often indicating metaphorical mappings (Lakoff and Johnson, 1980). Thus gestures help activate, manipulate, and package information for speech, reflecting their role in the integrated cognitive system that underlies both thinking and speaking (Kita et al., 2017). Interactionist accounts emphasize that gesture–speech timing is socially organized through multiple semiotic resources, with gesture, talk, gaze, and other modalities functioning as coordinated parts of interaction (Goodwin, 2007; Kendon, 2004; Parisse et al., 2022). Both cognitive and interactional approaches converge in viewing fluency as an integrative outcome, realized through the smooth expression of intrapersonal (mind–body) and interpersonal (speaker–interlocutor) coordination. Fluency is thus multidimensional, encompassing speech, interaction, and gesture, a perspective that Kosmala (2024) dubs 'inter-fluency.'

In this study, gestures include both hand movements and vocal tract actions such as the tongue and lips. Articulatory Phonology (AP; Browman and Goldstein, 1986, 1989) defines gestures as vocal-tract actions whose spatiotemporal coordination underlies phonological structure and phonetic implementation. Although AP presents the elaborate system of vocal tract gestures rather than the general body gestures, this perspective dissolves the boundary between speech and gesture, showing that linguistic segments and prosodic patterns emerge from the temporal coordination of bodily actions in an integrated communicative system (McNeill, 1992; Goldin-Meadow and Alibali, 2013). This kind of view is supported by evolutionary and experimental research, which suggests that manual gestures shape speech development and performance (Shattuck-Hufnagel and Ren, 2018; Pouw et al., 2021; Vainio, 2019; Gentilucci and Volta, 2008), with disfluencies often mirrored across modalities (Kosmala et al., 2023).

Phonology is a cognitive representation of physical actions. From this type of perspective, AP's gestural score specifies the embodied primitives of vocal tract actions. These primitives are hierarchically built up to syllables, foot, and prosodic words (see Selkirk, 1980 et seq., for 'Prosodic Hiearchy'). Prosodic words are the basis for phonological phrases and other larger categories, which function as a domain of various phonological rules. For example, rhythm rules were explained in Metrical Theory (Liberman and Prince, 1977; Hayes, 1995), which provides the cognitive scaffold that structures their temporal organization. The metrical grid encodes hierarchies of strong and weak beats that act as attractors for attention and timing, thereby licensing gestures at prominent positions such as stressed syllables, prosodic word boundaries, and phrasal edges. Recent models of oscillatory entrainment (Cummins and Port, 1998; Doelling et al., 2014) reinforce this interpretation by showing that prosodic rhythm reflects timing mechanisms, which align the execution of oral and manual gestures with rhythmic beats.

Although research on multimodality has grown steadily, systematic investigations linking gestures overall to phonological forms remain limited. While many gestures synchronize with pitch accents (Wagner et al., 2014), other articulators—the lips, tongue, cheeks, eyes, eyebrows, and head—appear to coordinate with different aspects of linguistic structure. Cross-linguistic studies illustrate this complexity: eyebrow raising, for instance, follows distinct temporal patterns in English and Japanese (de La Cruz-Pavía et al., 2020), and the tongue and lips help establish language-specific articulatory settings across utterances (Gick et al., 2004; Wilson et al., 2025). For EFL learners, the lack of explicit guidance on how such articulatory gestures should be timed and integrated risks reinforcing unnatural rhythm and persistent accentedness. What emerges, then, is a clear pedagogical imperative: gesture-informed teaching practices—drawing on both articulatory and manual cues—must be incorporated into pronunciation instruction, not as an optional supplement, but as an essential means of fostering naturalistic fluency and prosody.

L2 based work

L2-based work has employed gestures into pronunciation instruction to boost learners' understanding of English suprasegmental traits. For prosody training, 'beat' gestures—cyclic up and down movements of a hand—when aligned with stressed syllables of English, has been found to help regulate speech rhythm (McCafferty, 2002), and to facilitate the students' identification and production of syllables, word stress, and the rhythm of speech (Smotrova, 2017), since the beat gestures synchronize with prosodic peaks in English (Leonard and Cummins, 2011). Empirical studies report benefits for learners, such as reduced perceived accentedness (Gluhareva and Prieto, 2017), improved memory for pitch accents (Kushch et al., 2018), wider pitch range and durational contrast (Yamane et al., 2019), and enhanced pitch control and fluency (Cavicchio and Busà, 2023). Learner-produced beat gestures also show improvements of L2 English pronunciation, particularly among Catalan learners, where training with beat gestures yielded significantly lower accentedness than training without them (Llanes-Coromina et al., 2018; Prieto et al., 2025).

Compared to suprasegmental trainings, hand gesture benefits to segmental improvements seem to be more limited. Xi et al. (2024) found that learners using hand gestures mimicking lip aperture (wide for /æ/, narrow for /ʌ/) outperformed those mimicking tongue position or those using no gestures, suggesting that lip-focused cues are particularly effective. Hand gestures have been applied to vowel length contrasts as well (Hirata and Kelly, 2010; Hirata et al., 2014; Li et al., 2020, 2021), which we classify as suprasegmental (i.e., prosodic) feature. Within a framework of Autosegmental Phonology (e.g., Goldsmith, 1976; Hayes, 1995; Kubozono, 2017), vowel length is a property of its association to the prosodic (moraic) tier, where the length contrast is characterized in the number of morae nested by syllable unit (i.e., short vowel has one mora, while long vowel consists of two moras). This interpretation aligns with previous studies showing that manual gestures are particularly effective for suprasegmental features such as rhythm and fluency, whereas segmental accuracy is more directly supported by articulatory feedback. These findings suggest that lower-level segmental gestures, such as consonants and vowels, may benefit less from hand gestures than higher-level prosodic units. Instead, visual feedback on learners' own oral articulatory gestures may provide a more effective pathway for improving segmental accuracy (Suemitsu et al., 2015; Antolík et al., 2019; Kocjančič et al., 2024; Yamane et al., 2025), a possibility that warrants further investigation in future research.

Although gestures have been examined at both segmental and suprasegmental levels, systematic comparisons of objective outcomes across these domains remain underexplored, highlighting the need for studies that directly evaluate their relative effectiveness. Furthermore, though some gesture-based pedagogies have been shown to benefit learners in other Asian EFL contexts (Ma and Jin, 2022; Wang et al.,

2023), their specific impact on Japanese learners' fluency development has yet to be systematically examined.

The present experiment is designed to address this gap by testing training effects at both levels of phonology, targeting Japanese learners of English. Our focus is not to capture effects at all levels claimed under the prosodic hierarchy, but to explore two domains—suprasegmental and segmental levels—as an initial step in understanding gesture–speech integration. Neurocognitive research further supports this perspective, as delta- (0.5–3 Hz) and theta-band (3–9 Hz) rhythms have been shown to align with prosodic and syllabic cycles (Giraud and Poeppel, 2012; Doelling et al., 2014), providing a biological bridge between abstract prosodic structure and gesture–speech integration. Gestures appear to pattern with this same rhythmic system. For example, beat gestures frequently precede word onsets by approximately 100 ms, effectively resetting listeners' neural oscillations to sharpen temporal prediction and facilitate speech segmentation (Biau and Soto-Faraco, 2015; Biau et al., 2015). Together, these findings indicate that speech and gesture are not independent channels but coordinated expressions of a shared timing mechanism that underlies both perception and communication. Importantly, the present study integrates both suprasegmental and segmental targets within a single experimental design. By contrasting gesture types—hand gestures associated with suprasegmental development and mouth gestures with segmental refinement—it seeks to advance theoretical understanding of the rhythm–articulation interface while also offering pedagogical guidance for optimizing gesture-based L2 pronunciation training.

Purpose

The purpose of this study is to compare the effects of two gesture-based training methods—manual gestures and articulatory gestures—on distinct linguistic features of Japanese EFL learners' pronunciation. We also consider how the integration of these methods may provide complementary benefits for suprasegmental and segmental development.

Japanese learners, whose first language is based on a mora-timed rhythm (Port et al., 1987), tend to produce English with less durational variability in across all vowels in words, mirroring the more regular rhythm of Japanese. This produces English that sounds overly even and less natural to native listeners, often giving the impression of a slowed overall speech pace. The unnaturalness arises from the absence of vowel reduction in unstressed syllables and cliticization, processes through which the stress-timed rhythm of English facilitates phrasing and accelerates speech tempo. Thus, if gestures are carefully designed to guide learners toward temporal alignment with the prosodic peaks of English, they may come to chunk phrases, accelerate speech tempo, and thereby facilitate the development of 'speed fluency' (Lennon, 1990; de Jong, 2023), an area where Japanese speakers often face persistent difficulties (Tajima and Port, 2004; Kawase et al., 2024).

As for segmental skills, Japanese learners show consistently struggle with the vowel /æ/ ('ash'; low front vowel), which is absent from their native five-vowel system /a, i, u, e, o/, and is often substituted with /a/ ('lower-case a'; low central/back vowel) (Lambacher et al., 2005). This substitution arises because these two vowels share tongue height and show overlap in F1, although they differ in tongue backness. English /æ/ typically has F2 values around 1700–2050 Hz (Peterson and Barney, 1952; Hillenbrand et al., 1995), whereas Japanese /a/ F2 averages only 1,283 Hz for males and 1,530 Hz for females (Yazawa and Kondo, 2019). These values place Japanese /a/ much closer to English /ʌ/ ('wedge'; mid back vowel) than to /æ/. Orthographic conventions in the Japanese kana loanword system, clearly reflecting this merger (e.g., lab / love → ラブ), seem to be a contributing factor to both perceptual and articulatory confusions. Given these challenges, vowel contrasts such as /æ/ versus /a/ emerge as ideal targets for gesture-based training interventions, on par with the importance of speed fluency training.

This study investigates how manual (hand) and articulatory (mouth) gestures can facilitate the acquisition of specific phonological features in L2 learners, advancing an embodied, multimodal approach to pronunciation instruction. The research addresses two primary questions:

i) How do different types of gesture training differentially influence segmental and suprasegmental aspects of L2 speech?
ii) How does the timing of gesture training (hand-first vs. mouth-first) influence the trajectory of improvement across training phases?

We predict level-specific outcomes: learners trained with hand gestures will show greater improvement in suprasegmental fluency measured by speech rate, whereas learners trained with mouth gestures will demonstrate greater gains in segmental (vowel) accuracy measured by F2. Furthermore, we expect the timing of training to shape the trajectory of improvement: introducing hand training earlier will yield earlier fluency gains, while introducing mouth training earlier will yield earlier segmental gains.

Method

To test these predictions, we implemented a counterbalanced training design. Learners were divided into two groups that differed in the order of training: one group received hand training followed by mouth training (Hand–Mouth, HM), and the other group received mouth training followed by hand training (Mouth–Hand, MH). Training effects were assessed across three test phases: Pre, Mid, and Post. This design allowed us to examine not only the overall benefits of each type of gesture training (hand vs. mouth), but also whether the timing of training (earlier vs. later in the sequence) influenced the trajectory of improvement across phases.

Participants

Fifty Japanese university students from two classes of the English communication course participated in this study. Ten participants were excluded from the analysis because they failed to complete the entire experimental procedure. As a result, data from the remaining 40 participants (aged 18–19) were included in the analysis. All participants reported no history of hearing or speech impairments, were informed about the experimental guidelines, and provided written informed consent prior to participation. This study received ethical approval from the institutional review board of Hiroshima University.

Before the pretest session, participants were assigned to two distinct experimental groups according to their class affiliations. Two

training methods were implemented over a four-week period. Because students were taught in intact classes determined by the institution, we assigned the two training methods in reverse order across classes to counterbalance potential order effects, ensuring that any improvements could not be attributed solely to the method presented first (Note: While assigning intact classes to different training orders helped mitigate potential order effects, we acknowledge that this quasi-experimental design does not provide the same level of control as full randomization, and we note this as a limitation of the study).

Specifically, the group that first received Hand Gesture Training (HGT) (Figure 1) followed by Mouth Gesture Training (MGT) (Figure 2) was designated as the Hand-Mouth group (HM group, $n = 19$), while the group that followed the reverse order was labeled as the Mouth-Hand group (MH group, $n = 21$) (see Table 1).

Regarding their English proficiency, all participants reported having learned English as a second language through school-based instruction and indicated no experience of long-term residence in an English-speaking country. As first-year students, none had received formal training in English pronunciation or taken courses in linguistics or phonetics. Although the HM and MH groups differed significantly in their TOEIC scores ($p = 0.006$), the overall proficiency of the participants was relatively low ($M = 455.0 \pm 105.9$), approximately corresponding to CEFR levels A2–B1 and typical of Japanese first-year university students educated primarily through school-based instruction. Moreover, because the TOEIC (L&R) primarily assesses receptive skills (listening and reading) and do not fully represent overall English proficiency—particularly productive skills—we did not consider it appropriate to classify participants into high- and low-proficiency group on this basis. For practical reasons, they were instead assigned to two groups based on their institution-assigned class affiliations rather than their TOEIC scores.

Speech materials

In both training methods, a tongue twister titled "Betty Botter" was employed as the speech material. This tongue twister was selected because it includes the target vowels /æ/ in "batter" and /ʌ/ in "butter," which consistently appear in the same phonetic environment, surrounded by two consonants /b/ and /t/. Furthermore, it consisted of 63 syllables, with each word containing no more than two syllables. Such a design ensured that the participants, who were EFL learners, would not be overwhelmed by the potential complexity. The complete text content of "Betty Botter" is as follows.

> Betty Botter bought some butter;
> "But" she said "This butter's bitter!"
> If I put it in my batter,
> it would make my batter bitter.
> But a bit of better butter will make my batter better.
> So t'was better Betty Botter bought a bit of better butter.

Gesture training procedures

To examine how different gesture modalities contribute to L2 pronunciation development, we implemented two training conditions that target distinct phonological levels. Hand Gesture Training (HGT) was designed to support suprasegmental development by aligning manual movements with rhythmic and stress patterns, thereby reinforcing learners' awareness of timing and fluency. In contrast, Mouth Gesture Training (MGT) focused on segmental refinement by drawing learners' attention to tongue and lip configurations that differentiate the difficult vowel contrasts /æ/ and /ʌ/. The HGT, aimed at improving the fluency of English oral reading, and the MGT, aimed at enhancing pronunciation accuracy of the target sounds, were assigned to participants in two groups (the HM group and the MH group) with a reversed training sequence to ensure counterbalancing. Both conditions used the *Betty Botter* passage as practice material, enabling a direct comparison of how suprasegmental versus segmental gesture-based instruction facilitates L2 pronunciation learning.

For HGT, we used 'circular' gestures. Circular gestures occur naturally in everyday speech to emphasize rhythm and prosody, and are particularly used in music performances such as choral music to enhance expression and the quality of the overall performance (Jansson et al., 2021; Kilpatrick, 2020). In the field of conducting, circular motions are a type of beat gesture, and form a foundational part of gestural vocabulary. These circular and rounded motions are commonly found in almost all types of beat patterns, such as a 4/4 beat pattern, and 2/4 beat pattern in conducting *legato*, or melodious, smooth and continuous melodic lines (Figure 3). Most notably within the Ilya Musin method, a

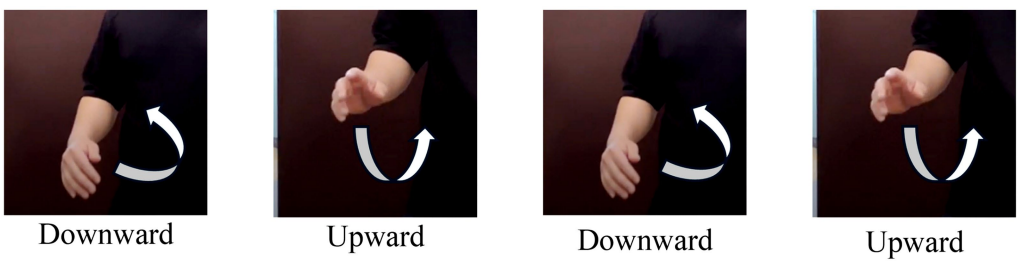

FIGURE 1
Hand-gesture-based training (HGT). A hand moves up and down in a circular motion, with the maximum downward extension synchronized with underlined words.: "Betty Botter bought some butter, but she said this butter's bitter, if I put it in my batter, ..." One movement cycle roughly corresponds to a phonological phrase (see Figures 5, 6 for the details).

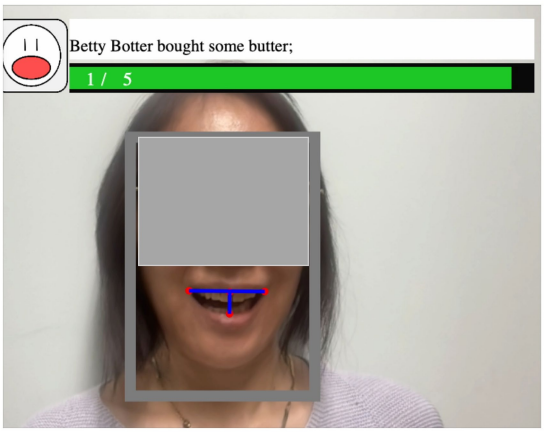

FIGURE 2
Mouth-gesture-based training (MGT). When the start button is pressed on the web-based face mesh program (Shitara et al., 2023), blue vertical and horizontal bars appear over participant's mouth opening area to indicate whether the mouth corners are appropriately raised. Learners then practice the sentence displayed at the top of the screen. A green bar shows the progress of the task.

TABLE 1 TOEIC (L&R) scores by group and sex.

Group	Sex	N	M	SD
HM group	Female	9	507.8	76.9
	Male	10	496.5	86.7
	All	19	501.8**	80.1
MH group	Female	15	455.3	9.3
	Male	6	305.8	70.0
	All	21	412.6**	110.0

M, mean; SD, standard deviation. An independent-samples *t*-test revealed a significant difference between HM group and MH Group in TOEIC scores, $p = 0.006$ (**).

highly respected school of conducting that emphasizes clarity and expressiveness through wrist-led movement (Musin, 1967; Ogrizovic-Ciric, 2009), the circular motions are also part of the beat gestures used in conducting single beats, compared to other common traditions of only beating up and down (Figure 4). In all circular gestures, consistent beat points were positioned at the onset of the hand's upward motion, reflecting conductors' metaphorical mapping of spatial rise onto musical crescendo (Meissl et al., 2022) or pitch rise (Morett et al., 2022). Drawing from this rationale and tradition, we integrated the circular motions from the Musin method into the training protocol, as they provide a controlled yet naturalistic means of coordinating physical gestures with rhythmic patterns in speech.

The details of HGT and MGT are given below.

Hand gesture training (HGT)

In Week 1, learners received hand-gesture training to enhance awareness of stressed syllables, followed in Week 2 by training focused on phonological phrases (typically noun phrases and verb phrases). This progression from lower- to higher-level suprasegmental units is aligned with the principles of prosodic hierarchy.

WEEK 1 (Strokes at stressed syllable level):

- An instructor showed circular strokes at every stressed syllable; 'raising' phase (x) aligned with every stressed syllable (e.g., Betty Botter bought some butter), reinforcing their awareness of cycles of stressed syllables (Figure 5).
- Students stood along walls and read aloud in unison while imitating the instructor's gestures. The instructor approached each student, and checked their hand shape, orientation and tension, and gave them verbal and haptic feedback.
- Students and the instructor read aloud with doing circular motion in unison about 5 times in total.

WEEK 2 (Strokes at phonological phrase level):

- An instructor showed circular strokes at every phonological phrase: 'raising' phase (x) aligned with the first stressed syllables within phonological phrases (e.g., [Betty Botter] [bought some butter]), reinforcing their awareness of cycles of phrases (Figure 6).
- Students and the instructor read aloud with doing circular motion in unison about 5 times in total. When the instructor notice students' erroneous hand motion, they were given verbal and haptic feedback.

Mouth gesture training (MGT)

In Week 1, learners engaged in listening and imitation tasks to build awareness of articulatory differences between two vowels. In Week 2, they progressed to lingual and lip shaping drills with real-time visual feedback, moving from awareness-raising to self-modulated practice to support changes in articulatory behavior.

WEEK 1 (Listening and imitation):

- An instructor conducted listening quiz contrasting /æ/ and /ʌ/ using *English Accent Coach* (Thomson, 2012).
- An instructor showed *Jolly Phonics* videos (Jolly Learning, 2013) illustrating *ant* (/æ/) vs. *umbrella* (/ʌ/), and explain the articulatory differences between the two vowels:
  - /æ/: front tongue is visible from the front, and lip shape is reverse triangle.
  - /ʌ/: tongue is positioned like Japanese /o/, but lip shape is similar to Japanese /a/.
- Students imitated vowels, practiced in pairs, and checked each other's pronunciation, tongue and lip positions.

WEEK 2 (Face-mesh software training):

- An instructor introduced web-based face mesh program (Shitara et al., 2023).
- Training emphasized:
  - Open the mouth wider than Japanese /a/, and raise mouth corners for /æ/.
  - Advance the front of the tongue for /æ/, and confirm the movement via visual feedback.
- Students practiced individually with real-time webcam feedback and scoring. The instructor observed students activities, and gave them oral feedback.

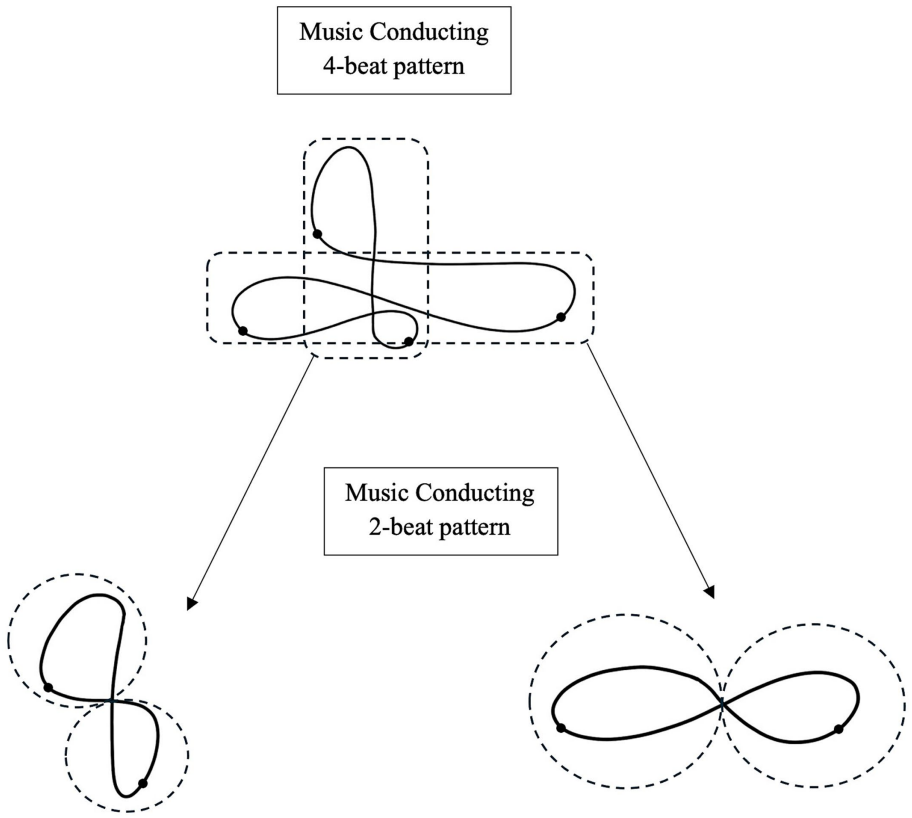

FIGURE 3
Circular motions in legato music conducting. Circular motions can be found in the conducting patterns. 2-beat patterns are simplified from the 4-beat pattern. The dots represent the rhythmic point, which systematically corresponds to where the hand starts to rise.

Production testing session

The pronunciation testing session was conducted in a soundproof booth. During the session, participants were seated at a table equipped with a condenser microphone (Audio Technica AT2020) for recording and a monitor for displaying prompt words. The pronunciation data captured by the microphone were transmitted to a laboratory computer via an audio interface (Focusrite Scalett Solo 2nd Gen) and recorded using Praat (Boersma and Weenink, 1992–2024) with a sampling frequency of 44,100 Hz.

In the paragraph-reading task, the tongue twister "Betty Botter," which was used in the training session, was also employed in the pronunciation testing. During the testing, the tongue twister was displayed on the monitor, and participants were instructed to read it aloud one time. They were not instructed to use any hand gestures, allowing us to assess their performance independently of gesture use and thereby isolate the effects of the training. In the picture-naming task, three pictures for "batter" and three for "butter" were selected. These pictures were presented on the monitor once each in random order. Participants were required to name the item depicted in each picture, thereby determining whether it was "batter" or "butter."

Experimental procedure

The entire experimental procedure consisted of two training phases and three test sessions over a four-week period. The schedule of the experiment is given in Figure 7. Before the first training phase, participants completed a pre-test, which included a paragraph-reading task and a picture-naming task, lasting approximately 20 min in total. Following the pre-test, the HM group underwent hand gesture training (HGT), while the MH group received mouth gesture training (MGT). Training sessions were conducted twice weekly in a classroom setting under instructor supervision, with each session lasting 20 min. After completing four training sessions (totalling 80 min), participants took a mid-test, which was identical to the pre-test. Participants then entered a second two-week training phase, in which the other type of training was implemented: the HM group received MGT, and the MH group underwent HGT. After completing another four training sessions, participants took a post-test, which was consistent with the pre-test and mid-test procedures.

Measurements and analyses

In the reading task, 120 tokens (40 participants × 3 training phrases × 1 repetition) were collected. The total reading duration for each participant of the "Betty Botter" text was calculated, including all types of pauses and repetitions. About the types of pause, only silent pauses were observed. An examination for filled pauses yielded none, likely because the speakers had already become familiar with the texts. In the picture-naming task, a total of 720 tokens (40 participants × 2 vowel stimuli × 3 training phrases × 3 repetitions) were collected. In the words "batter" and "butter," second formant (F2) of the vowels /æ/ and /ʌ/ was manually annotated and measured using Praat.

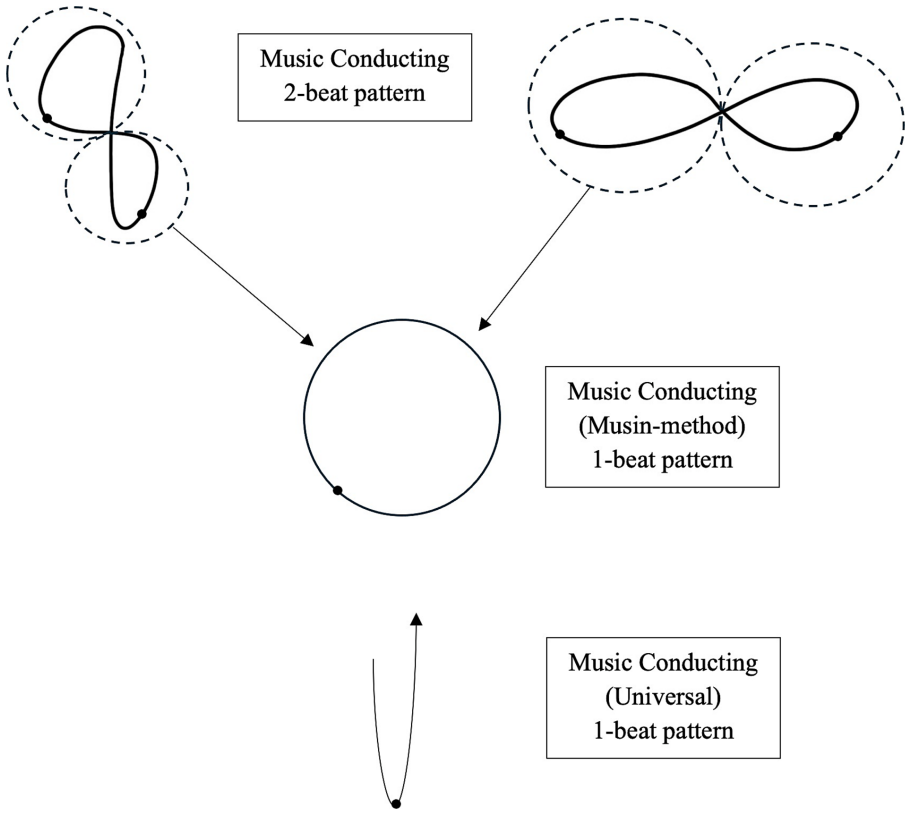

FIGURE 4
Circular motions for 1-beat pattern in the Musin-method. 1-beat patterns are decoupled from the 2-beat patterns in Musin-method. The dot representing the beat point remains at the same place in the gesture, which corresponds to where the hand starts to rise. The circular motion provides a continuous movement, compared to the universal pattern of beating up and down.

Speech rate

The speech rate was calculated by dividing the fixed total of 63 syllables in the "Betty Botter" text by each participant's total reading duration (in seconds). The result is expressed as syllables per second (SPS).

Second formant

When using Praat for F2 measurements, parameter settings were as follows: the number of formants was set to 5, and the window length was configured at 40 milliseconds. Furthermore, according to the frequency characteristics of participants' voices, the formant ceiling value was fine-tuned within the range of 5,000–6,000 Hz to achieve optimal formant tracking.

The F2 values were measured at the midpoint of the intervals annotated for the vowels /æ/ and /ʌ/ in the words "batter" and "butter." The onset of the intervals was defined as the first appearance of a periodic waveform following the consonant /b/, and the offset was marked as the last point of the periodic waveform prior to the consonant /t/. The raw F2 values were normalized using the Lobanov method, as implemented in NORM (Thomas and Kendall, 2007), based on each participant's F2 values measured three times under all conditions, to reduce individual differences due to physiological structure. Subsequently, the normalized F2 values obtained from the three measurements under all conditions were averaged and utilized for statistical analysis.

Statistical analyses

For the speech rate measurements, a two-way mixed-design ANOVA was conducted to examine the effects of training phase (pre, mid, post) as one within-subjects factor, and training sequence (HM group vs. MH group) as one between-subjects factor. For the normalized F2 measurements, a three-way mixed-design ANOVA was conducted to examine the effects of vowel type (/æ/, /ʌ/) and training phase (pre, mid, post) as two within-subjects factors, and training sequence (HM group vs. MH group) as one between-subjects factor.

Both statistical analyses were conducted under the assumption of sphericity, as confirmed by Mauchly's test ($p > 0.05$). Bonferroni correction was applied in *post hoc* pairwise comparisons to control the family-wise error rate. The significance level (α) was set to 0.05. All statistical analyses were performed using JASP (version 0.19.2).

Result

As hypothesized, hand training facilitates suprasegmental improvement, as speech rate increased only following hand training in both groups (Figure 8). A two-way mixed ANOVA revealed a significant Group × Time interaction on speech rate ($F(2,76) = 9.28$, $p < 0.001$, $\eta^2 = 0.044$). The main effect of Time and that of Group were also significant. Post-hoc tests revealed training- and Group-specific effects on the speech rate. For MH group, speech rate was higher in Post test than in Pre test ($t = 3.48$, $p = 0.004$, Cohen's d = 0.69) or in

10.3389/fcomm.2025.1620465

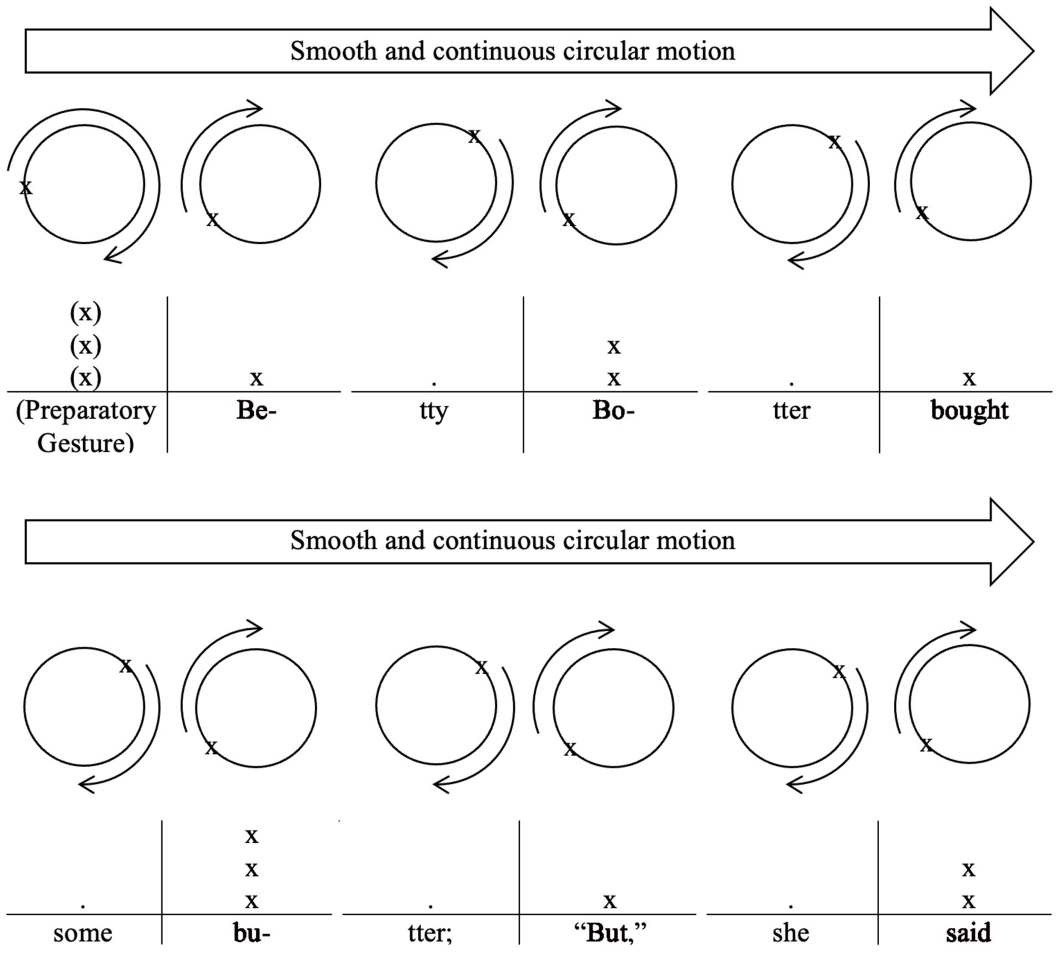

FIGURE 5
Stressed syllable alignment. The figure illustrates a sequence for a spoken phrase with smooth and continuous circular manual motion, with strokes placed on each stressed, similarly to the raising phase of 1-beat pattern in Figures 3, 4. The *raising* phase (x) of the circular gesture coincides with stress (e.g., *Betty Botter bought some butter*), highlighting the cyclic rhythm of English.

Mid test ($t = 3.45$, $p = 0.004$, d = 0.56), whereas it was not different between Pre and Mid tests ($t = 0.79$, $p = 0.105$, d = 0.14). For HM group, speech rate was higher in Mid test than in Pre test ($t = 3.64$, $p = 0.002$, d = 0.66) or in Post test ($t = 3.27$, $p = 0.007$, d = 0.56). The difference between the Pre and Post tests were not significant ($t = 0.446$, $p = 0.661$, d = 0.10). These results suggest that only Hand training improved the speech rate for both Groups.

The mouth training improved F2 value of the vowels /æ/ and /ʌ/ in the words "batter" and "butter" (Figure 9). The 3-way mixed ANOVA revealed a significant Vowel × Time interaction on Lobanov-normalized F2 value ($F(2,76) = 6.26$, $p = 0.003$, $\eta^2 = 0.038$). The main effects of Time ($F(2,76) = 16.38$, $p < 0.001$, $\eta^2 = 0.13$) and Vowel ($F(2,76) = 22.33$, $p < 0.001$, $\eta^2 = 0.098$) were also statistically significant.

Note that we collapsed the training Group since none of the Group-related interactions nor the main effect of Group were significant. For the vowel /ʌ/ in "butter," no significant differences were revealed by the post-hoc pairwise comparisons. For the vowel /æ/ in "batter," significant differences were observed between Pre and Post tests ($t = 5.49$, $p < 0.001$, d = 1.53) and Mid and Post tests ($t = 4.09$, $p < 0.001$, d = 0.99).

Discussion

This study tested two predictions: first, that hand gesture training would enhance suprasegmental fluency while mouth gesture training would improve segmental accuracy; and second, that the timing of training would shape the trajectory of improvement. The results supported the first prediction: hand training facilitated speech-rate gains, and mouth training contributed to F2 improvements. The second prediction, however, was not supported, as no significant group-specific timing effects were observed. Each research question and its corresponding results are discussed in detail below.

We postulated research question (i) How do different types of gesture training differentially influence segmental and suprasegmental aspects of L2 speech? Regarding this question, we hypothesized that different types of motor training would selectively facilitate distinct aspects of speech production: hand training would enhance suprasegmental features (e.g., speech rate), while mouth training would enhance segmental features (e.g., vowel articulation, measured by F2). The findings suggest that hand- and mouth-gesture training exert selective influences on different levels of the phonological hierarchy. Mouth gestures facilitated segmental improvement, as

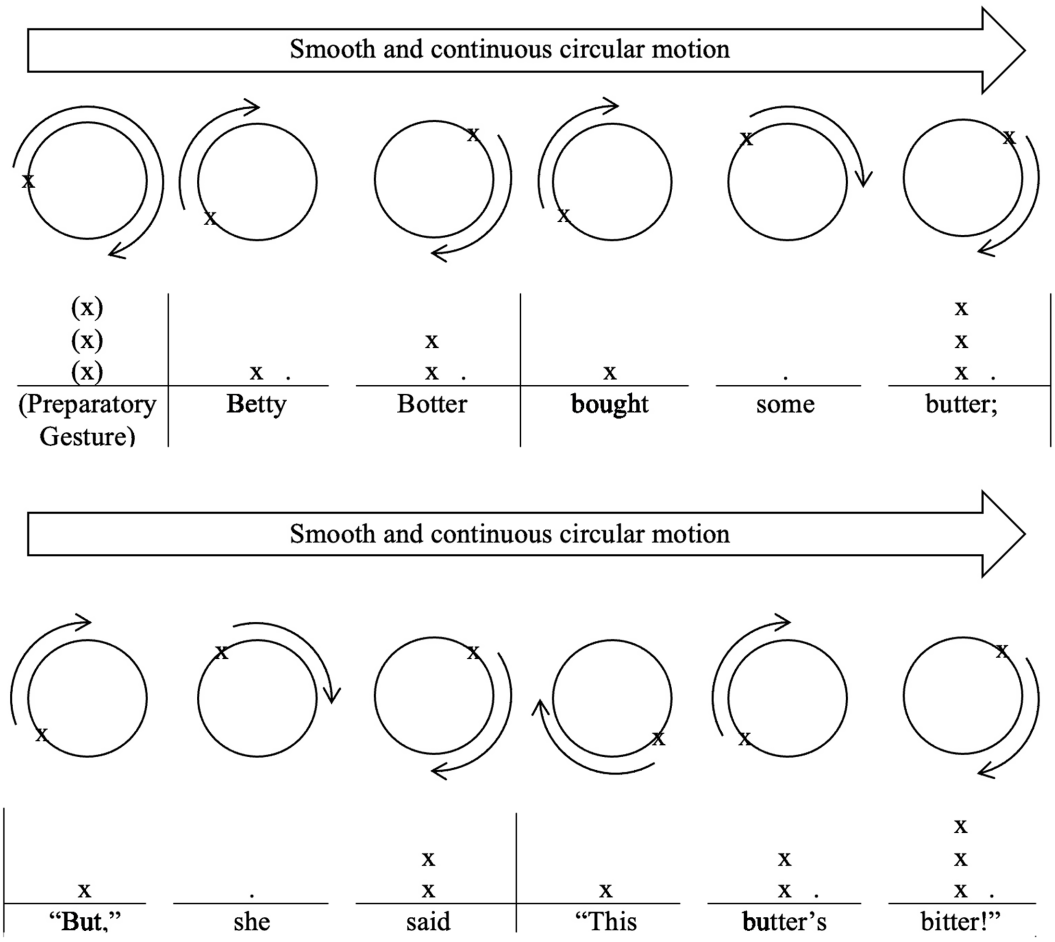

FIGURE 6
Phonological phrase alignment. The figure illustrates a sequence for a spoken phrase with smooth and continuous circular manual motion, with strokes placed on each phonological, similarly to the raising phase of 1-beat pattern in Figures 3, 4. The *raising* phase (x) of the circular gesture coincides with the first stressed syllable within a phrase (e.g., *[Betty Botter] [bought some butter]*), highlighting the cyclic rhythm of English.

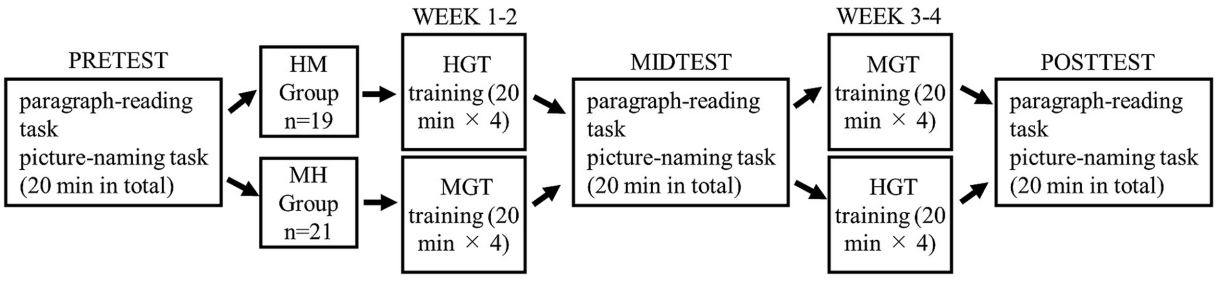

FIGURE 7
Flowchart detailing an experimental design. After completing the pretest (paragraph-reading task and picture-naming task), participants were assigned to either the Hand-first (HM) group or the Mouth-first (MH) group. Each group received their initial gesture training during Weeks 1 and 2, followed by a mid-test. In Weeks 3 and 4, the HM group received mouth gesture training and the MH group received hand gesture training. The posttest was administered after the second phase of training.

shown by clearer /æ/–/ʌ/ contrasts, while hand gestures enhanced suprasegmental fluency, reducing disfluencies and promoting smoother prosodic flow. These results align with neurophysiological evidence that delta-band oscillations (supporting phrasal rhythm and fluency) entrain rapidly to external gestures, whereas theta-band oscillations (supporting syllable-level articulation) require finer motor control and more sustained practice. In this sense, the two gesture modalities appear to direct learners' attention to different units of speech—mouth gestures to vowel articulation within syllables, and hand gestures to phrasing and timing at higher prosodic levels.

Recurrent hand gestures intersect with structural, semantic, and embodied dimensions of our learning targets, providing multiple

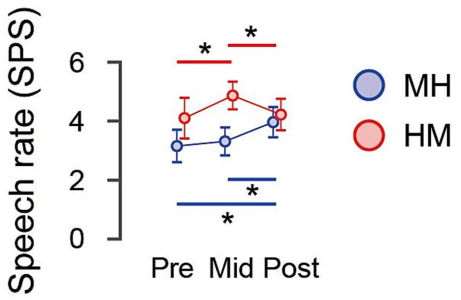

FIGURE 8
Comparison of speech rates of two groups. Speech rate was assessed as the number of syllables per second (SPS). MH group trained with mouth gesture in the first training period (between the Pre and Mid tests), and trained with hand gesture in the second training period (between the Mid and Post tests). *$p < 0.05$.

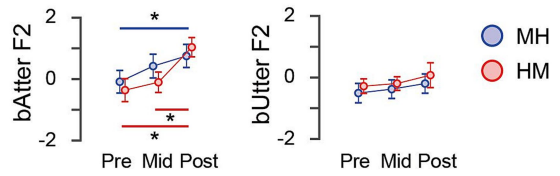

FIGURE 9
Comparison of F2 of two groups. Lobanov-normalized F2 values for the vowels /æ/ and /ʌ/ in the words "batter" and "butter." MH group trained with mouth gesture in the first training period (between the Pre and Mid tests), and trained with hand gesture in the second training period (between the Mid and Post tests). *$p < 0.05$.

layers of support: they can highlight prosodic structure, convey metaphorical meaning, and engage sensorimotor systems that reinforce learning. Structurally, cyclic hand motions align with prosodic phrases, embodying the recursive organization of language; one-circle-per-phrase movement, adapted from music conducting, marks phrase boundaries and facilitates prosodic flow (Selkirk, 1984; Hauser et al., 2002; Martins et al., 2017; Temperley, 2022). Semantically, circular motion metaphorically represents smoothness and continuity, reinforcing the sense of flow in speech; such gestures also appear spontaneously in conversation, when speakers searching for words often employ cyclic hand movements (Ladewig, 2011; Ladewig and Bressem, 2013), consistent with conceptual metaphor theory (Lakoff and Johnson, 1980; Johnson and Lakoff, 2002). Embodied, the physical enactment of such gestures provides proprioceptive and visual feedback that integrates manual and oral movements, strengthening the sensorimotor foundations of fluent speech production (Acton et al., 2013; Pouw et al., 2021; Yu et al., 2024). Together, these dimensions illustrate how recurrent gestures serve as a multimodal scaffold for speech learning.

Mouth-gesture training improved segmental accuracy, as learners enhanced the /æ/–/ʌ/ contrast by advancing the tongue and spreading the lips, supported by real-time biovisual feedback. Although such feedback on lingual gestures has rarely been reported, these results align with evidence that visual monitoring and self-correction can effectively refine segmental production (Gick et al., 2008; Suemitsu et al., 2015; Katz and Mehta, 2015; Yamane et al., 2025). In this way, mouth- and hand-gesture training yielded selective benefits: mouth training improved vowel contrast, while hand training enhanced fluency. These distinct effects reflect the prosodic hierarchy, where segmental and suprasegmental units are governed by separate rules. By directing learners' attention to the relevant level, gesture-based training facilitated targeted gains in L2 pronunciation.

We also posed question (ii) How does the timing of gesture training (hand-first vs. mouth-first) influence the trajectory of improvement across training phases? We predicted that introducing hand training earlier would yield earlier fluency gains, while introducing mouth training earlier would yield earlier segmental gains. The results showed that gesture training overall facilitated improvement in both domains: hand gestures enhanced suprasegmental fluency, and mouth gestures contributed to segmental accuracy. However, no Group-related interactions were significant. This indicates that the order of training did not affect the trajectory of improvement. In other words, although different types of gesture training benefitted different aspects of pronunciation, their effectiveness was not dependent on whether they were introduced first or second.

Although no Group × Time interactions reached significance, two descriptive patterns warrant brief discussion, as they may inform future research on gesture-based training. First, although mouth training was expected to yield immediate F2 gains when introduced early, such improvements did not seem to emerge in the MH group. One possible contributing factor is learner proficiency: as noted in the Method section, the MH group had lower average TOEIC scores than the HM group. Descriptive data of raw subgroup means (Table A1) further suggest that subgroup variability, particularly among MH males, may have influenced the trajectory of vowel accuracy gains. Learners with lower proficiency may require more extensive practice and auditory–motor feedback before segmental adjustments such as /æ/ can be reliably achieved (Flege et al., 1997; Alshangiti and Evans, 2024). Nonetheless, given the absence of significant Group-related interactions, these observations remain exploratory and should be investigated in future research.

Second, the HM group appeared to show a descriptive decline in speech rate at posttest. One possible explanation is that learners may face attentional limits when balancing fluency and segmental refinement, resulting in a temporary trade-off. Although the Group-related interactions were not significant, these descriptive observations likewise remain exploratory and should be investigated in future research with larger and more homogeneous samples.

Pedagogical implications

The results point to the potential of gesture-based training as a targeted supplement to pronunciation instruction. Rather than treating pronunciation as a uniform skill, training can be designed to address suprasegmental and segmental development in complementary ways. Hand gestures may provide an accessible entry point for building fluency across proficiency levels, whereas mouth gestures may be more effective for learners who already possess the proficiency needed for fine-grained articulatory adjustments. The HM group's improvement in vowel accuracy may have benefited from the fluency gains fostered by hand gestures, a pattern consistent with finding by Li et al. (2023),

who showed that hand-based gesture training targeting suprasegmental features also led to improvements at the segmental level. Although the specific suprasegmental measures differed—intonation in Li et al. and speech rate in the present study—both sets of findings converge on the idea that suprasegmental-focused training can create favorable conditions for subsequent segmental improvement. More broadly, these findings echo evidence that suprasegmental-focused training often produces more noticeable gains in listener judgments than segmental drilling alone (Gordon and Darcy, 2019).

Sequencing hand and mouth training could further maximize their complementary benefits. At the same time, descriptive patterns observed in this study suggest that learner variability may influence training outcomes, underscoring the need for flexible instructional designs. Taken together, these findings demonstrate how gesture-based training can differentially support segmental and suprasegmental development, offering a basis for more nuanced and effective approaches to L2 pronunciation pedagogy.

Limitations

While the present findings offer important insights into the developmental relationship between suprasegmental and segmental features in adult L2 speech, several limitations should be acknowledged. This study focused on two dependent variables—speech rate as a proxy for suprasegmental development and F2 values as a proxy for segmental articulation—which, although informative, cannot capture the full range of prosodic and articulatory changes involved in pronunciation learning. Future research should therefore include additional measures such as intonation, pitch range, stress placement, syllable duration, or consonant clarity to provide a more comprehensive picture of learning trajectories. Moreover, the training types (hand versus mouth gestures) were operationalized as broad modalities, yet the cognitive load, motor demands, and degree of linguistic integration likely varied across participants. Further studies could explore how differences in task complexity and attentional demands influence outcomes, helping to clarify the mechanisms that support learning. Finally, the relatively small sample size may have limited statistical power, possibly obscuring interaction effects or moderating influences. Addressing these issues will be essential for advancing our understanding of how gesture-based training supports L2 pronunciation development.

Conclusion

This study provides evidence that gesture-based training can differentially support suprasegmental and segmental aspects of L2 pronunciation. Hand gestures facilitated gains in fluency, while mouth gestures contributed to improvements in vowel articulation, as reflected in F2 values. Although no group-specific timing effects were observed, the overall pattern suggests that hand and mouth gestures provide distinct yet complementary benefits.

From a cognitive perspective, these findings are consistent with the view that speech and gesture form an integrated system in which multiple rhythmic and articulatory processes jointly shape language production. The distinct effects of hand and mouth gestures suggest that prosodic framing and articulatory refinement engage partly independent yet coordinated sensorimotor routines.

Pedagogically, when viewed through an interactional lens, gesture-based instruction can heighten learners' awareness of fluency as an integrative outcome—emerging from the coordination of intrapersonal (mind–body) and interpersonal (speaker–interlocutor) processes. This view further reinforces the notion of inter-fluency (Kosmala, 2024), which conceptualizes fluency as a multidimensional phenomenon encompassing speech, interaction, and gesture. Through haptic feedback from manual and lingual–labial gestures, learners can monitor and adjust their articulatory movements to maximize visible cues that support mutual intelligibility. Such embodied and socially attuned adjustments help synchronize gesture, speech, and facial expression, promoting fluency as a jointly managed, multimodal skill that integrates precision, rhythm, and interactive alignment. Descriptive patterns further suggest that individual differences in learners' sensitivity to gesture or articulatory feedback may influence training effectiveness, highlighting a valuable direction for future research.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Ethics statement

The studies involving humans were approved by Research Ethics Review Board, Graduate School of Humanities and Social Sciences. The studies were conducted in accordance with the local legislation and institutional requirements. The participants provided their written informed consent to participate in this study. Written informed consent was obtained from the individual(s) for the publication of any potentially identifiable images or data included in this article.

Author contributions

NY: Investigation, Supervision, Project administration, Writing – review & editing, Funding acquisition, Writing – original draft, Methodology, Validation, Visualization, Formal analysis, Conceptualization, Data curation, Resources. MS: Conceptualization, Resources, Validation, Funding acquisition, Writing – review & editing, Methodology, Visualization, Writing – original draft. XT: Supervision, Investigation, Data curation, Writing – original draft, Writing – review & editing, Visualization. AC: Conceptualization, Resources, Writing – original draft, Visualization, Funding acquisition, Methodology, Writing – review & editing.

Funding

The author(s) declare that financial support was received for the research and/or publication of this article. This work was supported by JSPS KAKENHI Grant Number 22K00621, 25K04168, and Hiroshima University Promotion grant of Integrated Arts and Sciences project.

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 no Gen AI was used in the creation of this manuscript.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

Publisher's note

References

Acton, W., Baker, A., and Burri, M. and Teaman, B., (2013). "Preliminaries to haptic-integrated pronunciation instruction." In: J. Levis and K. LeVelle, (eds.), Proceedings of the 4th Pronunciation in Second Language Learning and Teaching Conference, August 2012, Ames, IA: Iowa State University, pp. 234–244.

Alshangiti, W., and Evans, B. (2024). Learning English vowels: the effects of different phonetic training modes on Arabic learners' production and perceptiona. *J. Acoust. Soc. Am.* 156, 284–298. doi: 10.1121/10.0026451

Antolík, T. K., Pillot-Loiseau, C., and Kamiyama, T. (2019). The effectiveness of real-time ultrasound visual feedback on tongue movements in L2 pronunciation training: Japanese learners' progress on the French vowel contrast/y/—/u/. *J. Second Lang. Pronunc.* 5, 72–97. doi: 10.1075/jslp.16022.ant

Barsalou, L. W. (2008). Grounded cognition. *Annu. Rev. Psychol.* 59, 617–645. doi: 10.1146/annurev.psych.59.103006.093639

Biau, E., and Soto-Faraco, S. (2015). Synchronization by the hand: the sight of gestures modulates low-frequency activity in brain responses to continuous speech. *Front. Hum. Neurosci.* 9:527. doi: 10.3389/fnhum.2015.00527

Biau, E., Torralba, M., Fuentemilla, L., de Diego-Balaguer, R., and Soto-Faraco, S. (2015). Speaker's hand gestures modulate speech perception through phase resetting of ongoing neural oscillations. *Cereb. Cortex* 26, 246–256. doi: 10.1093/cercor/bhu207

Boersma, P., and Weenink, D., (1992–2024). Praat: doing phonetics by computer [computer program]. Available online at: http://www.praat.org. Version 6.4.22 (Accessed October 5, 2024).

Browman, C., and Goldstein, L. (1986). Towards an articulatory phonology. *Phonology* 3, 219–252. doi: 10.1017/S0952675700000658

Browman, C., and Goldstein, L. (1989). Articulatory gestures as phonological units. *Phonology* 6, 201–251. doi: 10.1017/S0952675700001019

Cavicchio, F., and Busà, M. G. (2023). Lending a hand to speech: gestures help fluency and increase pitch in second language speakers. *Lang. Interact. Acquis.* 14, 218–246. doi: 10.1075/lia.22023.cav

Cummins, F., and Port, R. (1998). Rhythmic constraints on stress timing in English. *J. Phon.* 26, 145–171. doi: 10.1006/jpho.1998.0070

de La Cruz-Pavía, I., Gervain, J., Vatikiotis-Bateson, E., and Werker, J. F. (2020). Coverbal speech gestures signal phrase boundaries: a production study of Japanese and English infant- and adult-directed speech. *Lang. Acquis.* 27, 160–186. doi: 10.1080/10489223.2019.1659276

de Jong, N. H. (2023). Fluency in speaking as a dynamic construct. *Lang. Teach. Res. Q.* 37, 179–187. doi: 10.32038/ltrq.2023.37.09

Doelling, K. B., Arnal, L. H., Ghitza, O., and Poeppel, D. (2014). Acoustic landmarks drive delta–theta oscillations to enable speech comprehension by facilitating perceptual parsing. *NeuroImage* 85, 761–768. doi: 10.1016/j.neuroimage.2013.06.035

Flege, J., Bohn, O., and Jang, S. (1997). Effects of experience on non-native speakers' production and perception of English vowels. *J. Phon.* 25, 437–470. doi: 10.1006/JPHO.1997.0052

Gentilucci, M., and Volta, R. (2008). Spoken language and arm gestures are controlled by the same motor control system. *Q. J. Exp. Psychol.* 61, 944–957. doi: 10.1080/17470210701625683

Gick, B., Bernhardt, B. M., Bacsfalvi, P., and Wilson, I. (2008). "11. Ultrasound imaging applications in second language acquisition" In: J. G. Hansen Edwards and M. L. Zampini, (eds.) Phonology and second language acquisition (John Benjamins Publishing Company), 309–322.

Gick, B., Wilson, I., Koch, K., and Cook, C. (2004). Language-specific articulatory settings: evidence from inter-utterance rest position. *Phonetica* 61, 220–233. doi: 10.1159/000084159

Giraud, A. L., and Poeppel, D. (2012). "Speech perception from a neurophysiological perspective" In: D. Poeppel, T. Overath, A. N. Popper and R. R. Fay, eds. The human auditory cortex (New York, NY: Springer New York), 225–260.

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

Goldin-Meadow, S., and Alibali, M. (2013). Gesture's role in speaking, learning, and creating language. *Annu. Rev. Psychol.* 64, 257–283. doi: 10.1146/annurev-psych-113011-143802

Goldsmith, J. A. (1976). Autosegmental phonology (Doctoral Dissertation): Massachusetts Institute of Technology, Massachusetts.

Goodwin, C. (2007). Participation, stance and affect in the organization of activities. *Discourse Soc.* 18, 53–73. doi: 10.1177/0957926507069457

Gordon, J., and Darcy, I., (2019). Teaching segmentals vs. suprasegmentals: different effects of explicit instruction on comprehensibility. In: J. Levis, C. Nagle and E. Todey, eds. Proceedings of the 10th pronunciation in second language learning and teaching conference, Ames, IA, September 2018. Ames, IA: Iowa State University, pp.116–126.

Hauser, M., Chomsky, N., and Fitch, W. (2002). The faculty of language: what is it, who has it, and how did it evolve? *Science* 298, 1569–1579. doi: 10.1126/science.298.5598.1569

Hayes, B. (1995). Metrical stress theory: principles and case studies. Chicago: University of Chicago Press.

Hillenbrand, J., Getty, L. A., Clark, M. J., and Wheeler, K. (1995). Acoustic characteristics of American English vowels. *J. Acoust. Soc. Am.* 97, 3099–3111. doi: 10.1121/1.411872

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

Jansson, D., Balsnes, A., and Durrant, C. (2021). The gesture enigma: reconciling the prominence and insignificance of choral conductor gestures. *Res. Stud. Music Educ.* 44, 509–526. doi: 10.1177/1321103X211031778

Johnson, M., and Lakoff, G. (2002). Why cognitive linguistics requires embodied realism. *Cogn. Linguist.* 13, 245–264. doi: 10.1515/cogl.2002.016

Jolly Learning, (2013). Jolly phonics letter sounds (American English). Jolly learning – The home of jolly phonics. Available online at: https://youtu.be/3LD7m3luv0Y?si=PQ-HddN7ygWsMCk6 (Accessed September 6, 2025).

Katz, W. F., and Mehta, S. (2015). Visual feedback of tongue movement for novel speech sound learning. *Front. Hum. Neurosci.* 9:612. doi: 10.3389/fnhum.2015.00612

Kawase, S., Davis, C., and Kim, J. (2024). Impact of Japanese L1 rhythm on English L2 speech. *Lang. Speech* 68, 118–140. doi: 10.1177/00238309241247210

Kendon, A. (2004). Gesture: visible action as utterance. Cambridge: Cambridge University Press.

Kilpatrick, C. E. (2020). Movement, gesture, and singing: a review of literature. *Update Appl. Res. Music Educ.* 38, 29–37. doi: 10.1177/8755123320908612

Kita, S., Alibali, M., and Chu, M. (2017). How do gestures influence thinking and speaking? The gesture-for-conceptualization hypothesis. *Psychol. Rev.* 124, 245–266. doi: 10.1037/rev0000059

Kocjančič, T., Bořil, T., and Hofmann, S. (2024). Acoustic and articulatory visual feedback in classroom L2 vowel remediation. *Lang. Speech*:00238309231223736. doi: 10.1177/00238309231223736

Kosmala, L., Horgues, C., and Scheuer, S. (2023). 'A multimodal study of how pronunciation-induced communication breakdowns are managed during tandem interactions' *Research in Language* 21, 291–312. doi: 10.18778/1731-7533.21.3.05

Kosmala, L. (2024). Beyond disfluency: the interplay of speech, gesture, and interaction. Amsterdam / Philadelphia: John Benjamins.

Kubozono, H. (2017). "Mora and syllable" in The handbook of Japanese linguistics, Oxford: Blackwell. 31–61. doi: 10.1002/9781405166225

Kushch, O., Igualada, A., and Prieto, P. (2018). Prominence in speech and gesture favour second language novel word learning. *Lang. Cogn. Neurosci.* 33, 992–1004. doi: 10.1080/23273798.2018.1435894

Ladewig, S. H. (2011). "Putting the cyclic gesture on a cognitive basis" In: M. Lemmens, ed. CogniTextes. Revue de l'Association française de linguistique cognitive. Villeneuve-d'Ascq, France.

Ladewig, S. H., and Bressem, J. (2013). New insights into the medium hand: discovering recurrent structures in gestures. *Semiotica* 2013, 203–231. doi: 10.1515/sem-2013-0088

Lakoff, G., and Johnson, M. (1980). The metaphorical structure of the human conceptual system. *Cogn. Sci.* 4, 195–208. doi: 10.1016/S0364-0213(80)80017-6

Lambacher, S. G., Martens, W. L., Kakehi, K., Marasinghe, C. A., and Molholt, G. (2005). The effects of identification training on the identification and production of American English vowels by native speakers of Japanese. *Appl. Psycholinguist.* 26, 227–247. doi: 10.1017/S0142716405050150

Lennon, P. (1990). Investigating fluency in EFL: a quantitative approach. *Lang. Learn.* 40, 387–417. doi: 10.1111/j.1467-1770.1990.tb00669.x

Leonard, T., and Cummins, F. (2011). The temporal relation between beat gestures and speech. *Lang. Cogn. Process.* 26, 1457–1471. doi: 10.1080/01690965.2010.500218

Li, P., Baills, F., Baqué, L., and Prieto, P. (2023). The effectiveness of embodied prosodic training in L2 accentedness and vowel accuracy. *Second. Lang. Res.* 39, 1077–1105. doi: 10.1177/02676583221124075

Li, P., Baills, F., and Prieto, P. (2020). Observing and producing durational hand gestures facilitates the pronunciation of novel vowel-length contrasts. *Stud. Second. Lang. Acquis.* 42, 1015–1039. doi: 10.1017/S0272263120000054

Li, P., Xi, X., Baills, F., and Prieto, P. (2021). Training non-native aspirated plosives with hand gestures: learners' gesture performance matters. *Lang. Cogn. Neurosci.* 36, 1313–1328. doi: 10.1080/23273798.2021.1937663

Liberman, M., and Prince, A. (1977). On stress and linguistic rhythm. *Linguist. Inq.* 8, 249–336. doi: 10.2307/4177987

Llanes-Coromina, J., Prieto Vives, P., and Rohrer, P. L. (2018). 'Brief training with rhythmic beat gestures helps L2 pronunciation in a reading-aloud task', in *Proceedings of the 9th International Conference on Speech Prosody (SpeechProsody 2018)*, Poznań, Poland, 13–16 June, pp. 498–502. doi: 10.21437/SpeechProsody.2018-101

Ma, S., and Jin, G. (2022). The relationship between different types of co-speech gestures and L2 speech performance. *Front. Psychol.* 13:941114. doi: 10.3389/fpsyg.2022.941114

Martins, M., Gingras, B., Puig-Waldmüller, E., and Fitch, W. T. (2017). Cognitive representation of "musical fractals": processing hierarchy and recursion in the auditory domain. *Cognition* 161, 31–45. doi: 10.1016/j.cognition.2017.01.001

McCafferty, S. G. (2002). Gesture and creating zones of proximal development for second language learning. *Mod. Lang. J.* 86, 192–203. doi: 10.1111/1540-4781.00144

McNeill, D. (1992). Hand and mind: what gestures reveal about thought. Chicago: University of Chicago Press.

Meissl, K., Sambre, P., and Feyaerts, K. (2022). Mapping musical dynamics in space. A qualitative analysis of conductors' movements in orchestra rehearsals. *Front. Commun.* 7:986733. doi: 10.3389/fcomm.2022.986733

Morett, L. M., Feiler, J. B., and Getz, L. M. (2022). Elucidating the influences of embodiment and conceptual metaphor on lexical and non-speech tone learning. *Cognition* 222:105014. doi: 10.1016/j.cognition.2022.105014

Musin, I. (1967). The technique of conducting. Moscow: Muzyka Publishing House.

Ogrizovic-Ciric, M. (2009). Ilya Musin's language of conducting gestures. Doctoral Dissertation. Athens, Georgia: University of Georgia.

Parisse, C., Morgenstern, A., and Caët, S. (2022) Annotating multimodal data: Interactions across semiotic resources, in Proceedings of the thirteenth language resources and evaluation conference. Marseille: European Language Resources Association (ELRA), pp. 2755–2764. Available online at: https://aclanthology.org/2022.lrec-1.297

Peterson, G. E., and Barney, H. L. (1952). Control methods used in a study of the vowels. *J. Acoust. Soc. Am.* 24, 175–184. doi: 10.1121/1.1906875

Port, R. F., Dalby, J., and O'Dell, M. (1987). Evidence for mora timing in Japanese. *J. Acoust. Soc. Am.* 81, 1574–1585. doi: 10.1121/1.394510

Pouw, W., de Jonge-Hoekstra, L., Harrison, S. J., Paxton, A., and Dixon, J. A. (2021). Gesture–speech physics in fluent speech and rhythmic upper limb movements. *Ann. N. Y. Acad. Sci.* 1491, 89–105. doi: 10.1111/nyas.14532

Prieto, P., Kushch, O., Borràs-Comes, J., Gluhareva, D., and Pérez-Vidal, C. (2025). Training ESL students to reproduce beat gestures in discourse leads to L2 pronunciation improvements. *Anu. Semin. Filol. Vasca "Julio Urquijo"* 57, 805–823. doi: 10.1387/asju.25982

Selkirk, E. O. (1980). The role of prosodic categories in English word stress. *Linguist. Inq.* 11, 563–605, Available online at: https://www.jstor.org/stable/4178106

Selkirk, E. (1984). "On the major class features and syllable theory," In: M. Aronoff and R. T. Oehrle, (eds.) *Language Sound Structure: Studies in Phonology Presented to Morris Halle by His Teachers and Students*, Cambridge, MA: MIT Press, pp. 107–136. Available online at: https://www.ai.mit.edu/projects/dm/featgeom/selkirk84-sonor.pdf

Shattuck-Hufnagel, S., and Ren, A. (2018). The prosodic characteristics of non-referential co-speech gestures in a sample of academic-lecture-style speech. *Front. Psychol.* 9:1514. doi: 10.3389/fpsyg.2018.01514

Shitara, T., Kimura, T., Makino, T., and Yamane, N., (2023). Feedback effects of mouth shape during speech training. Acoustical Society of Japan, Kansai branch, youth interaction meeting, December 2023, Kindai University. (In Japanese). Available online at: https://os3-314-46534.vs.sakura.ne.jp/text_modifiable/(Accessed November 22, 2025).

Smotrova, T. (2017). Making pronunciation visible: gesture in teaching pronunciation. *TESOL Q.* 51, 59–89. doi: 10.1002/TESQ.276

Suemitsu, A., Dang, J., Ito, T., and Tiede, M. (2015). A real-time articulatory visual feedback approach with target presentation for second language pronunciation learning. *J. Acoust. Soc. Am.* 138, EL382–EL387. doi: 10.1121/1.4931827

Tajima, K., and Port, R., (2004). 'Speech rhythm in English and Japanese, in Local', In: J. Ogden and R. Temple, (eds.), Papers in Laboratory Phonology VI: Phonetic Interpretation. Cambridge: Cambridge University Press, pp. 322–339. doi: 10.1017/CBO9780511486425.020

Temperley, D. (2022). Music and language. *Annu. Rev. Linguist.* 8, 153–170. doi: 10.1146/annurev-linguistics-031220-121126

Thomas, E.R., and Kendall, T., (2007). NORM: the vowel normalization and plotting suite. Online Resource. Available online at: http://lingtools.uoregon.edu/norm/norm1.php (Accessed April 29, 2025).

Thomson, R.I., (2012). English accent coach [computer program], version 2. Available online at: http://www.englishaccentcoach.com (Accessed September 6, 2025).

Vainio, L. (2019). Connection between movements of mouth and hand: perspectives on development and evolution of speech. *Neurosci. Biobehav. Rev.* 100, 211–223. doi: 10.1016/j.neubiorev.2019.03.005

Wagner, P., Malisz, Z., and Kopp, S. (2014). Gesture and speech in interaction: an overview. *Speech Comm.* 57, 209–232. doi: 10.1016/j.specom.2013.09.008

Wang, J., Gao, Y., and Cu, Y. (2023). Classroom gesture instruction on second language learners' academic presentations: evidence from Chinese intermediate English learners. *J. Engl. Acad. Purp.* Vol 66. doi: 10.1016/j.jeap.2023.101304

Wilson, I., Perkins, J., Sato, A., and Ishii, D. (2025). Articulatory settings of Japanese–English bilinguals. *Lang. Speech*:00238309251353727. doi: 10.1177/00238309251353727

Xi, X., Li, P., and Prieto, P. (2024). Improving second language vowel production with hand gestures encoding visible articulation: evidence from picture-naming and paragraph-reading tasks. *Lang. Learn.* 74, 884–916. doi: 10.1111/lang.12647

Yamane, N., Shinya, M., Teaman, B., Ogawa, M., and Akahoshi, S., (2019). Mirroring beat gestures: effects on EFL learners. In: Proceedings of the 19th International Congress of Phonetic Sciences (ICPhS 2019). Melbourne, Australia, pp.3523–3527.

Yamane, N., Sun, K., Perkins, J., Wilson, I., and Tan, X. (2025) Ultrasound pronunciation training: preset-posttest production and discrimination results. *Journal of Monolingual and Bilingual Speech*, University of Toronto Press. 6, doi: 10.3138/jmbs-25261-yamane

Yazawa, K., and Kondo, M., (2019). Acoustic characteristics of Japanese short and long vowels: formant displacement effect revisited. In Proceedings of the 19th international congress of phonetic sciences (pp. 671–675). Canberra, ACT: Australasian Speech Science and Technology Association Inc.

Yu, K., Zhang, J., Li, Z., Zhang, X., Cai, H., Li, L., et al. (2024). Production rather than observation: comparison between the roles of embodiment and conceptual metaphor in L2 lexical tone learning. *Learn. Instr.* 92:101905. doi: 10.1016/j.learninstruc.2024.101905

10.3389/fcomm.2025.1620465

Appendix

TABLE A1 Raw F2 values (Hz) of /æ/ and /ʌ/ across Pre-, Mid-, and Post-tests by group (HM, MH) and gender.

Group	Sex	N	Vowel	F2 (Hz)		
				Pre (M ± SD)	Mid (M ± SD)	Post (M ± SD)
HM group	Female	9	/æ/	1414.05 ±168.21	1506.95 ±181.43	1648.85 ±164.50
			/ʌ/	1485.67 ±133.47	1503.34 ±94.90	1478.77 ± 160.43
	Male	10	/æ/	1322.31 ±265.11	1298.03 ±259.79	1474.09 ± 208.40
			/ʌ/	1204.03 ±116.02	1229.65 ±126.21	1294.64 ± 219.74
MH group	Female	15	/æ/	1503.74 ±180.95	1747.80 ±354.12	1752.95 ±377.24
			/ʌ/	1411.00 ±165.32	1484.42 ±223.70	1429.33 ±226.30
	Male	6	/æ/	1326.59 ±260.52	1338.41 ±194.97	1584.42 ±479.53
			/ʌ/	1185.97 ±67.53	1177.88 ±51.87	1297.94 ±343.05