

Facial Video Cues in Shadowing: Effects on English Speech Rhythm Acquisition by Japanese EFL Learners

MATSUSHITA, Rinako

(Graduate School, Seinan Gakuin University)

NAKANISHI, Hiroshi

(Seinan Gakuin University)

1. Introduction

Shadowing, in which learners repeat spoken input immediately after hearing it, is widely recognized as an effective method for improving second language (L2) listening and speaking fluency (Foote & McDonough, 2017; Hamada, 2018a). Notably, recent studies suggest that shadowing supports the acquisition of prosodic features such as rhythm and intonation—often more efficiently than segmental features (Nakanishi, 2024)—because prosody is perceptually salient and processed holistically (Trofimovich & Baker, 2006).

Several empirical studies have specifically examined how shadowing contributes to the development of L2 rhythm. Hamada (2018b), for instance, implemented a 15-week “haptic-shadowing” program in which 58 Japanese university students at near-intermediate to intermediate levels performed physical gestures synchronized with English stress patterns while shadowing model speech. The training, conducted once a week for 30 minutes, resulted in significant improvements in both segmental accuracy and prosodic fluency, as evaluated by trained, advanced non-native English speakers. From a proficiency-based perspective, Matsushita (2025) demonstrated that only higher-proficiency learners (CEFR B1–B2), but not their lower-proficiency counterparts (A1–A2), significantly reduced the duration of inter-stress intervals (ISIs)—particularly in longer ISI types (ISI 4 and 5)—after shadowing each target sentence five times. The training involved five repetitions of target sentences that systematically varied the number of unstressed syllables between stressed words (e.g., “hear” and “met”), including “*I am glad to hear you met with him*”(ISI2), “*I am glad to hear that you met with him*”(ISI3), “*I am glad to hear that you have met with him*”(ISI4), and “*I am glad to hear that you will have met with him by then*” (ISI5).

With regard to rhythm assessment, while some studies have relied on expert judgment (e.g., Hamada, 2018b), others have adopted objective acoustic metrics such as the inter-stress interval (ISI) and its standard deviation (ISI-SD). ISI refers to the temporal duration between successive stressed syllables. While native English

speakers typically produce ISIs that remain relatively stable across different rhythmic structures (Roach, 1982), Japanese EFL learners—particularly those at lower proficiency levels—tend to produce longer ISIs as more unstressed syllables are inserted, reflecting a syllable-timed (or mora-timed) rhythm pattern (Mochizuki-Sudo & Kiritani, 1991; Konishi & Kondo, 2019). In addition to ISI, the present study adopts ISI-SD, the standard deviation of inter-stress intervals, as a quantitative index of rhythmic variability. Konishi and Kondo (2019) demonstrated that ISI-SD values were smallest for native speakers, followed by advanced learners, and largest for beginner learners. This finding suggests that ISI-SD can serve as a valid measure of rhythmic acquisition in L2 speech.

In addition to auditory input, recent research has explored the role of visual prosodic cues—particularly those derived from a speaker’s face—in facilitating L2 speech perception and production. Nakano et al. (2024) conducted an eye-tracking study with Japanese EFL learners and found that participants reproduced more words during shadowing when viewing a speaker’s moving face than a still image. Gaze data further revealed that lower-proficiency learners predominantly fixated on the speaker’s mouth, whereas higher-proficiency learners distributed their gaze more evenly between the eyes and mouth. These behavioral findings suggest that facial cues play a key role in enhancing auditory-motor integration and fluency during shadowing. Complementing these behavioral findings, Jeong et al. (2023) conducted an fMRI study showing that Japanese EFL learners exhibited greater activation in speech-related brain regions—including the left inferior frontal gyrus (IFG), premotor cortex, and supramarginal gyrus—when shadowing audiovisual facial videos compared to mosaic-masked videos. Furthermore, increased activity was observed in the right fusiform gyrus and the left hippocampus, regions associated with face processing and memory encoding, respectively.

From an articulatory-phonetic perspective, jaw movement reflects the stress patterns of English speech in a systematic way (Erickson et al., 2012). In particular, native English speakers were found to open their jaws more widely when producing syllables with higher metrical stress. This pattern of jaw displacement followed a consistent strong–weak alternation that corresponded to the rhythmic structure of the utterance. The degree of jaw opening also showed a strong correlation with acoustic cues such as the first formant (F1), indicating a link between articulation and perceived stress. These results suggest that jaw movement contributes directly to the timing and rhythm of stress in English speech. Based on this evidence, audiovisual shadowing—by visually presenting articulatory gestures such as jaw opening—may help learners internalize the timing of stress and produce rhythm that more closely approximates native-like English speech.

Taken together, these studies suggest that shadowing training supports the development of L2 speech rhythm, and that incorporating visual input through

audiovisual shadowing may further enhance this effect by promoting multimodal integration. The present study builds on this line of research by investigating whether audiovisual shadowing—featuring visible articulatory gestures such as jaw movement—facilitates the acquisition of English rhythm among Japanese EFL learners. Using both ISI and ISI-SD as acoustic metrics, the study explores how facial cues modulate rhythmic timing across different proficiency levels.

2. Purpose of This Study

This study investigates the role of visual prosodic cues in the acquisition of English speech rhythm by Japanese EFL learners. Specifically, it investigates whether shadowing with visible facial articulation facilitates greater rhythmic improvement compared to shadowing with a face-masked model in which the mouth region is obscured by mosaic processing. The study followed a short-term repeated-measures design consisting of a pretest, two training phases, and two posttests (Post1 and Post2). In each training phase, participants shadowed four target sentences, each repeated five times. To quantify rhythmic development, the study employed temporal measures, namely ISI and ISI-SD at each testing point.

Two primary research questions were addressed:

1. Does shadowing with visible facial movement result in greater improvement in English rhythm than shadowing with mouth-masked video, as measured by ISI or ISI-SD across testing phases?
2. Does learners' proficiency level influence how visual facial cues are utilized during training?

This study aims to clarify the contribution of visible articulatory cues—particularly jaw and lip movements—to the internalization of prosodic timing in L2 speech, and to explore whether the effectiveness of these cues varies depending on learner proficiency.

3. Method

3.1 Participants

A total of 68 Japanese university students (19 males, 49 females) participated in this study. Participants' English listening proficiency was assessed using the Oxford Quick Placement Test – Listening (maximum score = 25), with scores ranging from 8 to 21 ($M = 15.43$, $SD = 3.07$). According to the Common European Framework of Reference for Languages (CEFR), 2 participants were classified as A1, 24 as A2, 35 as B1, and 7 as B2. Participants were assigned to one of two audiovisual shadowing conditions: a full-face video condition (FF group, $n = 36$), in which the speaker's entire face was visible, or a mouth-masked video condition (MM group, $n = 32$), in which the mouth region was obscured by a mosaic filter. An independent-samples t -test revealed no significant difference in listening proficiency between the two groups, $t(66) = 0.84$, ns . Within the FF group, 1 participant was categorized as A1, 11 as A2, 21 as B1, and 3

as B2. In the MM group, the distribution was 1 A1, 13 A2, 14 B1, and 4 B2 learners.

3.2 Procedure

The experiment consisted of the following sequence: a pretest, a first training phase (five shadowing sessions), a posttest (Post1), a second training phase (five additional shadowing sessions), and a second posttest (Post2). During the pretest, Post1, and Post2, participants completed an oral reading test. In each shadowing session, participants were instructed to repeat English utterances while watching videos presented on individual monitors. Depending on the assigned condition, participants performed shadowing using either a full-face video (FF condition) or a mouth-masked video (MM condition). The entire experiment was conducted in a Computer-Assisted Language Learning (CALL) classroom.

3.2.1 Pretest, Posttest1, and Posttest2

Participants were instructed to read aloud eight short sentences (e.g., *My mom will be stopping her car in the park*) printed on a sheet of paper, at their normal speaking pace, pausing briefly between each sentence. Although the same set of sentences was used across all tests to ensure comparability in subsequent analyses, the order of sentence presentation was randomized for each session. Participants' oral readings were audio-recorded for later analysis.

3.2.2 Shadowing Training Sessions

Two shadowing training sessions were conducted: the first was administered between the Pretest and Posttest1, and the second between Posttest1 and Posttest2. During each session, participants were instructed to shadow the speech of a native English speaker (a female American English speaker) while watching video recordings of her articulating the sentences. The same set of eight sentences used in the pre- and posttests was employed in these sessions, with the order randomized within each session.

Each shadowing trial began with the target sentence displayed on the computer monitor, and participants were instructed to memorize the sentence before the shadowing task. Following this, they performed shadowing while watching a video in which the model speaker pronounced the same sentence five times. Participants were assigned to one of two conditions that differed in the visual accessibility of the speaker's mouth. In the FF condition (Full-Face Video), participants viewed a video in which the speaker's entire face, including her mouth, was visible, thereby allowing access to visual articulatory cues such as lip and jaw movement. In contrast, in the MM condition (Mouth-Masked Video), the speaker's mouth area was obscured with a mosaic filter, preventing participants from accessing articulatory information while keeping other facial features visible (See Figure 1).

Figure 1 Video Conditions in the Shadowing Training Session

FF (Full-Face) condition



MM (Mouth-Masked) condition



3.2.3 Materials

For the pretest, Posttest1, and Posttest2, participants were instructed to read four target sentences that systematically manipulated the number of unstressed syllables inserted between two stressed syllables—specifically, between *mom* and *stop*—to create four inter-stress interval (ISI) types: ISI2, ISI3, ISI4, and ISI5. The ISI2 sentence was *My mom is stopping her car in the park*; the ISI3 sentence was *My mom will be stopping her car in the park*; the ISI4 sentence was *My mom will have been stopping her car in the park*; and the ISI5 sentence was *My mommy will have been stopping her car in the park*. These sentences were carefully constructed so that primary phrasal stress would consistently fall on syllables containing the low vowel /a/ (e.g., *mom*, *stop*, *car*, *park*), which is articulated with considerable jaw lowering and thus offers enhanced visual salience for prosodic cues. In addition to the experimental sentences described above, four filler sentences (e.g., *Lonely men were in warm rain*) were also included.

3.3 Analysis and Evaluation Method

Prior to statistical analysis, a Winsorization procedure was applied to reduce the influence of extreme outliers. For each experimental condition, values exceeding ± 2.5 SD from the mean were replaced with the corresponding threshold. This approach is commonly used in psycholinguistic research to reduce outlier impact (e.g., Nicklin & Plonsky, 2020). All linear mixed-effects model (LMM) analyses were performed on the Winsorized dataset.

To assess learners' acquisition of English rhythm, two objective measures were used: inter-stress interval (ISI) and ISI standard deviation (ISI-SD). These metrics capture both the duration and consistency of timing between stressed syllables.

Inter-Stress Interval (ISI)

ISI refers to the duration (in milliseconds) between two stressed syllables within a sentence. In this study, ISI was measured from the onset of the vowel in the first stressed syllable (e.g., *mom*) to the onset of the vowel in the second stressed syllable (e.g., *stop*), across four sentence types with one to four intervening unstressed syllables.

For example, ISI2 corresponds to “*My mom is stopping her car in the park,*” whereas ISI5 corresponds to “*My mommy will have been stopping her car in the park.*”

Inter-Stress Interval Standard Deviation (ISI-SD)

All experimental sentences (ISI2–ISI5) were designed to contain four stressed syllables with the vowel /a/ (e.g., *mom, stop, car, park*). The durations (in milliseconds) between each pair of successive stressed syllables were measured, and the standard deviation of these inter-stress intervals was calculated for each utterance. ISI-SD is assumed to be used as an index of rhythmic stability, with lower values indicating more consistent timing between stressed syllables.

4. Results

4.1 Model Specifications

Four linear mixed-effects models (LMMs) were constructed to examine the effects of training condition (FF vs. MM), testing period (Post1 vs. Post2), ISI type (ISI2–5), and learner proficiency (Upper vs. Lower) on ISI duration and rhythmic variability (ISI-SD). Participants were treated as random intercepts, and pretest scores (Pre_duration or Pre_SD) were entered as covariates to account for baseline differences.

The model formulas were as follows:

ISI Duration: Two-Factor model

$$\text{ISI_duration} \sim \text{Pre_ISI_duration} + \text{Condition} * \text{Period} + \text{ISI} + (1 \mid \text{Participants})$$

Rhythmic Variability (ISI-SD): Two-Factor model

$$\text{ISI_SD} \sim \text{Pre_ISI_SD} + \text{Condition} * \text{Period} + \text{ISI} + (1 \mid \text{Participants})$$

ISI Duration: Three-Factor model

$$\text{ISI_duration} \sim \text{Pre_ISI_duration} + \text{Condition} * \text{Period} * \text{Level} + \text{ISI} + (1 \mid \text{Participants})$$

Rhythmic Variability (ISI-SD): Three-Factor model

$$\text{ISI_SD} \sim \text{Pre_ISI_SD} + \text{Condition} * \text{Period} * \text{Level} + \text{ISI} + (1 \mid \text{Participants})$$

4.2 ISI Duration: Two-Factor Model (Condition × Period)

The two-factor model for ISI duration yielded a marginal R^2 of .760 and a conditional R^2 of .851. As shown in Table 1, pretest ISI duration (Pre_ISI_duration) was a significant predictor of posttest performance ($p < .001$), indicating that baseline inter-stress timing strongly influenced subsequent performance. ISI types also had a positive effect (all $p < .001$), with longer durations observed for sentences with more intervening unstressed syllables. No significant main effects were found for Condition or Period, nor was their interaction significant.

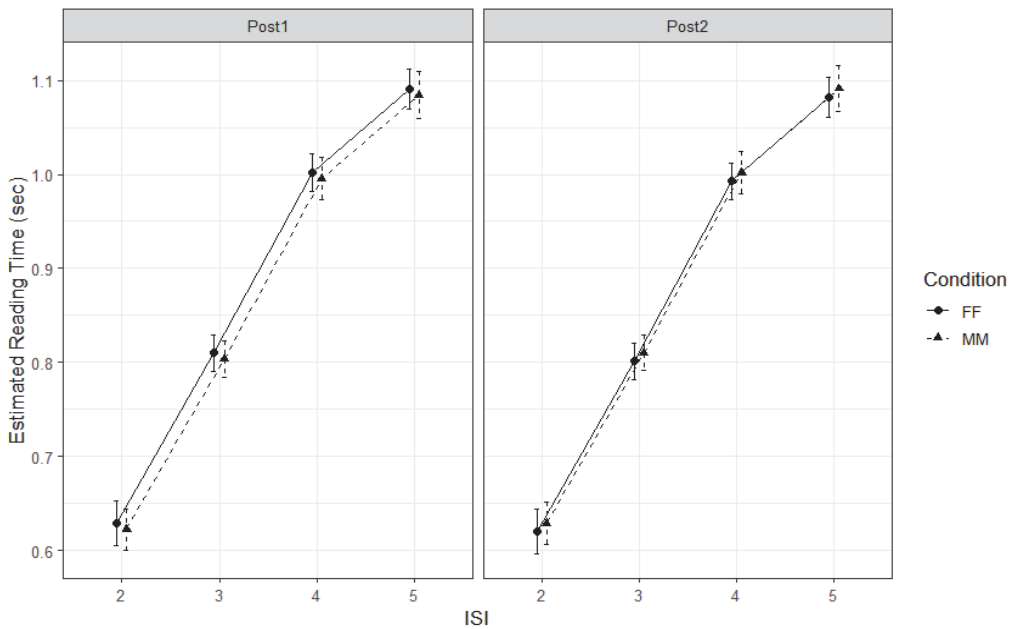
Post hoc comparisons revealed no significant differences between the FF and MM conditions at any ISI type or testing period. Figure 2 displays the estimated ISI durations for each ISI type across the two test periods (Post1 and Post2), separately for

each condition.

Table 1 Fixed Effects for ISI Duration (Two-Factor LMM)

Predictor	Estimate	SE	Df	t	p
(Intercept)	0.477	0.025	264	19.14	< .001
Pre_ISI_duration	0.170	0.037	468	4.61	< .001
Condition (MM)	-0.006	0.024	90	-0.27	.787
Period (Post2)	-0.009	0.013	402	-0.70	.486
ISI 3	0.181	0.015	442	11.74	< .001
ISI 4	0.373	0.027	470	13.97	< .001
ISI 5	0.462	0.030	471	15.42	< .001
Condition × Period	0.015	0.019	402	0.83	.410

Figure 2 Estimated ISI Duration at Post1 and Post2 by Condition



4.3 ISI-SD: Two-Factor Model (Condition × Period)

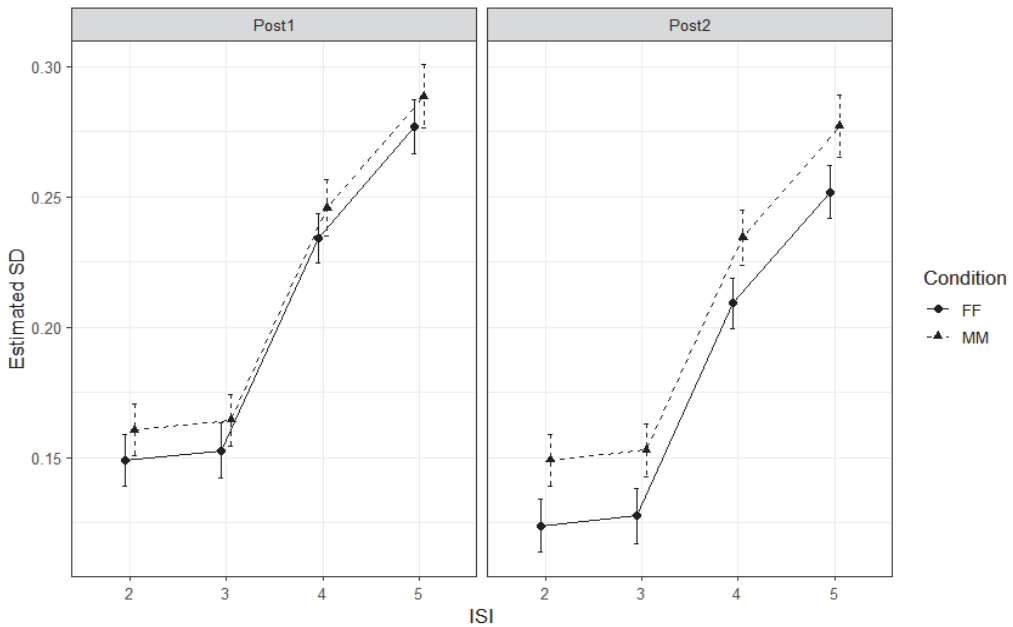
The two-factor model for inter-stress interval variability (ISI-SD) yielded $R^2_m = .536$ and $R^2_c = .648$. The fixed effects reported in Table 2 reveal that both Pre_ISI_SD and ISI type significantly predicted posttest ISI-SD values ($p < .001$), indicating that learners with more variable initial timing and more complex ISI conditions tended to show higher posttest variability. A small but significant reduction in ISI-SD values was observed in Post2 compared to Post1 ($p = .001$), suggesting that repeated shadowing practice led to improved rhythmic stability over time. No significant main effects of Condition or Condition × Period interaction were found.

Table 2 Fixed Effects for ISI-SD (Two-Factor LMM)

Predictor	Estimate	SE	df	t	p
(Intercept)	0.102	0.011	223	9.50	< .001
Pre_ISI_SD	0.200	0.040	413	4.99	< .001
Condition (MM)	0.012	0.012	108	0.98	.329
Period (Post2)	-0.025	0.008	390	-3.23	.001
ISI 3	0.004	0.008	395	0.48	.631
ISI 4	0.085	0.010	452	8.58	< .001
ISI 5	0.128	0.011	458	11.15	< .001
Condition × Period	0.014	0.011	390	1.20	.233

Post hoc tests showed that at ISI 2–5 in Post2, the FF condition had significantly lower ISI-SD values than MM ($p = .035$), suggesting that visual cues may help stabilize rhythm with repeated exposure. Figure 3 illustrates this trend, showing ISI-SD values by condition (FF vs. MM) across the two posttest periods. These findings suggest that cumulative exposure to visible articulatory gestures facilitates the stabilization of speech rhythm. Accordingly, the response to Research Question 1 may be regarded as partially affirmative: repeated shadowing with facial visual input facilitates more native-like prosodic regularity, as reflected in reduced ISI-SD values.

Figure 3 Estimated ISI-SD at Post1 and Post2 by Condition



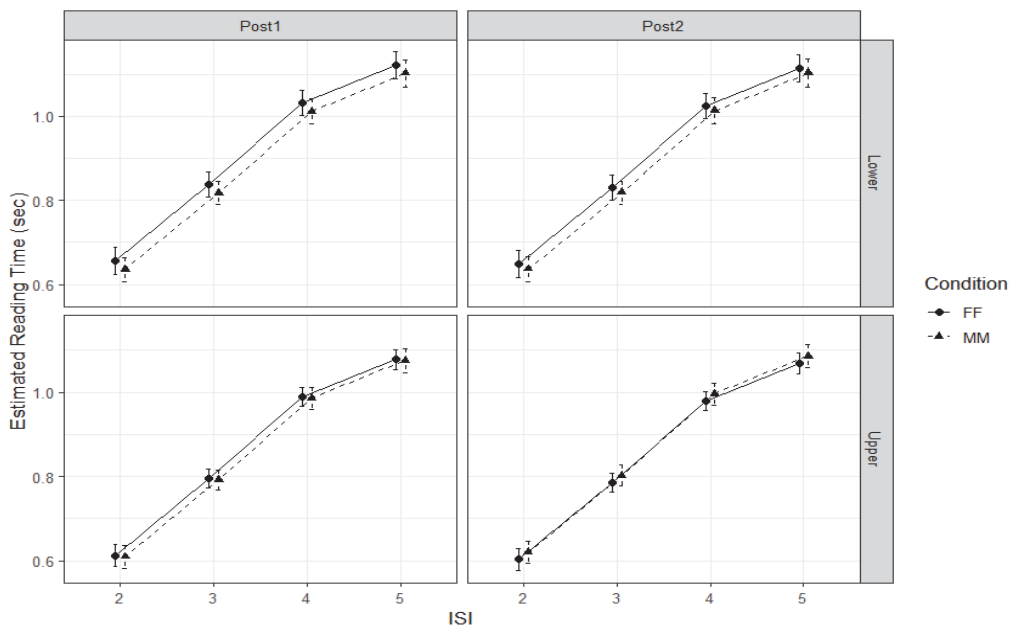
4.4 ISI Duration: Three-Factor Model (Condition × Period × Level)

The three-factor linear mixed-effects model for ISI duration showed $R^2_m = .761$ and $R^2_c = .851$. As presented in Table 3, both Pre_ISI_duration and ISI type remained significant predictors ($p < .001$), confirming the influence of baseline ISI duration and rhythmic complexity on subsequent performance. No significant main effects were found for Condition, Period, or Proficiency Level, nor were any of their interactions significant.

Table 3 Fixed Effects for ISI Duration (Three-Factor LMM)

Predictor	Estimate	SE	Df	t	p
(Intercept)	0.508	0.035	161	14.69	< .001
Pre_ISI_duration	0.165	0.037	465	4.45	< .001
Condition (MM)	-0.020	0.038	88	-0.51	.611
Period (Post2)	-0.007	0.022	400	-0.32	.749
Level (Upper)	-0.043	0.034	85	-1.26	.212
ISI 3	0.182	0.015	438	11.78	< .001
ISI 4	0.376	0.027	466	14.01	< .001
ISI 5	0.466	0.030	467	15.45	< .001
Condition × Period	0.008	0.031	401	0.26	.793
Condition × Level	0.017	0.049	85	0.35	.730
Period × Level	-0.002	0.027	400	-0.09	.929
Condition × Period × Level	0.012	0.039	401	0.30	.761

Figure 4 Estimated ISI Duration by Condition, Period, and Proficiency Level



Post hoc comparisons revealed no significant differences in ISI duration between the FF and MM conditions across any combination of Period, Proficiency Level, and ISI type. Figure 4 demonstrates a consistent pattern of in ISI durations across both test periods and proficiency levels, with no substantial differences observed between the FF and MM conditions.

4.5 ISI-SD: Three-Factor Model (Condition \times Period \times Level)

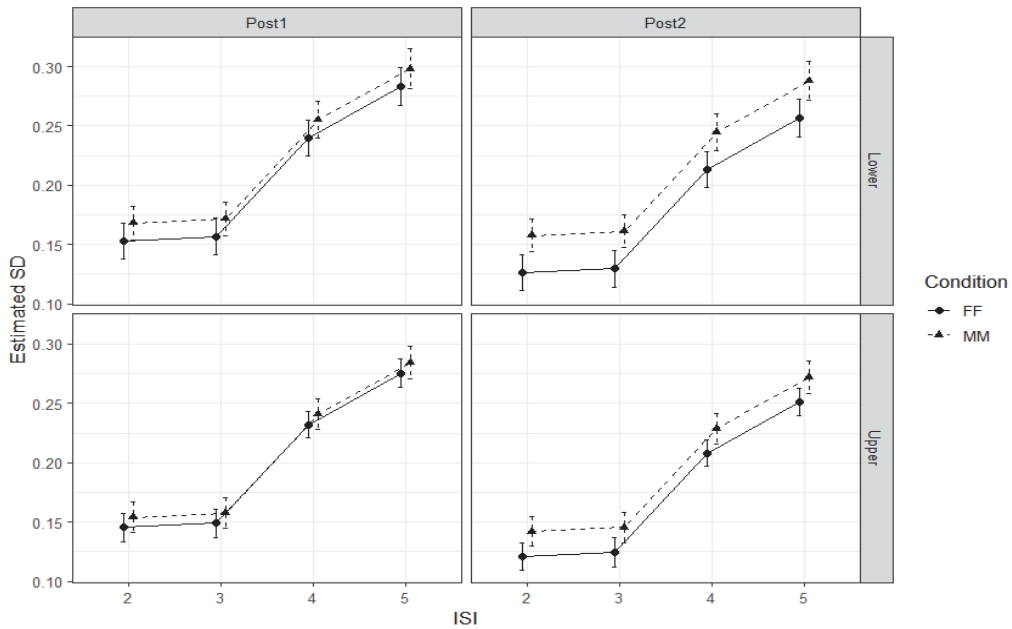
The three-factor linear mixed-effects model for ISI-SD yielded $R^2_{_m} = .534$ and $R^2_{_c} = .649$. As shown in Table 4, both Pre_ISI_SD and ISI type were significant predictors ($p < .001$), indicating that rhythm variability across posttest sessions was influenced by baseline timing stability and prosodic complexity. No other fixed effects or interactions reached significance. Post hoc comparisons revealed no significant differences between the FF and MM conditions; however, a consistent trend toward lower ISI-SD values was observed in the FF condition during Post2 (see Figure 5).

These results provide limited support for Research Question 2. No statistically significant effects or interactions involving proficiency level were observed for either ISI duration or ISI-SD, indicating that learner proficiency did not systematically affect the use of visual prosodic information during training. However, the trend illustrated in Figure 5 suggests that lower-proficiency learners may benefit more from visual prosodic input, as reflected in lower ISI-SD values observed in the FF group.

Table 4 Fixed Effects for ISI-SD (Three-Factor LMM)

Predictor	Estimate	SE	df	t	p
(Intercept)	0.108	0.016	152	6.69	< .001
Pre_ISI_SD	0.191	0.041	428	4.65	< .001
Condition (MM)	0.015	0.020	109	0.78	.437
Period (Post2)	-0.027	0.014	387	-1.96	.051
Level (Upper)	-0.008	0.017	101	-0.45	.657
ISI 3	0.004	0.008	394	0.45	.653
ISI 4	0.087	0.010	445	8.64	< .001
ISI 5	0.130	0.012	452	11.16	< .001
Condition \times Period	0.016	0.019	389	0.86	.391
Condition \times Level	-0.007	0.024	103	-0.27	.790
Period \times Level	0.002	0.017	388	0.15	.883
Condition \times Period \times Level	-0.004	0.024	389	-0.17	.862

Figure 5 Estimated ISI-SD by Condition, Period, and Proficiency Level



5. Discussion

This study investigated the effect of visual prosodic cues—specifically, facial video showing prominent jaw opening—on the acquisition of English speech rhythm in L2 learners during shadowing training. A key finding was that in the two-factor model (Condition \times Period), the FF condition (with visible mouth movements) yielded significantly reduced rhythmic variability (as measured by ISI-SD) compared to the MM condition (with the mouth region masked) at Post2, while no significant difference was observed at Post1. These results suggest that visual articulatory cues—particularly jaw and lip movements—facilitate the acquisition of speech rhythm with sufficient exposure. This finding addresses Research Question 1, showing that shadowing with visible facial movements improves rhythmic control, especially as reflected in ISI-SD.

This interpretation is supported by the findings of Erickson et al. (2012), who demonstrated that jaw movement in native English speakers synchronizes with metrical structure: greater jaw opening was observed during stressed syllables, and jaw oscillation patterns aligned with prosodic groupings. These articulatory features may offer visible cues that help learners perceive and reproduce rhythmic timing more accurately, potentially contributing to more stable inter-stress intervals and reduced ISI-SD in L2 speech production. Furthermore, Park et al. (2016) showed that visual speech—particularly lip movements—can synchronize with the listener’s low-frequency brain rhythms, helping to improve speech intelligibility. This is consistent with Pelle and Sommers (2015), who argued that visual cues convey

temporal information, such as the amplitude envelope, which enhances attention and sensitivity to incoming speech. Together, these findings suggest that rhythmic visual input contributes to accurate speech timing, especially through repeated audiovisual exposure that promotes multimodal integration.

This study also investigated whether learners' proficiency level influenced how visual facial cues were utilized during training (Research Question 2). A three-way linear mixed-effects analysis (Condition \times Period \times Proficiency) revealed no significant interactions for either ISI duration or ISI-SD, suggesting that proficiency did not significantly affect the impact of visual cues. One possible explanation for the lack of a proficiency effect lies in the design of the stimuli. The target utterances featured stressed syllables with the low vowel /a/, which is characterized by wide jaw opening. This articulatory feature likely enhanced the visibility and perceptual salience of prosodic cues for all participants, thereby reducing the differential impact of visual information across proficiency groups in the FF condition.

Nevertheless, Figure 5 indicates a trend wherein lower-proficiency learners benefited more from the FF condition, as evidenced by consistently reduced ISI-SD values relative to the MM condition. This tendency supports the view that less proficient learners depend more on visual prosodic information, particularly when auditory input alone is insufficient to identify stress patterns or prosodic boundaries (Sueyoshi & Hardison, 2005). For these learners, dynamic facial gestures such as jaw lowering may serve as reliable temporal cues for speech segmentation and prosodic interpretation. In this sense, visual information operates not merely as redundant input but as a compensatory modality that supports perceptual processing when auditory input is insufficiently informative.

These findings are consistent with the complementary visual cues hypothesis within the broader audiovisual integration framework (Jesse & Massaro, 2010; Munhall et al., 2004), which holds that visual articulatory signals can supplement or enhance auditory speech perception, especially under conditions of perceptual ambiguity or degraded input. From this perspective, learners with incomplete L2 prosodic representations may rely more heavily on external multimodal input to construct rhythmic patterns.

6. Conclusion and Implication

This study examined whether visual access to a speaker's facial articulation—specifically, jaw movement—during shadowing practice contributes to the stabilization of English speech rhythm in Japanese EFL learners. The results showed that learners in the FF condition (facial video with clearly visible jaw motion) exhibited significantly lower rhythmic variability (ISI-SD) than those in the MM condition (mouth-masked video) by the second posttest. This finding suggests that visible articulatory movement can assist learners in managing prosodic timing more

consistently over repeated training sessions.

The findings have important pedagogical implications for English language instruction targeting Japanese EFL learners. Audiovisual shadowing practice using facial videos—particularly stimuli that contain large mouth-opening vowels /a/ in stressed syllables—demonstrates measurable benefits for English rhythm acquisition. Even low-proficiency Japanese EFL learners, whose native mora-timed rhythm differs from English stress-timed patterns, can benefit from materials incorporating visible articulatory movements. These visual cues appear to facilitate timing control by providing learners with observable jaw movements that mark prosodic prominence. This result is consistent with Wilson et al. (2020), who introduced a "jaw dancing" technique that explicitly teaches English stress-timing through jaw displacement patterns for intermediate-level Japanese EFL learners.

However, several limitations should be acknowledged. First, this study constituted a short-term intervention that assessed immediate outcomes; the durability of the observed gains remains unknown and should be tested with delayed follow-ups (e.g., at 1 month and 3 months post-training). Second, the same sentence stimuli were used across all test phases to maintain control over ISI type; thus, it remains to be determined whether rhythm gains transfer to novel texts or spontaneous speech production. Third, the current analysis focused solely on ISI and ISI-SD. To capture rhythmic variation more comprehensively, future research should incorporate additional metrics such as the normalized pairwise variability index for vowels (nPVI-V) and VarcoV. These measures reflect durational variability while controlling for speech rate and have been shown to effectively distinguish between stress-timed and syllable-timed rhythm patterns in L2 speech (Grabe & Low, 2002; Liu & Takeda, 2021; White & Mattys, 2007).

Nevertheless, this study serves as an initial empirical investigation into the role of visual prosodic cues—particularly jaw movement—during shadowing in facilitating the acquisition of English speech rhythm. It offers valuable insights into the cognitive mechanisms underlying multimodal prosody processing and provides pedagogically implications for second language instruction.

Acknowledgment

This study was partially supported by JSPS Grant-in-Aid for Scientific Research (C), (PI: Hiroshi Nakanishi, No.23K00786). We are deeply grateful to Professor Cynthia Daugherty for her generous assistance with the English-language audiovisual recordings. We also appreciate the helpful suggestions provided by the three anonymous reviewers.

References

- Erickson, D., Suemitsu, A., Shibuya, Y., & Tiede M (2012). Metrical structure and production of English rhythm. *Phonetica*, 69(3), 180-190.
<https://doi.org/10.1159/000342417>
- Foote, J. A., & McDonough, K. (2017). Using shadowing with mobile technology to improve L2 pronunciation. *Journal of Second Language Pronunciation*, 3(1), 34-56.
<https://doi.org/10.1075/jslp.3.1.02foo>
- Grabe, E., & Low, E. L. (2002). Durational variability in speech and the rhythm class hypothesis. *Laboratory Phonology*, 7, 515-546.
<https://doi.org/10.1515/9783110197105.2.515>
- Hamada, Y. (2018a). Shadowing: What is it? How to use it. Where will it go? *RELC Journal*, 50(3), 386-393. <https://doi.org/10.1177/0033688218771380>
- Hamada, Y. (2018b). Shadowing for pronunciation development: Haptic-shadowing and IPA-shadowing. *Asia TEFL Journal*, 15(1), 167-183.
<https://doi.org/10.18823/asiatefl.2018.15.1.11.167>
- Jeong, H., Kazai, K., Kajiura, M., Nakano, Y., & Kadota, S. (2023, March). *The effect of speaker's face on brain mechanisms during second language shadowing: An fMRI study* [Conference presentation]. AAAL 2023 Conference, Portland, OR.
- Jesse, A., & Massaro, D. W. (2010). The temporal distribution of information in audiovisual spoken-word identification. *Attention, Perception, & Psychophysics*, 72(1), 209-225. <https://doi.org/10.3758/APP.72.1.209>
- Konishi, T., & Kondo, M. (2019, September). *A study on rhythm control in Japanese EFL learners using inter-stress interval duration* [Conference presentation]. The 33rd General Meeting of the Phonetic Society of Japan, Seisen University, Tokyo, Japan.
- Liu, S., & Takeda, K. (2021). Mora-timed, stress-timed, and syllable-timed rhythm classes: Clues in English speech production by bilingual speakers. *Acta Linguistica Academica*, 68(3), 350-369. <https://doi.org/10.1556/2062.2021.00469>
- Matsushita, R. (2025). The Effect of Shadowing on the Acquisition of English Rhythm by Japanese EFL Learners—Analysis Based on the Duration of ISI—, *Seinan Gakuin University Graduate School Research Papers*, 20, 27-37.
<http://repository.seinan-gu.ac.jp/handle/123456789/2660>
- Mochizuki-Sudo, M., & Kiritani, S. (1991). Production and perception of stress-related durational patterns in Japanese learners of English. *Journal of Phonetics*, 19(2), 231-248. [https://doi.org/10.1016/S0095-4470\(19\)30219-0](https://doi.org/10.1016/S0095-4470(19)30219-0)
- Munhall, K. G., Jones, J. A., Callan, D. E., Kuratate, T., & Vatikiotis-Bateson, E. (2004). Visual prosody and speech intelligibility: Head movement improves auditory speech perception. *Psychological Science*, 15(2), 133-137.
<https://doi.org/10.1111/j.0963-7214.2004.01502010.x>
- Nakanishi, H. (2024, June). *The effects and instruction of shadowing* [Conference

- presentation]. The 30th Annual Conference of the Kansai English Language Education Society (KELES), Ryukoku University, Kyoto, Japan.
- Nakano, Y., Kadota, S., Kawasaki, M., Nakanishi, H., Hase, N., & Shiki, O. (2024). An eye-tracking study on the effects of speaker's face on shadowing performance. *Language Education & Technology, 61*, 29-58. https://doi.org/10.24539/let.61.0_29
- Nicklin, C., & Plonsky, L. (2020). Outliers in L2 research in applied linguistics: A synthesis and data re-analysis. *Annual Review of Applied Linguistics, 40*, 26-55. <https://doi.org/10.1017/S0267190520000057>
- Park, H., Kayser, C., Thut, G., & Gross, J. (2016). Lip movements entrain the observers' low-frequency brain oscillations to facilitate speech intelligibility. *eLife, 5*, e14521. <https://doi.org/10.7554/eLife.14521>
- Peelle, J. E., & Sommers, M. S. (2015). Prediction and constraint in audiovisual speech perception. *Cortex, 68*, 169-181. <https://doi.org/10.1016/j.cortex.2015.03.006>
- Sueyoshi, A., & Hardison, D. M. (2005). The role of gestures and facial cues in second language listening comprehension. *Language Learning, 55*(4), 661-699. <https://doi.org/10.1111/j.0023-8333.2005.00320.x>
- Trofimovich, P., & Baker, W. (2006). Learning second language suprasegmentals: Effect of L2 experience on prosody and fluency characteristics of L2 speech. *Studies in Second Language Acquisition, 28*(1), 1-30. <https://doi.org/10.1017/S0272263106060013>
- White, L., & Mattys, S. L. (2007). Calibrating rhythm: First language and second language studies. *Journal of Phonetics, 35*(4), 501-522. <https://doi.org/10.1016/j.wocn.2007.02.003>
- Wilson, I., Erickson, D., Vance, T., & Moore, J. (2020). Jaw dancing American style: A way to teach English rhythm. In *Proceedings of Speech Prosody 2020* (pp. 556-560). International Speech Communication Association. <https://doi.org/10.21437/SpeechProsody.2020-114>

