Language and Cognition (2024), 16: 4, 1718–1755
doi:10.1017/langcog.2024.24

ARTICLE

Multimodal training on L2 Japanese pitch accent: learning outcomes, neural correlates and subjective assessments

Yukari Hirata[1], Erica Friedman[2], Caroline Kaicher[3] and Spencer D. Kelly[1]

[1]Center for Language and Brain, Colgate University, Hamilton, NY, USA; [2]Department of Speech, Language & Hearing Sciences, Boston University, Boston, MA, USA and [3]Department of Psychology, Stanford University, Stanford, CA, USA
Corresponding author: Yukari Hirata; Email: yhirata@colgate.edu

(Received 16 October 2022; Revised 25 February 2024; Accepted 15 March 2024)

Abstract

Japanese pitch accent is phonemic, making it crucial for second-language learners to acquire. Building on theories of multimodal learning, the present study explored how auditory, visual and gestural training of Japanese pitch accent affected behavioral, neural and meta-cognitive aspects of pitch perception across two experiments. Experiment 1 used a between-subjects pre/posttest design to train native English speakers to perceive Japanese pitch accents in one of the following three conditions: (1) baseline (audio + flat notation), (2) pitch height notation (audio + notation mimicking pitch height) and (3) pitch height notation + a left-hand gesture (L-gesture) (to engage the contralateral right hemisphere specialized for suprasegmental pitch processing). Our results indicated that (2) pitch height notation training was most robust in its benefits, as participants in this condition improved on trained and novel words alike. Experiment 2 used a within-subjects design to extend Experiment 1 in three ways: adding a right-hand gesture (R-gesture) condition (to engage more segmental language areas in the left hemisphere), introducing a neural correlate of cognitive load (measured by EEG alpha and theta power) and performing a metacognitive subjective assessment of learning (e.g., 'Which training did you find the most helpful?'). The results showed that although there were no differences among our four training conditions on learning outcomes or EEG power, participants made the most positive subjective evaluations about pitch height notation and R-gesture training. Together, the results suggest that there may be a 'just right' amount of multimodal instruction to boost learning and increase engagement during foreign language pitch instruction.

Keywords: Japanese pitch accent; multimodal training; second language (L2); hand gesture

1. Introduction

Japanese pitch accent is crucial for second language (L2) learners to master in order to communicate effectively with native speakers. Pitch accent in Japanese is phonemic, meaning that it varies by word and can mark the difference between otherwise identical words. For example, *kami* means 'god' with a high-low (HL) pitch pattern, while it means 'hair' with a low-high (LH) pitch pattern. Pitch accent distinction in Japanese is solely realized in fundamental frequency, while the English language does not make lexical distinction based solely on the fundamental frequency (Beckman, 1986). Thus, it can be very difficult for many native English speakers to acquire (Hirano-Cook, 2011; Muradás-Taylor, 2022; Sakamoto, 2011). The present study investigates whether multimodal training is effective for native English speakers' learning of Japanese pitch accent perception, and if so, to what extent multimodal input is optimal in assisting perception of these phonemic pitch distinctions.

1.1. Multimodality in L2 learning

Despite the traditional emphasis on auditory input in L2 instruction, multimodal input offers many learning benefits (McCafferty & Stam, 2009). Specifically, visual input in the form of waveform displays and visual pitch markers accompanying speech are advantageous for L2 phonetic learning, and gesture input, while proven to be helpful for semantic and pragmatic components of L2 learning, may extend to phonetic aspects of L2 acquisition as well (Allen, 1995; Baills et al., 2019; Church et al., 2017; Hannah et al., 2017; Hirata et al., 2014; Kelly et al., 2017; Liu et al., 2011; Morett et al., 2022; Motohashi-Siago & Hardison, 2009; Pi et al., 2021; Sueyoshi & Hardison, 2005; Tellier, 2008). Below, we review some of the relevant research on the benefits of multimodal instruction on L2 learning.

1.1.1. Speech + visual input

Theories about how the brain understands information provide mechanisms for effective teaching and learning. Dual coding theory (DCT) holds that verbal (linguistic) information and nonverbal information (imagery) are processed in two separate systems (Clark & Paivio, 1991), with stimuli containing words activating verbal representations, and stimuli that contain images activating image-based representations. Presenting information in a way that integrates both verbal and image representations is thought to help learning, as coding stimuli in two different ways can increase the likelihood of remembering it.

Findings from the literature on multimodal learning support DCT by demonstrating that multimodal audio and visual input helps in various aspects of L2 learning (Liu et al., 2011; Motohashi-Siago & Hardison, 2009). For example, Japanese geminate consonants, having slightly longer duration than their singleton (shorter) counterparts, were more accurately identified by native English speakers following multimodal training with audio and visual speech waveform displays showing the segmental duration of the consonant, compared to audio-only training (Motohashi-Siago & Hardison, 2009). This has also been shown with Chinese tones, where training with audio input accompanied by pinyin spelling of the spoken syllables plus visual pitch markers showing the shape of the tones reduces errors in identification compared with other forms of training with less multimodal input (Liu et al., 2011). The authors theorize that the multimodal training of the contour + pinyin condition was most effective for accurate tone perception because the visual modality

https://doi.org/10.1017/langcog.2024.24 Published online by Cambridge University Press

was intentionally designed to support learner attention to tonal information. Thus, high and low tone heights seem to evoke an up-down metaphor, suggesting that stimuli that visually represent tonal contours are an isomorphic analogue to spoken tonal contours (Bolinger, 1983; Liu et al., 2011; Morett et al., 2022). Such stimuli may allow the cognitive system to utilize the natural congruence between the spectral and spatial processing of auditory and visual information in a highly beneficial manner.

However, not all studies have shown uniformly positive effects of visual input on learning to perceive lexical tones. For example, Morett et al. (2022) presented native English speakers with training videos comprised of Mandarin lexical tones coupled with animated dots metaphorically tracing the pitch of the tones. When the dots were incongruent with the tones during training, they decreased performance from pretest to posttest relative to a no motion baseline; however, when the dots were congruent with the tones, they increased performance no better than the baseline training. Interestingly, these congruent and incongruent metaphoric dots produced similar learning outcomes as metaphoric hand gestures.

1.1.2. Speech + gestural action

A more ready-made type of multimodal expression comes in the form of co-speech hand gestures.[1] McNeill (1985) argues that the hand gestures that accompany speech combine to create a tightly coupled semantic system, and there is a large body of research showing that these gestures play a significant role in language production (Hostetter & Alibali, 2008) and comprehension (Dargue et al., 2019; Hostetter, 2011) in a native language. Recently, there has been a theoretical push to explore this gestural benefit at the phonetic level as well (Kelly, 2017).

Indeed, there is good evidence that gesture affects prosodic components of language in L1 speech production and comprehension (Hubbard et al., 2009; Krahmer & Swerts, 2007). For example, Krahmer and Swerts (2007) showed that producing beat gestures with certain words not only changed how those words were produced, but even when the acoustic properties of speech were controlled for, beat gestures affected listeners' perception of the acoustic prominence of those words.

Given their prominent role in L1 speech production and comprehension, one might expect gestures to also have benefits for speakers of an L2. Indeed, it is now well established that hand gestures serve multiple positive functions in the context of L2 production, comprehension and learning (Gullberg, 1998, 2006; Lazaraton, 2004; McCafferty, 2002; McCafferty & Stam, 2009; Sime, 2006; Smotrova & Lantolf, 2013; Yoshioka & Kellerman, 2006). However, with specific regard to gesture's phonetic function in processing and learning in L2, there appears to be mixed results in the literature (Baills et al., 2019; Church et al., 2017; Gluhareva & Prieto, 2017; Hannah et al., 2017; Hirata & Kelly, 2010; Hoetjes & Van Maastricht, 2020; Morett et al., 2022; Morett & Chang, 2015; Smotrova, 2017; Xi et al., 2020; Zhen et al., 2019; Zheng et al., 2018). For instance, Hirata and Kelly (2010) found that auditory training in which participants viewed beat/metaphoric gestures did not help improve the perception of phonemic vowel length in L2 Japanese learners beyond that of an audio-only condition. This is consistent with the study by Morett et al. (2022) showing that training with congruent pitch gestures did not improve novice learners' ability to

[1]Lip movements are another natural form of multimodal input (McGurk & MacDonald, 1976), but in the interest of space, we do not include them here (but for more on this, see Hirata & Kelly, 2010; Hardison & Pennington, 2021).

https://doi.org/10.1017/langcog.2024.24 Published online by Cambridge University Press

perceive Mandarin tones any better than a no gesture baseline. In contrast, Gluhareva and Prieto (2017) found that intermediate L2 learners of English (native Catalan speakers) were judged by native English speakers to have more native accents when trained with beat gestures versus no gesture – however, this pattern held only for hard items, but not easy ones. With regard to pitch perception, Bailly et al. (2019) showed that observing and producing metaphoric pitch gestures helped L2 speakers learn novel Mandarin tonal distinctions and vocabulary items. Together, these results demonstrate how gestures can be beneficial for L2 phonetic learning in some contexts, but not in others (Kelly, 2017).

Perhaps the *production* of gestures may also assist L2 phonological learning in some contexts (Bailly et al., 2019; Zhen et al., 2019). For example, Bailly et al. (2019) directly compared the effect of gesture observation and production on the perception of Chinese lexical tones, finding that both were effective in improving accuracy (for the prosodic benefits of producing hand claps during L2 learning, see Zhang et al., 2020). However, the effects are not always robust. Hirata et al. (2014) conducted a similar comparison of gesture observation and production on L2 perception of Japanese vowel-length contrasts, and also compared syllabic- versus moraic-rhythm gestures. They found that there was similar auditory improvement for all combinations of trainings, but observing syllable gestures had a slight advantage over the other conditions. Thus, further examination of gesture production may be warranted, which we will address in the present study.

Finally, there is evidence that producing gestures may serve to prime different neural networks to facilitate learning. Because the hands are controlled by contralateral hemispheres – which specialize in different aspects of perceptual and cognitive processing (Poeppel, 2003; Zatorre et al., 2002) – it is possible that gesturing with the left and right hands may help learning in different ways. For example, left-hand gestures (L-gestures) would more directly activate a right lateralized network, which is specialized for processing prosodic dimensions of speech, such as rhythm, intonation, tone and pitch (Lattner et al., 2005; Loui et al., 2011; Schlaug et al., 2009; Sidtis, 1980; for music: Peretz & Zatorre, 2005). This might be especially useful for early learners of a pitch/tone-based language for at least three reasons: (1) the right hemisphere arcuate facilicus is a purported mechanism for processing pitch sequences (Loui et al., 2011), (2) previous research has shown that naïve nontonal language speakers process lexical tones primarily in the right hemisphere (Klein et al., 2001) and (3) the left hemisphere becomes specialized for pitch processing only after extensive experience with a tonal/pitch-based language (in infants: Sato et al., 2010; and adults: Wang et al., 2004), making the right hemisphere a possible more viable early target.

In contrast, right hand gestures (R-gestures) would more directly activate a left lateralized network, which is specialized for fine-grained processing of smaller units, such as syllables and phonemes (Blumstein et al., 1977; Burton et al., 1998; Caplan et al., 1995; Fiez et al., 1995). Given that phonemic pitch/tonal processing is lateralized to left-hemisphere networks in native speakers (Japanese: Sato et al., 2010; Mandarin: Wang et al., 2004; Thai: Van Lancker & Fromkin, 1973), it is possible that directly targeting the left hemisphere network would help novice L2 learners to perceive the tones in a more linguistic way. In other words, by encouraging novices to process pitch patterns in the same left-lateralized way as native speakers, it may be possible to give them a head start in the learning process. Another possible advantage of R-gestures is that they are more easily produced by right-handed individuals, and past research has shown that there is a positive association with

https://doi.org/10.1017/langcog.2024.24 Published online by Cambridge University Press

using one's dominant hand to perform manual actions and gestures (Casasanto, 2009, 2011). Because no study (to our knowledge) has explored how pitch training with L- and R-gestures may differentially facilitate the early stages of L2 phonemic perception, the present study aimed to be a first step in exploring these two different mechanisms.

1.2. The present study

While many previous studies have investigated the effect of gesture perception on various aspects of L2 learning, there is relatively little research exploring gesture production and its effect on L2 *phonetic* learning (Baills et al., 2019; Hirata et al., 2014; Zhen et al., 2019), and gesture is rarely studied in tandem with other visual–spatial representations of phonology in an L2 context (but for gesture *perception*, see Morett et al., 2022). Moreover, the neural mechanisms and subjective impressions of multimodal instruction have also been overlooked, and no study (to our knowledge) has compared L-gesture and R-gesture in L2 pitch training. The current preregistered experiments address these gaps in the literature by determining if multimodal learning with various forms of visual–spatial representation of pitch accent (through spatial notation and hand gestures) improves Japanese pitch accent perception, which occurs at the phonemic level.

The Japanese pitch accent shares some similarities with English lexical stress, but there are some important differences. In both languages, one part of the word is produced and perceived more prominently than other parts of the word, and the location of the prominent part is lexically determined. In Japanese, the pitch accent is located where high (H) is followed by low (L). For example, in Appendix 1, the pitch accent is on the first mora of the four-mora words in Type 1, and it is on the second mora in Type 2 and so on. However, we note a major difference between the two languages in ways that the prominence is realized. In English, lexical stress is realized in multiple ways, such as by the prominent syllable being longer in duration, higher in the fundamental frequency, higher in intensity and/or by the vowel quality changes (e.g., a different quality of the first vowel in *to recórd* as a verb versus in *a récord* as a noun) (Beckman, 1986; Sluijter & van Heuven, 1996a, 1996b). In contrast, the realization of pitch accent in Japanese is only by the use of pitch height, which is a perceived or produced height of fundamental frequency, and other properties of the word such as syllable duration or vowel quality remain relatively the same (Vance, 2008). For example, in *kámi* 'god' with a HL pitch pattern versus *kamí* 'hair' with a LH pitch pattern, the duration and quality of the first vowel /a/ do not differ drastically regardless of the accent presence.[2]

Many studies have shown that L2 perception and production of Japanese pitch accent is a challenge for native English learners of Japanese (Goss, 2020; Hirata, 2015; Muradás-Taylor, 2022). Pedagogically speaking, a practical challenge for learners is that many Japanese language textbooks (e.g., Banno et al., 2020) do not mark lexical

[2]Another major difference between the two languages is that Japanese has words without pitch accent, which means that there is no pitch *fall* from high to low within the word (as shown in Type 0, i.e., the LHHH pitch type, as in Appendix 1). It must be noted that in Japanese phonology, the change from L to H is not processed as a pitch accent. This pattern of 'no pitch fall' contrasts with English in which no word can get by without a lexical stress when pronounced in isolation (see Vance (1987, 2008) for more details of Japanese phonetics and phonology).

https://doi.org/10.1017/langcog.2024.24 Published online by Cambridge University Press

pitch accent in their vocabulary lists, and that the acquisition of pitch accent is typically left up to individual instructors and learners. Even with some textbooks that do mark pitch accent (e.g., Jorden & Noda, 1987; Noto, 1992), there is little scientific research investigating what type of pitch accent notations are helpful or effective for learners. This motivated our comparison between the first two conditions using different notations, as described below.

By studying multiple forms of visual–spatial pitch in a phonetic learning task, we aim to investigate the possibility that multimodality is important to varying extents for different levels of language. Experiment 1 uses a between-subjects and pre/-posttest design to explore the efficacy of combining layers of multimodal input to create three training conditions: (1) baseline (audio + flat notation), (2) notation (audio + notation spatially mimicking pitch height) and (3) notation used in (2) + L-gestures. Training (2) has visual information that is more directly relatable to the pitch accent patterns than training (1), and training (3) has an additional modality of hand gesture production while also having the same visual information on pitch accent patterns as training (2). Experiment 2 expands on Experiment 1 by adding a R-gesture training and also extending our dependent measures for learning outcomes. Specifically, it measures not only pitch identification accuracy but also neural activity (EEG) and subjective assessments following the various levels of multimodal training.

If multimodal training assists L2 phonetic learning, then more multimodal information during training will boost Japanese pitch accent learning (Hardison & Pennington, 2021). However, it is possible that multimodal training boosts L2 phonetic learning only when there is the right amount of multimodal input, with too much visual information perceptually distracting learners and decreasing effectiveness (Kelly, 2017). The results of this study will help elucidate which of these possibilities is the case and will clarify the mechanisms behind the benefits of multimodality in L2 phonetic training.

2. Experiment 1

As described in the previous section, Experiment 1 compares the three types of training with identical audio materials: (1) a flat notation baseline displaying pitch patterns of H and L with text, (2) a notation displaying pitch patterns in a corresponding visual–spatial arrangement and (3) the notation used in (2) + L-gesture production in which participants traced the pitch contour of the words with their left hands.[3] Based on previous research showing the facilitative effects of visual information corresponding to critical auditory characteristics (Hardison, 2005; Hardison & Pennington, 2021), we predicted that training (2) would result in more improvement than training (1). Based on theories of multimodal processing (Clark & Paivio, 1991) and empirical research showing that left-hand movements boost learning by activating a right hemisphere prosodic network (Loui et al., 2011; Schlaug et al., 2009), we predicted that training (3) would result in the most improvement in pitch accent perception.

[3]We chose not to add a R-gesture training for Experiment 1 because of power constraints; too many subjects were needed for our between-subjects design if the study were to have four training conditions.

https://doi.org/10.1017/langcog.2024.24 Published online by Cambridge University Press

2.1. Methods

2.1.1. Participants

This study included 66 participants as determined by a power analysis (power = .95, effect size f = .25, alpha = .05). All were right-handed, monolingual English speakers (51 females and 15 males) who were between the ages of 17 and 22. None had any formal exposure to the Japanese language. To ensure everyone was a monolingual English speaker, we gave a short survey about language background and excluded those who were exposed to any language other than English in their household growing up. Participants were recruited via posters around Colgate University's campus and social media posts, and all participants were compensated with $30 for their participation.

Participants were divided into the following three groups: (1) BASELINE (flat notation) training, (2) NOTATION (spatially representing pitch height) training and (3) L-GESTURE training. All participants completed a pretest, underwent their respective training type (1, 2 or 3), and then completed a posttest to assess the effect of that particular training condition on Japanese pitch accent perception.

2.1.2 Materials

2.1.2.1. Pretest and posttest stimuli. Both the pretest and posttest consisted of 36 target words within carrier sentences. Each target word had four morae (which roughly correspond to syllables) and one of four pitch patterns: HLLL (e.g., *mominoki* 'fir tree'), LHLL (e.g., *kudamono* 'fruit'), LHHL (e.g., *tamanegi* 'onion') or LHHH (e.g., *niwatori* 'chicken'). We chose these words as opposed to shorter words that have fewer pitch pattern alternatives, for example, HL or LH. This was because four mora words are common in Japanese vocabulary and the chance level of correct responses would be 25%, leaving plenty of room to see participants' improvement. It also avoids any possible ceiling effects for some individuals (see large individual variations of L2 learners' abilities in Muradás-Taylor, 2022).

Each word was spoken by two native Japanese speakers from the Tokyo Metropolitan areas, in their 50s, one male and one female.[4] Thus, there were a total of 72 trials. The audio stimuli were presented concurrently with visual stimuli of the written sentence using PowerPoint. The 36 words were of two types: 16 were words that would be trained and 20 were words that would not be included in the training (Appendix 1). A combination of trained and novel words was included to investigate the generalizability of each training type. The visual stimuli consisted of the sentence written out on the screen with a space between each mora and a blue box around the target word (shown in Figure 1). The four answer options were displayed below the sentence, and the question number was displayed in the top left corner. This word-in-a-sentence format was chosen for our testing and training because it has more facilitative and generalizable effects than a word-in-isolation format (Hirata, 2004a, 2004b). While the 72 target words were the same for the pretest and posttest, the carrier sentences varied. For the pretest, the carrier sentences *mazu* ___ *ja nai* 'First of all, it is not ___.' and *soko de* ___ *ga mieta* 'At that point, you could see ___.' were used, while for the posttest, the carrier sentences *kore wa* ___ *desu yo* 'This is ___.' and *sorede* ___ *datta* 'So, it was ___.' were used. Audio stimuli were edited in

[4]Speaker variability is important in enhancing the learning of nonnative speech sounds, especially in generalizing to novel stimuli (Zheng et al., 2018).

https://doi.org/10.1017/langcog.2024.24 Published online by Cambridge University Press

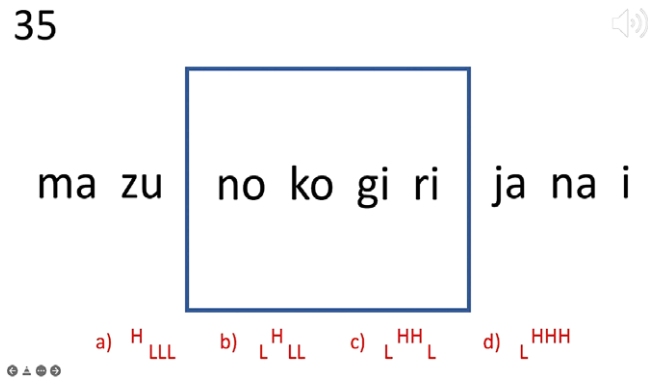


Figure 1. Pretest stimuli. This visual slide (stimulus number 35) is an example of the pretest stimuli presented along with the audio of the whole sentence '*mazu nokogiri janai*'. The box shows the target word. The four pitch patterns (a)–(d) written in red at the bottom are the response alternatives for participants to choose from for the target word they had heard. The Ls and Hs represent lows and highs of pitch accent, respectively. Note that the Ls and Hs were used in the baseline flat notation training, and the spatial arrangement of those Ls and Hs captures the visual–spatial representation used in notation training and L-gesture training.

Praat so that each sentence was played twice with one second of silence at the beginning of the sentence and 2.5 seconds of silence between repetitions.

2.1.2.2. Training stimuli. Audio clips for the training session were recorded by the same male and female native Japanese speakers as the pretest and posttest. Twenty words in total were used for training, allowing for five of each pitch pattern to be trained. Of these 20 trained words, 16 of them were used in testing (see Appendix 1). Four of the trained words were excluded from testing in order to have an equal number of vowel-beginning words in each pitch type category. Each was presented to the participant in four different carrier sentences, for a total of 80 trials. The carrier sentences used for training were *ima* ___ *ga suki desu* 'I like ___ now.', *sore wa* ___ *de wa nai desu* 'It is not ___.', *are wa* ___ *desu* 'The one over there is ___.' and *koko wa* ___ *da to omoimasu* 'I think that this is ___.' The audio was trimmed with one second of silence at the beginning and 2.5 seconds of silence in between repetitions, and each sentence was repeated three times. Please see the next section on the procedure for how each audio repetition came with the presentation of varying visual stimuli.

There were three training conditions (shown in Figure 2). Group A that received the baseline flat notation training first saw the sentences written horizontally on the screen along with the auditory stimuli, as shown in Figure 2 (1). On the second and third repetitions, the pitch pattern was revealed to them through Hs and Ls written below each mora of the target word, as shown in Figure 2 (2a). Participants were told that H and L referred to high and low pitch, respectively. Group B that received notation training first saw the sentence written horizontally for the first repetition, as shown in Figure 2 (1). Then, for the second and third repetitions, the morae of the target word shifted to create a visual–spatial representation of the pitch pattern, such that their vertical position indicated whether they had a high or low pitch (Figure 2 (2b)). Group C that received the L-gesture training saw the same visual stimuli as the

https://doi.org/10.1017/langcog.2024.24 Published online by Cambridge University Press

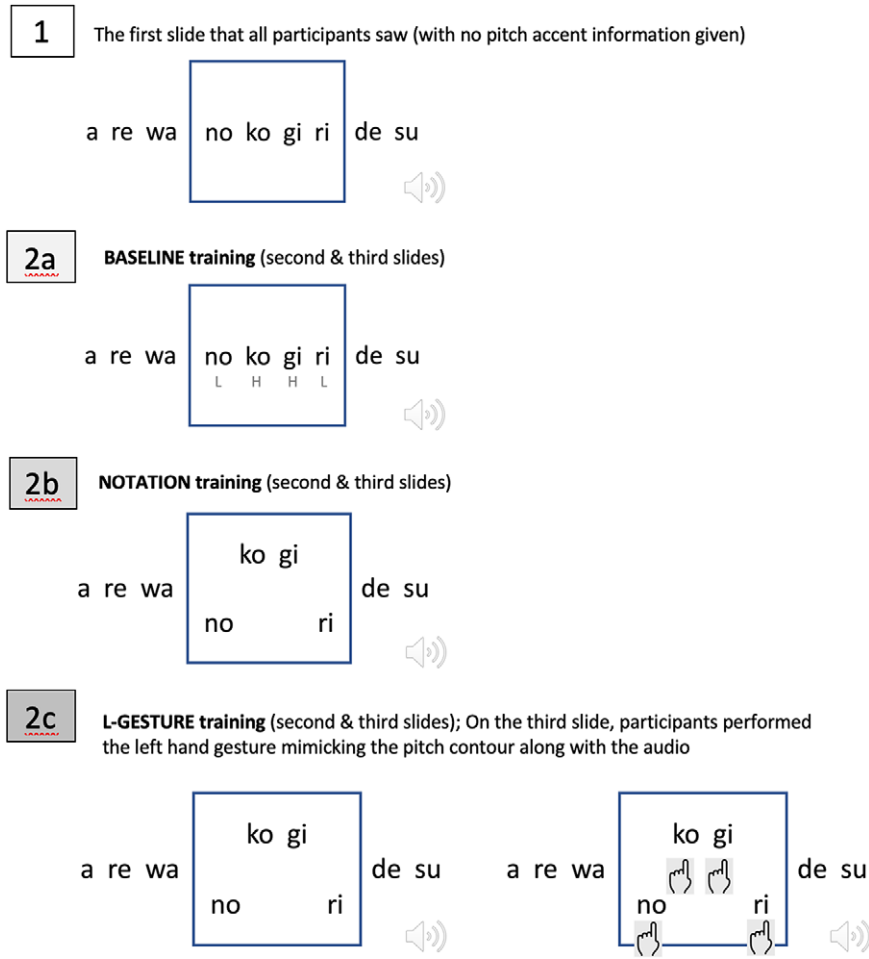


Figure 2. Training stimuli for each condition. (1) First slide for all participants, (2a) baseline flat notation training, (2b) notation training and (2c) L-gesture training. (The hand images were not displayed to the participant – they are used here to demonstrate the contour of the gesture produced by participants.)

notation training, but also produced L-gestures tracing the corresponding highs and lows of the pitch patterns (Figure 2 (2c)).

2.1.3. Procedure

Two pilot subjects were run before any data were formally collected to solidify the procedure. All participants attended a total of three sessions on three separate days, and a between-subjects pretest/posttest design was used to assess the efficacy of the three different training conditions in improving Japanese pitch accent perception. Because of the constraints of the Covid-19 pandemic, this experiment was conducted entirely over the virtual video meeting platform Zoom. In all training and testing sessions, participants were tested individually.

https://doi.org/10.1017/langcog.2024.24 Published online by Cambridge University Press

2.1.3.1. Day 1: Introduction and pretest. Participant consent was obtained by reviewing and signing a consent form that was sent to the participant. Participants took the pretest during the first session following a brief introduction to Japanese pitch accent. The introduction explained what Japanese pitch accent is by contrasting it with stress accent in English, and included examples spoken by a female native Japanese speaker for each of the four types of pitch patterns used in this study (HLLL, LHLL, LHHL and LHHH). The introduction also showed an example of minimal accent pairs using the Japanese word *hari*. Audio clips of *hari* spoken by a female native Japanese speaker with a HL and LH accent pattern were played to demonstrate to participants how differences in pitch accent can change the meaning of otherwise identical words, as the former means 'needle' and the latter means 'supple-surface'.

After the introduction, which took about 20 minutes, participants were instructed on their task for the pretest — to listen to audio carefully and to determine which pitch pattern out of the four pitch pattern options was correct for the target word, which was outlined in a blue box (as shown in Figure 1). Four example questions were shown to participants at the end of the introduction to make sure they understood the format of the pretest and their task.

The pretest consisted of 72 words in carrier sentences. Participants were instructed to take out a blank piece of paper and number it 1–72, leaving enough space to write the letter – either A, B, C or D – that corresponded to the pitch pattern they believed to be correct for the target word. Each slide was shown for a duration of ~13 seconds. The first eight seconds consisted of showing the sentence twice (a little over 2 seconds each) with 2.5 seconds pause between them, and this was followed by five seconds of silence for the participant to write down their answer. Halfway through (after question 36) a break of 2–5 minutes was mandated for all participants. Once the participant was ready (after a maximum of 5 minutes), the second half of the pretest was administered. When the pretest was complete, participants emailed a photo of their answer sheet to the experimenters. In total, the first session took about 45 minutes to complete.

2.1.3.2. Day 2: Training. The second session took place 1–3 days after the first session. During this session, participants underwent training that differed depending on the condition they were assigned, and they were asked to learn to identify Japanese pitch accent patterns as much as they could. In all training conditions, participants listened to Japanese words that were always embedded in carrier sentences and saw the sentence written across their screen, as shown in Figure 2 (1). For all training conditions, the first auditory presentation was accompanied by the identical visual slide that did not reveal the pitch accent of the target word (Figure 2 (1)). All participants were instructed to listen carefully and try to identify the correct target pitch pattern. On the second and third time the sentence was played, the target word's pitch pattern was shown to the participant in different ways depending on the training condition. Baseline training displayed the correct pitch pattern using Ls and Hs underneath each mora in the target word, as shown in Figure 2 (2a). Notation training used a visual–spatial notation in which the target word's pitch pattern was represented by the vertical position of the morae on the computer screen, as shown in Figure 2 (2b). On the second and third slides with the accompanying audio, the baseline and the notation groups were instructed to listen and make sure that the displayed answer made sense with their auditory impression. L-gesture training

https://doi.org/10.1017/langcog.2024.24 Published online by Cambridge University Press

employed the same notation visuals, as shown in Figure 2 (2b), but this group was instructed to listen and understand the pitch pattern indicated by the second slide, and additionally, on the third slide (which is the same as the second one), trace the target word's pitch contour in the air with their left hand as they heard the speaker say the target word (Figure 2 (2c)).

Before training started, a brief introduction was provided on the specific training conditions to which they were assigned. The introduction for the L-gesture condition included the experimenter demonstrating the correct gesture for an example sentence. Before modeling the gesture, the experimenter explained that the hand should trace the pattern that the heights of the morae make on the screen. Then, the experimenter produced the four-part gesture sequence with the left hand (e.g., in the same pattern as illustrated in Figure 2 (2c)), making sure to take up the whole zoom screen. Zoom's mirroring function was used so that the gestures would be displayed to the participant in the correct direction. Following the gesture, the experimenter emphasized how the hand went up/down along with each mora in the word. For the next example, the experimenter and participant both did the gesture to practice the movement and ensure the participant understood the task. Throughout the training, the experimenter watched the participants' L-gestures through the Zoom screen to make sure participants were producing the gesture at the right time, using the correct hand, and that the shape of the gesture was large enough and followed the high and low spatial arrangement of the morae presented in the notation visual stimuli. The experimenter offered suggestions to correct the participant gesture production when necessary.

During all three training conditions, participants heard 80 trials of a total of 20 Japanese words. Thus, each word was displayed to participants a total of four times, each time in a different carrier sentence. After every 5 words, participants had the option of taking a short break, and after every 20 words, a 2–5 minute break was mandated. In total, the second session took about 50 minutes.

2.1.3.3. Day 3: Posttest. In the third session, which occurred 1–3 days after the second session, the posttest was administered. Prior to the posttest, participants were reminded that this study is a learning experiment and to try their best. The posttest followed the same format as the pretest, with the same 36 words tested in the pretest (16 trained words and 20 untrained words) spoken by either the male or female native Japanese speaker and presented in a randomized order. Each word was spoken twice in a different carrier sentence for a total of 72 trials.

A 2–5 minute break was mandated halfway through, during which the experimenter checked in with the participant. After the posttest was completed, participants emailed a photo of their answer sheet to the experimenter, and the participant was debriefed and paid. In total, the third session took about 30 minutes.

2.1.4. Design and analysis

This experiment had a 2 (pre/posttest) × 3 (training condition) × 2 (trained/novel) mixed design. Pre/posttest and trained/novel were within-subjects variables, and training condition was a between-subjects variable. The dependent variable was accuracy of auditory identification on all of the testing items. This experiment was preregistered through Open Science Framework (https://osf.io/rbkuh). Our sampling plan, methods and analyses follow what was reported there, with the exception

https://doi.org/10.1017/langcog.2024.24 Published online by Cambridge University Press

of using a linear mixed effects (LME) model instead of an ANOVA, which was requested by a reviewer.

2.2. Results

Identification accuracy for items on the pretest and posttest was analyzed using mixed effects logistic regression models. The models were fit in R (version 4.3.2), implemented in RStudio, using the glmer() function of the lme4 package (Bates et al., 2015), and null hypothesis significance testing was conducted using the lmerTest package (Kuznetsova et al., 2015). *T* tests were conducted using the emmeans package in R (Lenth et al., 2019).

A model including fixed effects of group (baseline/notation/L-gesture), test (pretest/posttest) and item type (trained/novel) with random intercepts of participant and word displayed a significant effect of test ($\beta = .52$, SE = .12, $p < .001$), with participants improving from pretest to posttest. No significant effects of training group or item type on test performance were observed. Additionally, a significant three-way interaction was observed for the L-gesture group versus the baseline group ($\beta = .47$, SE = .22, $p < .05$) (Appendix 3.1). As outlined in our preregistration report, we followed up on this significant three-way interaction by conducting *a priori* t tests on all conditions from pretest to posttest. These t tests were on the estimated marginal means of the LME model. These analyses revealed that the improvement from pretest to posttest on trained and novel items depended on training condition. As shown in Figure 3, those who received baseline training, z ratio = 4.44, $p < 0.001$, and notation training, z ratio = 3.94, $p < 0.001$, displayed significant improvement on the trained items from pretest to posttest. In contrast, for the novel (untrained) items, the notation, z ratio = 3.82, $p < 0.001$ and L-gesture, z ratio = 3.42, $p < 0.005$, groups showed significant improvement from pretest to posttest. Thus, those who were trained using notation significantly improved on both trained and novel items from pretest to posttest, while those who received the flat notation baseline training significantly improved only on trained items, and those who completed L-gesture training significantly improved only on novel items.

2.3. Discussion

The notation training was beneficial for learning both trained and novel words, while the flat notation baseline was beneficial only for trained items, and the L-gesture training was beneficial only for novel items. Therefore, the notation training was most robust in its benefits, whereas the benefits of the other training groups appear to be more specific. This is partial support for our preregistered prediction: While notation training produced wider learning outcomes than our baseline training, the L-gesture condition was less robustly effective than notation alone.

Since the benefits of the flat notation training did not extend to novel words, perhaps participants shallowly encoded the pitch patterns of the trained words, thus relying on memorization as they completed the posttest. Meanwhile, participants who underwent the L-gesture training may not have been able to focus on the trained items as much as those who completed the flat notation training due to the distraction of producing gestures at the same time as listening. Indeed, producing gestures with

https://doi.org/10.1017/langcog.2024.24 Published online by Cambridge University Press

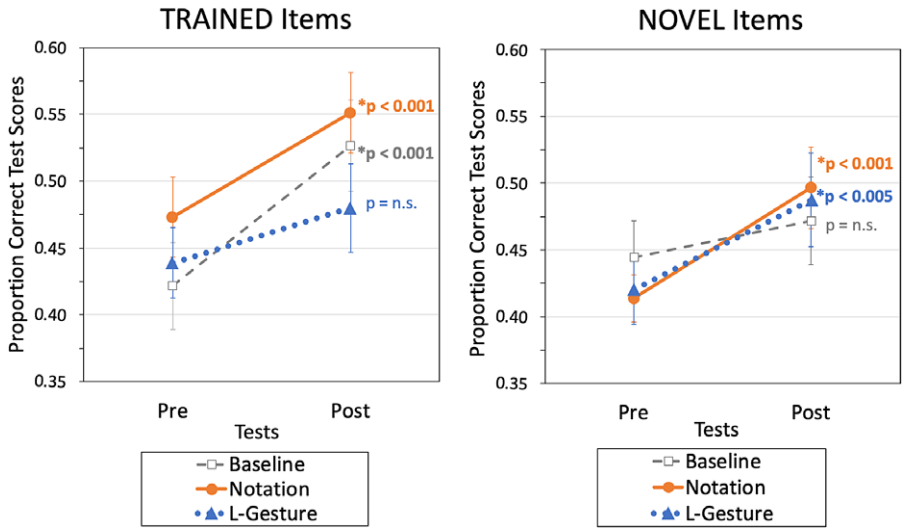


Figure 3. Proportion correct test scores of the three groups in Experiment 1. Only the notation group improved for both trained and untrained items.

one's nondominant hand may be a strain in and of itself, adding the challenge of an already difficult task.

Regardless of why left-hand gestures were distracting, it is interesting that their presence eliminated the positive effects of the spatial notation training. This pattern fits with research showing that gestures can occasionally be detrimental for L2 learning (vocabulary learning: Kelly & Lee, 2012; phonetic perception and production: Hirata & Kelly, 2010; Hoetjes & Van Maastricht, 2020). However, L-hand gestures were not all bad, as they did help with learning novel items, which also fits with previous research. For example, we know from work on mathematical instruction that gestures are particularly good at helping learners generalize what they have learned to novel problems (Novack et al., 2014), so perhaps left-hand pitch gestures served this function in our L2 learning context. Given that the literature is already mixed on the benefits of gesture in L2 phonetic learning, and given that left-hand gestures may not be the optimal choice to maximize the positive benefits of gesture, it is important to follow-up on these findings.

To do so, we conducted a second experiment that aimed to replicate the findings from Experiment 1 and extend it in three important ways. First, we added a right-hand gesture (R-gesture) training to explore whether they confer benefits that left-hand gestures do not. Second, we sought to investigate the possible neural mechanisms behind differences in performance across training types. Third, we added a subjective assessment of the training types, which may reveal benefits of multimodal input related to attitude, motivation and enjoyment. Two additional differences should also be noted. Experiment 2 was conducted in an in-person setting, which may be more suitable for our training paradigm than a virtual one in Experiment 1, and it used a within-subjects design to see if our findings from the between-subjects design in Experiment 1 could still be replicated.

https://doi.org/10.1017/langcog.2024.24 Published online by Cambridge University Press

3. Experiment 2

One novel contribution of Experiment 2 is that it investigates the neural mechanism of our results from Experiment 1. To do this, we use EEG to measure levels of 'cognitive load' during different training types. Cognitive load theory (CLT) posits that working memory has a limited capacity in terms of holding or processing new information, whereas the capacity of long-term memory for this is virtually unlimited (Miller, 1956; Peterson & Peterson, 1959; Sweller et al., 1998). Specifically, one particular type of cognitive load may be relevant: extraneous load, which is the mental effort used as a result of the design of the task (Antonenko et al., 2010). When extraneous load is high because of a cognitively ineffective presentation of information, fewer working memory resources are available to handle the intrinsic load of the task, resulting in less information learned and a higher total cognitive load (Sweller, 2010). Therefore, higher cognitive load occurs during difficult tasks that require more mental effort, while lower cognitive load reflects the opposite. CLT suggests that multimodality in L2 instruction could optimize cognitive load in a way that decreases extraneous load as much as possible (Pi et al., 2021). Differences in cognitive load across training conditions may explain some of the differences observed in Experiment 1.

Cognitive load can be assessed through alpha (8–12 Hz) and theta (4–7 Hz) band power in the EEG signal. A suppression of alpha activity is indicative of alert attention and reflects greater cognitive load (Antonenko et al., 2010). Meanwhile, theta activity is associated with enhanced internal attention and sustained neural activity, which allows for working memory representations to be maintained. Thus, increased theta activity reflects greater cognitive load. Accordingly, measuring the changes in alpha and theta brainwave rhythms provide insight into how the participant is processing information, even when they are unaware of such changes or unable to explain them (Başar et al., 1999; Klimesch et al., 2005; Pi et al., 2021).

There may also be a relationship between our multimodal trainings and participants' subjective experiences of enjoyment, motivation and mental effort. These subjective experiences were proposed to play a significant role in successful L2 acquisition in Dulay et al. (1982). The more enjoyable, relaxed and motivated learners feel, the more effectively language input is to be incorporated in their learning process (Dulay et al., 1982; Krashen, 1978; Ni, 2012). While the pitch notation training had the most robust improvement in Experiment 1, perhaps the pitch gesture training had benefits that were not captured in our dependent measure of pitch identification accuracy. Thus, a subjective assessment of participants' experiences during the training would provide useful information about the potential for gesture training to be utilized in the L2 classroom as a way to keep students engaged and enthusiastic about learning L2 phonetics.

Our aim was to replicate our original findings using the previously described training conditions in a within-subjects design, while also adding two novel dependent measures: (1) a subjective assessment of participant enjoyment/engagement during each training condition and (2) an EEG measure of alpha (8–12 Hz) and theta (4–7 Hz) power to measure levels of cognitive load during each training condition. We also added a right-handed gesture (R-gesture) training condition to our design to compare the impact of L-gesture and R-gesture on cognitive load and learning. This addition of R-gesture allowed us to address the possible limitations of gesturing with one's nondominant hand (Casasanto, 2009, 2011), in addition to

https://doi.org/10.1017/langcog.2024.24 Published online by Cambridge University Press

exploring the benefits of targeting left-hemisphere language networks involved in native speakers' pitch/tonal processing (Sato et al., 2010; Van Lancker & Fromkin, 1973; Wang et al., 2004).

With the additional measures, we predicted that beyond enhancing pitch perception, notation training will lower cognitive load and raise subjective assessments compared to the baseline flat notation. In addition, we will explore whether we replicate the first study, with L-gesture training being less effective than notation training. If confirmed, it would suggest that there may be a 'just right' amount of multimodal instruction to boost learning, reduce cognitive effort and increase motivation at the earliest stages of foreign language learning. For our new R-gesture condition, we predicted that it may be more effective than the L-gesture training for two reasons: First, R-gestures may prime more traditional left-hemisphere language areas rather than the right-hemisphere areas activated by L-gestures (Blumstein et al., 1977; Burton et al., 1998; Caplan et al., 1995; Fiez et al., 1995; Poeppel, 2003), which may be helpful for the phonemic level of learning. Second, gestures using the participants' dominant hand may be physically easier (Casasanto, 2011), potentially increasing positive subjective evaluations and reducing cognitive load.

3.1. Methods

3.1.1. Participants

Experiment 2 had 48 new right-handed, monolingual English speakers as our participants[5] (37 female, 11 male) with the same eligibility criteria as Experiment 1. Participants were recruited via posters around campus and social media posts, and all participants were compensated with $25 for their participation. People who participated in Experiment 1 were excluded. Participants also completed the Oldfield handedness inventory (Oldfield, 1971).

3.1.2. Overview

All participants underwent all four training types while wearing an EEG headset, and then completed a pitch identification test and subjective assessment in order to assess the efficacy of the four different training conditions in improving Japanese pitch accent perception, as well as the levels of effort, motivation and enjoyment each training induced.

3.1.3. Materials

3.1.3.1. Training stimuli. The training in Experiment 2 utilized the same recordings of the 20 words used in the training of Experiment 1. Because this was a within-subjects design, the words were organized into four sets of five words each, with each set being assigned to a different condition in the training. Training was blocked by condition, with the order of conditions counterbalanced across participants, resulting

[5]Power analysis was different for a within-subjects design with four conditions. Using an effect size (f) of .25, alpha level of .05 and power of .95, G*Power recommended a sample size of 36. Because the counterbalancing of condition orders created 24 unique order combinations, if we were to use 36 subjects, 10 participants would be receiving an order of training conditions already used by another participant, and 14 participants would be the only subject receiving that specific condition order. Thus, we decided to use 48 subjects, so each condition order could be evenly repeated on 2 participants.

https://doi.org/10.1017/langcog.2024.24 Published online by Cambridge University Press

in 24 unique orders. The set of items assigned to each condition were also counterbalanced so that items and condition were not confounded.

Just as in Experiment 1, each word was presented to the participant in four different carrier sentences, for a total of 20 unique sentences in each of the four training blocks. Half of the sentences were spoken by a male speaker, while the other half were spoken by a female speaker. This was repeated twice for a total of 40 trials in each training condition, with 160 trials total.

The visual stimuli for each training remained the same as in Experiment 1 (Figure 2) for baseline, notation and L-gesture training. In Experiment 2, however, we added R-gesture training which used identical visual slides as L-gesture training.

A notable difference in training in Experiment 2 compared to Experiment 1 was the sentence repetition at which the pitch pattern was revealed to participants. In Experiment 1, the target word did not show the answer on the first repetition of the sentence and then changed to reveal its pitch pattern for the second and third repetitions. This was the case for all three training conditions. However, in Experiment 2, the pitch pattern was revealed to participants on the first and second repetition of each sentence, then the indication of the pitch pattern disappeared for the third repetition. EEG measurements were only taken from the third repetition of each sentence in order to control for visual stimuli and physical movement in our EEG measurements.

All training materials were presented on Qualtrics as individual videos with the three sentence repetitions for each word. The screen loaded to a new video once the previous video finished playing. Presentation of videos was randomized within each condition block.

3.1.3.2. Pitch identification test. The pitch identification test consisted only of the 20 trained words, as performance for untrained words would not be indicative of the success of any particular training condition.[6] Each word was spoken twice, once by a female native Japanese speaker and once by a male native Japanese speaker, so that there were 40 total trials. Similar to the training, the words were presented on Qualtrics in the form of videos in which the audio was repeated twice along with the sentence written out on the screen with a space between each mora and a blue box around the target word (Figure 1). The four answer options were displayed below the sentence in the same manner as Experiment 1. We used the same two carrier sentences as we did for the posttest of Experiment 1 (*kore wa* ___ *desu yo* and *sorede* ___ *datta*).

3.1.3.3. Subjective assessment. The subjective assessment was a Google Form that participants took in front of the experimenter (See Appendix 2). The form was seven questions long, with questions aimed to measure perceived mental effort (1 = very easy; 5 = very difficult), enjoyment (1 = enjoy most; 4 = enjoy least) and attention span for each training condition (by choosing how many hours in one sitting, ranging from 0.25 hours to 2.25 hours, and days per week, ranging from 1 to 7 they would like to practice with each training condition). Participants were also instructed to select

[6]In Experiment 1, participants each only completed one training type. Thus, we could be sure that performance on novel words was due to the effect of that training type.

https://doi.org/10.1017/langcog.2024.24 Published online by Cambridge University Press

which condition was the most helpful, intuitive and motivating for them to continue learning Japanese pitch accent.

3.1.3.4. EEG apparatus. We used the Emotiv Epoc+ headset for our EEG recordings. This system allows us to measure the cumulative activity of many neurons and divide them into frequencies that serve as indirect indices of certain brain activity. The brain's electrical activity can be divided into five distinct wavebands: alpha, beta, theta, delta and gamma (Liu et al., 2013). Alpha activity encompasses the frequency range of 8–13 Hz and is pronounced in the parietal and occipital brain regions when in a state of consciousness, quiet or rest. A suppression of alpha activity is indicative of alert attention and reflects greater cognitive load (Antonenko et al., 2010). Theta activity refers to the frequency range of 4–7 Hz and occurs in the prefrontal cortices. Previous studies have shown that increased theta activity is associated with enhanced internal attention and sustained neural activity, which allows for working memory. The Emotiv software produces power values for each frequency band, reflecting the level of neuronal activity at that frequency. Thus, our EEG measure allows us to examine changes in neural activity elicited by our four different trainings, with a decrease in alpha power and an increase in theta power as our operational definition of an increase in cognitive load elicited by the trainings.

3.1.4. Procedure

Three pilot subjects were run to ensure that our procedure ran smoothly, that the EEG data were collected properly, and that the sessions were broken up in a way that minimized participant fatigue as much as possible.

3.1.4.1. Session 1: Training. Participant consent was obtained by reviewing and signing a consent form. Experimenters gave the same brief introduction to Japanese pitch accent at the beginning of the session as was used in Experiment 1. After the introduction, participants were instructed on their task during each of the training conditions and were shown examples of what the stimuli in each training condition were like. During this introduction, the experimenter demonstrated the correct gesture production for both the L-gesture and R-gesture conditions, and participants practiced these gestures while looking at the notation that was displayed during the gesture conditions. The experimenter and participant also did the gesture together to ensure that the participant understood the task before the participant was asked to practice gesturing on their own. For these introductory examples, a red line connected the vertically positioned morae of the target word to outline the shape that the participant's gesture production should follow. Just as in Experiment 1, participants were instructed to produce their L-gesture or R-gesture at the same time as they heard the speaker say the target word.

After the introduction, we placed the EEG headset on the participant and ensured high contact quality for the sensors from which we were taking EEG measurements. The Emotiv software gives contact quality measurements for each electrode, with the color green indicating high contact quality. If any of the electrodes we were taking measurements from had contact quality other than green, we made sure the hair around the sensor was moved to minimize the blocking of the sensor and/or added more saline solution to the felt pad to reduce impedance. This was done until all sensors were green. Beforehand, felt pads were

https://doi.org/10.1017/langcog.2024.24 Published online by Cambridge University Press

soaked in saline solution and placed in the sensors we planned to take recordings from. The participant was shown their EEG recordings to acclimate them to wearing the headset, and the experimenters ensured that the participant felt comfortable in the headset before the training began.

Participants then underwent training with four counterbalanced blocks, one for each condition. All participants were instructed to listen carefully and try to learn to distinguish between the difficult pitch contrasts. During each training section, participants heard 40 trials of Japanese words in carrier sentences (5 words × 4 carrier sentences × 2 repetitions), and each sentence was repeated 3 times. The four training conditions included the baseline flat notation training (shown in Figure 2 (2a)), the notation training (shown in Figure 2 (2b)), the L-gesture training (shown in Figure 2 (2c)) and the R-gesture training (same as Figure 2 (2c), but with the right hand). In the L- and R-gesture trainings, participants were instructed to create gestures with the corresponding hand that traced the shape that the word made with its vertical positioning of each mora. All participants underwent each of the four training conditions, with a 2–5 minute break after completing the first two training conditions to prevent fatigue. During the break time, the experimenter checked in with the participants and asked them how they felt it was going, and if the Epoc+ headset was still comfortable on their head. EEG contact quality was also checked during this time and adjusted as needed.

As mentioned previously for Experiment 2, the first and second time the sentence was played, the answer was displayed, either by vertical spacing of the morae (notation, L-gesture and R-gesture trainings) or by labeling with H and L (baseline flat notation training), with participants gesturing for these first two repetitions in the L-gesture and R-gesture condition. For the third repetition of each sentence, when we asked the participants to listen to the auditory stimulus carefully, the slide did not include the answer. On this third repetition of stimulus, one experimenter manually tagged its onset and offset on the Emotiv software. This marked where our alpha and theta power values would be taken from for our measurement of cognitive load. For the L-gesture and R-gesture trainings, participants were instructed not to gesture for the third repetition of each sentence. This design was intended to control for the movement of participants' hands in our EEG data, which could disrupt the EEG signal. During the gesture training conditions, the experimenter continually watched the participants' gesture production and offered feedback to correct any inaccurate gestures as necessary.

In total, the first session took about an hour and a half.

3.1.4.2. Session 2: Pitch identification test and subjective assessment. The second session occurred either the following day or 2 days after the first session. At the start of the second session, the experimenter reviewed a brief PowerPoint presentation explaining the instructions for the pitch identification test and showing examples of the questions. Then, the participant completed the test of the 20 trained words, each shown in two carrier sentences, so that there were 40 questions. As in Experiment 1, auditory stimuli were presented along with the visual slides (see Figure 1). The question presentation order was randomized, and the answer options were displayed below each sentence so that participants could refer to them as they listened to the audio input. After listening to each sentence twice, the participant selected their answer and clicked the arrow button to progress to the next question.

https://doi.org/10.1017/langcog.2024.24 Published online by Cambridge University Press

After the pitch identification test was completed, the subjective assessment was administered via Google forms. After the questionnaire, participants were debriefed and compensated with $25 over Venmo for their participation. In total, the second session took about 30 minutes to complete.

3.1.5. Design and analysis

The experiment has a one-way design with training condition as the repeated-measures independent variable. This variable has four levels: (1) baseline, (2) notation, (3) L-gesture and (4) R-gesture. There are three sets of dependent variables present in this experiment: (1) Pitch Identification: pitch accent perception accuracy of trained items, (2) Neural Cognitive Load: EEG measure of power of alpha and theta frequency bands and (3) Subjective Assessment: participant rated measures of preference and engagement of the different training conditions. This experiment was preregistered through Open Science Framework (https://osf.io/8qn9e). Our sampling plan, methods and analyses follow what was reported, with the exception of using a LME model instead of an ANOVA.

For our EEG measure of cognitive load, we utilized the alpha and theta power values measured by the Emotiv software. As per previous research, alpha power was taken from occipital electrodes (O1, O2) and theta power was taken from prefrontal electrodes (AF3, AF4) (Antonenko et al., 2010; Pi et al., 2021; Quandt et al., 2012). Pi et al. (2021) also use parietal regions in alpha power measurements, but the Emotiv Epoc+ headset does not have any of the parietal electrodes that they used, as P3 and P4 were reference sensors. Thus, alpha measurements were taken from electrodes in occipital regions only (O1 and O2). The two power values before the end tag were averaged for each trial because the Emotiv software calculated power values across an epoch of 2 seconds. Thus, the power values at the end of each sentence reflected power over the entire sentence. These power values were then averaged across trials for one alpha average and one theta average for each condition per participant. No baseline was used, as we are interested in comparing absolute power values across conditions. Additionally, a time-locked baseline would not be possible due to the nature of our stimuli; namely, the participants heard sentences without long pauses in between, resulting in a baseline that may not accurately differentiate a true resting state from an attentive state.

To eliminate artifacts caused by either overly fluctuating EEG quality or significant participant movement, we ran our data through an outlier rejection program. Participants' data were excluded if more than half of their trials were discarded by the outlier rejection program. Based on these criteria, data from 5 and 13 participants were excluded from all 4 conditions for alpha and theta measurements, respectively.

3.2. Results

3.2.1. Pitch identification

Pitch identification accuracy was analyzed using a mixed effects logistic regression model, with the same R packages listed for Experiment 1. A model with a fixed effect of condition and random intercepts of participant and word revealed no significant effect of condition. This indicates that the varying levels of multimodal input across conditions did not have an effect on learning outcomes (see Table 1 for learning outcome results and Appendix 3.2 for the model output).

https://doi.org/10.1017/langcog.2024.24 Published online by Cambridge University Press

Table 1. Pitch identification accuracy for words corresponding to the four training conditions in Experiment 2

Training condition	Average percent correct of corresponding words
Baseline	50.00% (SD = 22.2%)
Notation	47.50% (SD = 22.0%)
L–gesture	46.04% (SD = 21.2%)
R–gesture	48.13% (SD = 19.7%)

3.2.2. Neural cognitive load

3.2.2.1. Alpha EEG. A mixed effects model with participant as a random intercept revealed no significant effect of condition on alpha power[7] (see Appendix 3.3 for the model output). Additionally, there were no significant correlations between alpha EEG and pitch identification scores.

3.2.2.2. Theta EEG. A mixed effects model with the participant as a random intercept revealed no significant effect of condition on theta power, nor were there any significant correlations between theta EEG and pitch identification scores (see Appendix 3.4 for the model output).

3.2.3. Subjective assessments

Table 2 shows a summary of all results in subjective assessments.

3.2.3.1. Effort rating. A significant effect of condition was found among participants' perceived effort used to distinguish the pitch contrasts, $\chi^2(3) = 40.410$, $p < .001$. Wilcoxon tests revealed that perceived mental effort was reported as greater in the baseline condition compared to the notation ($Z = 4.01$, $p < .001$), L-gesture ($Z = 3.61$, $p < .001$) and R-gesture ($Z = 4.34$, $p < .001$) conditions. In addition, the L-gesture

Table 2. Means and standard deviations of participants' responses to the subjective assessment survey in Experiment 2

	Baseline		Notation		L-gesture		R-gesture	
	Mean/count	SD	Mean/count	SD	Mean/count	SD	Mean/count	SD
Effort during training								
Effort	3.54	1.16	2.40	1.14	2.71	1.01	2.31	0.93
Time Investment								
Hours	0.51	0.30	0.78	0.49	0.64	0.39	0.78	0.41
Days	2.02	1.30	2.92	1.29	2.52	1.37	3.00	1.34
Top choice (out of 48)								
Most intuitive	2	NA	21	NA	6	NA	19	NA
Most helpful	5	NA	16	NA	9	NA	18	NA
Most motivation	5	NA	22	NA	8	NA	13	NA
Most enjoyment	5	NA	17	NA	6	NA	20	NA

[7]We were unable to include a random intercept of word in our model due to constraints in our EEG software.

https://doi.org/10.1017/langcog.2024.24 Published online by Cambridge University Press

condition was judged to be more effortful than the R-gesture condition ($Z = 2.59$, $p = .02$).

3.2.3.2. Time investment: Hours. A repeated measures ANOVA on hours of investment revealed a main effect of training condition, $F(3,141) = 9.190$, $p < .001$, $\eta p^2 = 0.164$. Bonferroni *t* tests revealed that participants were willing to practice notation, $t(47) = -3.504$, $p < .001$ and R-gesture training, $t(47) = 4.826$, $p < .001$, for a significantly longer time than baseline flat notation. In addition, participants were willing to invest more time doing R-gesture training, $t(47) = 3.528$, $p < .001$, than L-gesture training.

3.2.3.3 Time investment: Days. A repeated measures ANOVA on days of practice revealed a main effect of training condition, $F(3,141) = 8.798$, $p < .001$, $\eta p^2 = 0.158$. Participants reported wanting to spend significantly fewer days practicing the flat notation baseline than the notation, $t(47) = 3.877$, $p < 0.001$ and R-gesture, $t(47) = 3.760$, $p < .001$, conditions. In addition, participants were willing to invest more time doing R-gesture training, $t(47) = 3.91$, $p < .001$, than L-gesture training.

3.2.3.4. Intuitiveness. There was a significant nonrandom distribution of what condition was considered most intuitive, $\chi^2(3) = 22.17$, $p < 0.001$. Participants overwhelmingly preferred the notation or R-gesture conditions (40 participants) over the baseline or L-gesture conditions (8 participants).

3.2.3.5. Helpfulness. There was a significant nonrandom distribution of what condition was considered most helpful, $\chi^2(3) = 9.17$, $p = .027$. Participants strongly preferred the notation or R-gesture conditions (34 participants) over the baseline or L-gesture conditions (14 participants).

3.2.3.6. Motivation. There was a significant nonrandom distribution of what condition was considered the most motivating, $\chi^2(3) = 13.83$, $p = 0.003$. Participants strongly preferred the notation or R-gesture conditions (35 participants) over the baseline or L-gesture conditions (13 participants).

3.2.3.7. Enjoyment. There was a significant nonrandom distribution of what condition was considered most enjoyable, $\chi^2(3) = 14.50$, $p = 0.002$. Participants strongly preferred the notation or R-gesture conditions (37 participants) over the baseline or L-gesture conditions (11 participants).

3.3. Discussion

The results are partially consistent with our preregistered predictions. Although the learning outcomes and EEG measures revealed no effects of training condition, the notation and R-gesture condition vastly outperformed the baseline and L-gesture conditions in all of the subjective assessments.

The overall lack of difference in both EEG power values and pitch identification accuracy across conditions suggests two possibilities: 1) there is no effect of the different trainings on cognitive load and learning or 2) our design limited our ability

https://doi.org/10.1017/langcog.2024.24 Published online by Cambridge University Press

to detect these differences. Regarding the second possibility, the lack of significance in our EEG data may be a result of selecting a time window for our EEG that was too far removed from the effects of the training. Recall that measures of alpha and theta power were taken during the third repetition of the sentence during which the correct pitch pattern was not detectable using any modality other than audio. Participants were instructed to think about the correct pitch pattern they had just learned from the two repetitions prior, but we have no way of verifying that they were actually doing this during the time interval that our alpha and theta power values were taken. We time locked in this way to control for hand movement creating artifacts in our EEG recording (Quandt et al., 2012), and we believed that changes in cognitive load would be detectable downstream immediately after multimodal presentation. However, while our conservative recording window makes sense for avoiding motion artifacts, it may have shifted focus away from when we would have seen larger training effects on our EEG power analysis. Perhaps recording during the training window when the multimodal training input was actually being delivered would have differentiated our four conditions.

As there was no difference in pitch identification accuracy for the R-gesture and L-gesture conditions, there does not appear to be a learning benefit of one type of gesture over the other. This, combined with there being no advantage of either gesture in reducing cognitive load compared to baseline, suggests that gesture may not target the neural mechanisms involved in the processing of Japanese pitch accents. This is surprising given the EEG research on native language processing showing that observing beat gestures (gestures that emphasize prosodic stress) affects early phonetic (Biau & Soto-Faraco, 2013) and later semantic (Morett et al., 2020) processing of L1 words. Moreover, past research has shown that producing right-handed pitch gestures can help with the perception of L2 Chinese lexical tones (Baills et al., 2019). One possible explanation for this inconsistency is that in the present study, the Japanese words and pitch gestures were more complex than in previous studies. Each Japanese word was four morae in length, and in addition, was presented in a carrier sentence, which is different from past gesture research on tonal languages (e.g., Mandarin in Baills et al., 2019; Zhen et al., 2019). Perhaps this complexity was too much for novice learners, which suggests that not all pitch gestures and languages are created equal. Future research should explore how linguistic and gestural differences across languages may modulate the extent to which the hands are integrated with speech during L2 perception and learning.

Despite our EEG data suggesting that training condition did not affect cognitive load, the other novel dependent variable of participant subjective assessment did show robust effects of the training condition. Participants indicated a strong preference for the notation and R-gesture conditions, and a strong dispreference toward the flat notation baseline condition, which participants judged to be the most effortful. The notation and R-gesture conditions were chosen by the highest number of participants for being the most intuitive, helpful, enjoyable and motivating. Additionally, the R-gesture and notation condition outperformed the L-gesture and baseline condition on the number of hours (in one sitting) and days per week participants would be willing to practice. We did not find a significant difference between the number of hours and days participants would be willing to practice in the notation compared to the R-gesture condition, further implicating these two conditions as the most engaging.

https://doi.org/10.1017/langcog.2024.24 Published online by Cambridge University Press

Thus, we can conclude that while multimodal training may not always have a more robust effect on auditory learning outcomes compared to instruction with lower levels of multimodality, it does, however, clearly affect learner engagement, motivation and enjoyment. Moreover, the fact that R-gesture training was preferred over L-gesture training suggests that not all multimodal inputs are created equal. We further explore these intriguing possibilities in Section 4.

4. General discussion

Our study examined the effects of increasing levels of multimodality in L2 Japanese pitch accent training: We compared the baseline notation displaying pitch patterns with H and L with other multimodal forms: the written notation that spatially mimicked the pitch patterns and the two types of hand gestures in which participants traced the pitch patterns on the screen. Experiment 1 had a between-subjects design and examined gestures using only the left hand (to engage the contralateral right hemisphere specialized for suprasegmental pitch processing). The results indicate that the notation that spatially mimicked pitch patterns is most robust in its benefits, as participants in this condition improved on trained and novel words alike. This supports the claim in the extant L2 phonetics research that perceptual learning takes place effectively when provided with notation that visually mimics speech characteristics (Hardison & Pennington, 2021). Experiment 2 extended Experiment 1 in three ways: adding a R-gesture condition (to engage more segmental language areas in the left hemisphere), introducing a neural correlate of cognitive load (measured by EEG alpha and theta power) and a subjective assessment of the trainings (e.g., 'Which training did you find the most helpful?'). The results showed that although there were no differences among our four training conditions on auditory learning outcomes or EEG power, participants made the most positive subjective evaluations about the pitch-mimicking notation and R-gesture training conditions. Together, the results suggest that there may be a 'just right' amount of multimodal instruction to boost learning and increase engagement during foreign language pitch instruction.

4.1. A multimodal sweet spot

The results across both experiments suggest that our notation and R-gesture training sessions offer an ideal balance of multimodal features to facilitate learning and foster engagement. We conclude this based on two key findings across our two experiments. First, Experiment 1 showed that training with the notation mimicking pitch-height improved pitch accent perception for both trained and novel words. Meanwhile, our baseline training improved only on trained words, and our L-gesture training improved only on novel words. Importantly, our L-gesture training was visually identical to the notation condition, and differed only by the production of hand gestures that traced the pitch patterns. This suggests that the gestures may have counteracted the benefits of the notation for trained items, potentially by distracting from the phonological input. In this sense, the notation may have just enough multimodality to boost learning compared to instruction with less rich multimodality, but not too much multimodal information so that participants are overwhelmed with information. Unfortunately, we did not have a R-gesture training condition in Experiment 1 due to power constraints of our

https://doi.org/10.1017/langcog.2024.24 Published online by Cambridge University Press

between-subjects design, so we do not know whether adding R-gestures to pitch notations would have produced similar learning outcomes as notation training alone.

Second, Experiment 2 showed that notation and R-gesture training were most commonly chosen as the most appealing. These two conditions were judged to be least effortful, while also being the most motivating, intuitive, helpful and enjoyable. For example, with regard to intuitiveness, participants were five times more likely to say that the notation and R-gesture condition were their favorites. These two conditions also spurred learners to want to continue learning about Japanese pitch contrasts (e.g., they would be willing to spend 50% more time on them than they would on baseline training). Given that emotional and attitudinal factors are key parts of what determines success in L2 learning (Krashen, 1978), these preferences are noteworthy even in the absence of differences in learning outcomes (at least in Experiment 2). From a purely practical point of view, even if different multimodal techniques do not differ much in their actual effectiveness, there is something to be said for keeping early L2 learners in their seats longer and eager to keep learning.

The picture that emerges is that there may be a 'just right' amount *and* type of multimodal input for helping people learn novel L2 prosody. On the one hand, this claim is compatible with DCT (Clark & Paivio, 1991) and also meshes with empirical findings showing that enjoyment and motivation are elicited by multimodal and embodied instruction and practice (Asher, 1966; Chicho, 2021; Gullberg, 1998; Smotrova, 2017). For example, in an actual English as a foreign language classroom, Chicho (2021) did interviews of student learners and found that embodied and multimodal learning elicited high learner motivation, confidence and self-development. Here is one student's positive reflection: 'Embodied learning approach changed my perception toward the learning. Previously, I thought that studying a language was very boring and difficult, but now I totally … think that language learning is more interesting when it happens naturally' (p. 55). This suggests that even though the present study used artificial and controlled experimental contexts, the positive results from the notation and R-gesture condition may generalize to more naturalistic contexts.

This 'just right' conclusion is consistent with research suggesting that in some L2 contexts, some forms of multimodal and embodied input can be detrimental (Bailles et al., 2019; Hirata & Kelly, 2010; Hoetjes & Van Maastricht, 2020; Kelly & Lee, 2012). Indeed, even though the amount of multimodal input was identical in the R-gesture and L-gesture conditions, participants showed a strong and consistent preference for the R-gesture condition. This finding fits well with Casasanto's body specificity hypothesis (2009), which holds that specific body characteristics (i.e., handedness) result in corresponding differences in mental representations and cognition. Specifically, Casasanto showed that right-handed people implicitly map positive valence toward rightward space, and his explanation was that people unconsciously link bodily action with good and bad, such that we associate good things more strongly with the side that we more fluently interact with and bad concepts with our nondominant side. So it is possible that when people produced L-gestures, they struggled with using their nondominant hand and made a more negative association with what they were learning, causing them to think the task was harder and less enjoyable. All in all, this suggests that it is the presence *and* quality of multimodal input that matter.

https://doi.org/10.1017/langcog.2024.24 Published online by Cambridge University Press

It is well established that hand gestures conveying semantic information, such as iconics, are tightly integrated with the meaning of speech during development and learning (Goldin-Meadow, 2005: Kelly, 2017). However, other types of gestures conveying prosodic information, such as deictics, metaphorics and beats, also play semantic, pragmatic and syntactic roles with speech during development and learning (for a review, see Hübscher & Prieto, 2019). For example, Hübscher and Prieto (2019) argue that prosodic information across speech and prosodic gesture is integrated in such a way to help children to understand the communicative intention and meaning of utterances. This integrated relationship stays strong even after language mastery, as adults use commonly gestures, such as beats, to draw attention to words within an utterance (Krahmer & Swerts, 2007). Thus, it seems that hand gestures are tightly integrated with speech at the suprasegmental level of language to communicate the meanings and intentions of full utterances.

However, the evidence is more mixed whether these prosodic benefits of the hands extend into smaller time windows, such as segmental properties within single words (Kelly, 2017). Indeed, while prosodic gestures, like beats, deictics or metaphorics, seem designed to emphasize words over the course of an utterance, these gestures seem less naturally suited to highlighted phonetic features *within* words (Hirata et al., 2014; Hirata & Kelly, 2010; Hoetjes & Van Maastricht, 2020; Morett et al., 2022).[8] Indeed, the results reported in the present study suggest that prosodic information in pitch gestures may overwhelm or distract from the predominantly auditory focus that is required for L2 pitch learning, especially among novice learners.

Given the high demand that L2 pitch learning places on learner's auditory focus, it may be important to introduce aspects of embodied learning to engage and motivate learners so that they are not fatigued from the intense auditory attention that pitch accent discrimination requires. Our subjective assessment has proven R-gestures to be an effective way to do this, implying that certain types of gestures that are comfortable and natural for learners can provide a powerful affective component for learning L2 prosody. Thus, the prosodic benefit of gesture for segmental learning of novel L2 pitch patterns may be more indirect than the natural benefits of gestures at the suprasegmental level, as it serves the purpose of increasing learner enjoyment, which will indirectly lead to increased investment, rather than boosting learning outcomes themselves.

4.2. Limitations and future studies

There are limitations of this study that deserve attention. First, the study focuses on completely naïve learners of Japanese in an artificial learning environment. It is not clear how these results would generalize to more advanced students in real language classrooms where there is a much heavier emphasis on learning L2 meaning in actual communicative contexts. Indeed, there are likely very different levels of investment and motivation between a controlled experimental setting, such as the present study, and the rich and dynamic context of a real language classroom. It would be interesting to investigate how phonetic training – for example, in a controlled L2 lab context –

[8]It is worth noting that at least one study has found that while gesture instruction does not help L2 segmental perception, it does help learners better *produce* novel segmental contrasts (Xi et al., 2020).

https://doi.org/10.1017/langcog.2024.24 Published online by Cambridge University Press

interacts with higher-level instruction of vocabulary and grammar in an actual classroom. Pedagogically, the present training might effectively combine with an approach that encourages the benefits of an initial 'silent period' – with careful listening of auditory input but without learning word meaning or engaging in oral practice (Dulay et al., 1982; Ervin-Tripp, 1974; Gary, 1978; Neufeld, 1978, 1988; Pardo, 1995; Postovsky, 1974, 1977). For example, Hirata and Kato's (2007) pilot study suggest that learners gain distinct benefit from engaging first in auditory perception learning (just like the present study) and then moving onto vocabulary learning, as opposed to the other way around. This is an interesting future direction since few language teaching approaches even consider this benefit despite the potentials suggested in the literature above.

Second, the within-subjects design of Experiment 2 may have obscured training differences among our conditions. Because participants received all levels of training, it is likely that posttest scores within one condition were influenced by the other conditions. For example, it is possible that being exposed to the notation condition could have generalized benefits for the baseline condition. Because we had a between-subjects design in Experiment 1, we could actually test for these generalized learning effects, but our within-subjects design in Experiment 2 did not allow this examination. Thus, the fact that there were no training effects in Experiment 2 is hard to interpret, especially in light of the significant differences we found across the training conditions in Experiment 1.

Third, our EEG design limited our ability to detect differences in pitch identification and alpha and theta power across the four training conditions. Because we recorded the EEG to identical sentences – that *followed* the actual multimodal input – it is possible that we underestimated cognitive load differences during training. It is possible that if the EEG signal were taken while the participants were gesturing and viewing the visual–spatial representation of pitch accent, we may have found a greater effect of training condition on cognitive load. Future studies should more directly measure EEG differences during the multimodal encoding phase of training (Pi et al., 2021).

A fourth limitation, specific to the two gesture conditions, is that the present study cannot distinguish between the costs and benefits of observing and producing gestures. Both gesture observation and production are important parts of language use (Church et al., 2017; Cienki & Müller, 2008), but they may have different functions in their assistance of L2 phonetic learning (Baills et al., 2019). Perhaps observing gestures eliminates the distraction that may occur by producing them, in turn inducing less cognitive load and enhancing learning (Pi et al., 2021). Then again, it is also possible that observing gestures along with a visual–spatial notation may overwhelm the participants with visual information when trying to differentiate difficult phonetic distinctions (Kelly & Lee, 2012). Directly comparing gesture observation and production in learning outcomes and cognitive load (and subjective evaluations) would provide valuable information for L2 teachers as they decide how gestures should be integrated into their classrooms.

Finally, on a related note, future research should explore the relationship between training conditions and testing conditions. Note that our gesture training involved producing gestures, but not perceiving them. This may explain why our gesture training did not influence our dependent measure of speech *perception.* Given that previous research (e.g., Xi et al., 2020) has shown that gesture training helps with the production – but not the perception – of novel L2 speech, it is possible that our

https://doi.org/10.1017/langcog.2024.24 Published online by Cambridge University Press

gesture training may have helped with the production of Japanese pitch contrasts. To uncover these sorts of nuanced effects, future research should combine measures of perception and production when testing the effectiveness of gesture training on novel L2 contrasts.

5. Conclusion

Our study contributes to the existing L2 learning literature by taking a multimodal perspective. Gesture is rarely studied in tandem with other visual–spatial representations of phonology, and it is even more rare to combine learning outcomes, brain measures and subjective evaluations into a single study. Additionally, little work has been done investigating how hand use modulates the benefits of gesture production in an L2 learning context. Our results suggest that *too little* multimodal input (flat notation of high and low) and the *wrong kind* of multimodal instruction (L-gestures) may yield nonoptimal learning outcomes and negative evaluations of that learning. Rather, it appears that a 'just right' amount of multimodal input – in our case, pitch notation alone or coupled with a right-handed gesture – produces the most beneficial learning opportunities for L2 learners.

Acknowledgements. We would like to thank Leo M. Shiner, Paige Avila, Tim Collett, Dr. Bruce Hansen, and Dr. Masaaki Kamiya for their assistance that was vital to our success.

Funding statement. The project was funded by the Research Council and the Center for Language and Brain of Colgate University.

References

Allen, L. Q. (1995). The effects of emblematic gestures on the development and access of mental representations of French expressions. *The Modern Language Journal*, 79(4), 521–529.

Antonenko, P., Paas, F., Grabner, R., & Van Gog, T. (2010). Using electroencephalography to measure cognitive load. *Educational Psychology Review*, 22(4), 425–438.

Asher, J. J. (1966). The learning strategy of the total physical response: A review. *The Modern Language Journal*, 50(2), 79–84.

Baills, F., Suárez-González, N., González-Fuente, S., & Prieto, P. (2019). Observing and producing pitch gestures facilitates the learning of Mandarin Chinese tones and words. *Studies in Second Language Acquisition*, 41(1), 33–58.

Banno, E., Ikeda, Y. Ohno, Y., Shinagawa, C., & Tokashiki, K. (2020). *Genki: An integrated course in elementary Japanese I*. The Japan Times Publishing.

Başar, E., Başar-Eroğlu, C., Karakaş, S., & Schürmann, M. (1999). Are cognitive processes manifested in event-related gamma, alpha, theta and delta oscillations in the EEG? *Neuroscience Letters*, 259(3), 165–168.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using ime4. *Journal of Statistical Software*, 67(1), 1–48.

Beckman, M. (1986). *Stress and non-stress accent*. Foris Publications.

Biau, E., & Soto-Faraco, S. (2013). Beat gestures modulate auditory integration in speech perception. *Brain and Language*, 124(2), 143–152.

Blumstein, S. E., Baker, E., & Goodglass, H. (1977). Phonological factors in auditory comprehension in aphasia. *Neuropsychologia*, 15(1), 19–30.

Bolinger, D. (1983). Intonation and gesture. *American Speech*, 58(2), 156–174.

Broadcasting Culture Research Institute (1998). *NHK日本語発音アクセント辞典 (NHK Japanese Pronunciation Accent Dictionary)*. NHK Publishing (in Japanese).

Broadcasting Culture Research Institute (2016). *NHK日本語発音アクセント新辞典 (NHK Japanese Pronunciation Accent New Dictionary)*. NHK Publishing (in Japanese).

https://doi.org/10.1017/langcog.2024.24 Published online by Cambridge University Press

Burton, M. W., Blumstein, S. E., & Small, S. L. (1998). The neural basis of phonological processing: The role of segmentation. *Brain and Language*, 65(1), 249–251.

Caplan, D., Gow, D., & Makris, N. (1995). Analysis of lesions by MRI in stroke patients with acoustic-phonetic processing deficits. *Neurology*, 45(2), 293–298.

Casasanto, D. (2009). Embodiment of abstract concepts: Good and bad in right-and left-handers. *Journal of Experimental Psychology: General*, 138(3), 351.

Casasanto, D. (2011). Different bodies, different minds: The body specificity of language and thought. *Current Directions in Psychological Science*, 20(6), 378–383.

Chicho, K. Z. H. (2021). Embodied learning implementation in EFL classroom: A qualitative study. *International Journal of Social Sciences & Educational Studies*, 8(1), 51–58.

Church, R. B., Alibali, M. W., & Kelly, S. D. (2017). *Why gesture?: How the hands function in speaking, thinking and communicating* (Vol. 7). John Benjamins Publishing Company.

Cienki, A., & Müller, C. (2008). Metaphor, gesture, and thought. In R. W. Gibbs Jr (Ed.), *The Cambridge handbook of metaphor and thought* (pp. 483–501). Cambridge University Press.

Clark, J. M., & Paivio, A. (1991). Dual coding theory and education. *Educational Psychology Review*, 3(3), 149–210.

Dargue, N., Sweller, N., & Jones, M. P. (2019). When our hands help us understand: A meta-analysis into the effects of gesture on comprehension. *Psychological Bulletin*, 145(8), 765–784.

Dulay, H., Burt, M., & Krashen, S. D. (1982). *Language two*. Oxford University Press.

Ervin-Tripp, S. (1974). Is second language learning like the first? *TESOL Quarterly*, 8, 111–127.

Fiez, J. A., Raichle, M. E., Miezin, F. M., Petersen, S. E., Tallal, P., & Katz, W. F. (1995). PET studies of auditory and phonological processing: Effects of stimulus characteristics and task demands. *Journal of Cognitive Neuroscience*, 7(3), 357–375.

Gary, J. O. (1978). Why speak if you don't need to? The case for a listening approach to beginning foreign language learning. In W. C. Ritchie (Ed.), *Second language acquisition research—issues and implications* (pp. 185–199). Academic Press.

Gluhareva, D., & Prieto, P. (2017). Training with rhythmic beat gestures benefits L2 pronunciation in discourse-demanding situations. *Language Teaching Research*, 21(5), 609–631.

Goldin-Meadow, S. (2005). *Hearing gesture: How our hands help us think*. Harvard University Press.

Goss, S. (2020). Exploring variation in nonnative Japanese learners' perception of lexical pitch accent: The roles of processing resources and learning context. *Applied Psycholinguistics*, 41, 25–49.

Gullberg, M. (1998). *Gesture as a communication strategy in second language discourse: A study of learners of French and Swedish*. Lund University Press.

Gullberg, M. (2006). Some reasons for studying gesture and second language acquisition (Hommage à Adam Kendon). *International Review of Applied Linguistics*, 44(2), 103–124.

Hannah, B., Wang, Y., Jongman, A., Sereno, J. A., Cao, J., & Nie, Y. (2017). Cross-modal association between auditory and visuospatial information in Mandarin tone perception in noise by native and non-native perceivers. *Frontiers in Psychology*, 8(1), 2051.

Hardison, D. M. (2005). Contextualized computer-based L2 prosody training: Evaluating the effects of discourse context and video input. *CALICO Journal*, 22(2), 175–190.

Hardison, D. M., & Pennington, M. C. (2021). Multimodal second-language communication: Research findings and pedagogical implications. *RELC Journal*, 52(1), 62–76.

Hirano-Cook, E. (2011). *Japanese pitch accent acquisition by learners of Japanese: Effects of training on Japanese accent instruction, perception, and production.* Doctoral dissertation, University of Kansas.

Hirata, Y. (2004a). Computer assisted pronunciation training for native English speakers learning Japanese pitch and durational contrasts. *Computer Assisted Language Learning*, 17(3–4), 357–376.

Hirata, Y. (2004b). Training native English speakers to perceive Japanese length contrasts in word versus sentence contexts. *Journal of the Acoustical Society of America*, 116(4), 2384–2394.

Hirata, Y. (2015). L2 phonetics and phonology. In H. Kubozono (Ed.), *Phonetics & phonology volume: The handbook of Japanese language and linguistics* (pp. 719–762). De Gruyter Mouton.

Hirata, Y. & Kato, H. (2007). Native English speakers' acquisition of Japanese quantity distinction: Interaction between phonemic and lexical processing. In *A Final Report on Research Activities and Findings at Advanced Telecommunications Research Institute International*, Kyoto, Japan. Unpublished manuscript available upon request from yhirata@colgate.edu.

Hirata, Y., & Kelly, S. D. (2010). Effects of lips and hands on auditory learning of second-language speech sounds. *Journal of Speech, Language, and Hearing Research*, 53(2), 298–310.

https://doi.org/10.1017/langcog.2024.24 Published online by Cambridge University Press

Hirata, Y., Kelly, S. D., Huang, J., & Manansala, M. (2014). Effects of hand gestures on auditory learning of second-language vowel length contrasts. *Journal of Speech, Language, and Hearing Research*, 57(6), 2090–2101.

Hirayama, T. (Ed.) (1960). *Zenkoku Akusento Jiten (Accent Dictionary of All Japan)*. Tokyodo (in Japanese).

Hoetjes, M., & Van Maastricht, L. (2020). Using gesture to facilitate L2 phoneme acquisition: The importance of gesture and phoneme complexity. *Frontiers in Psychology*, 11(1), 575032.

Hostetter, A. B. (2011). When do gestures communicate? A meta-analysis. *Psychological Bulletin*, 137(2), 297.

Hostetter, A. B., & Alibali, M. W. (2008). Visible embodiment: Gestures as simulated action. *Psychonomic Bulletin & Review*, 15(3), 495–514.

Hubbard, A. L., Wilson, S. M., Callan, D. E., & Dapretto, M. (2009). Giving speech a hand: Gesture modulates activity in auditory cortex during speech perception. *Human Brain Mapping*, 30(3), 1028–1037.

Hübscher, I., & Prieto, P. (2019). Gestural and prosodic development act as sister systems and jointly pave the way for children's sociopragmatic development. *Frontiers in Psychology*, 10, 1259.

Jorden, E. H., & Noda, M. (1987). *Japanese: The spoken language part 1*. Yale University Press.

Kelly, S. D. (2017). Exploring the boundaries of gesture-speech integration during language comprehension. In R. B. Church, M. W. Alibali, & S. D. Kelly (Eds.), *Why gesture? How the hands function in speaking, thinking and communicating* (pp. 243–265). John Benjamins Publishing Company.

Kelly, S., Bailey, A., & Hirata, Y. (2017). Metaphoric gestures facilitate perception of intonation more than length in auditory judgments of non-native phonemic contrasts. *Collabra: Psychology*, 3(1), 7.

Kelly, S. D., & Lee, A. (2012). When actions speak too much louder than words: Gesture disrupts word learning when phonetic demands are high. *Language and Cognitive Processes*, 27(6), 793–807.

Kindaichi, H. (1996). *明解日本語アクセント辞典第2版 (Lucid Japanese Accent Dictionary Second Edition)*. Sanseido Publishing (in Japanese).

Klein, D., Zatorre, R. J., Milner, B., & Zhao, V. (2001). A cross-linguistic PET study of tone perception in Mandarin Chinese and English speakers. *Neuroimage*, 13(4), 646–653.

Klimesch, W., Schack, B., & Sauseng, P. (2005). The functional significance of theta and upper alpha oscillations. *Experimental Psychology*, 52(2), 99–108.

Krahmer, E., & Swerts, M. (2007). The effects of visual beats on prosodic prominence: Acoustic analyses, auditory perception and visual perception. *Journal of Memory and Language*, 57(3), 396–414.

Krashen, S. (1978). Individual variation in the use of the monitor. In W. C. Ritchie (Ed.), *Second language acquisition research: Issues and Implications* (pp. 175–183). Academic Press.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2015). Package 'lmertest'. R package version 2.0, 734.

Lattner, S., Meyer, M. E., & Friederici, A. D. (2005). Voice perception: Sex, pitch, and the right hemisphere. *Human Brain Mapping*, 24(1), 11–20.

Lazaraton, A. (2004). Gesture and speech in the vocabulary explanations of one ESL teacher: A microanalytic inquiry. *Language learning*, 54(1), 79–117.

Lenth, R., Singmann, H., Love, J., Buerkner, P., & Herve, M. (2019). Package 'emmeans'. *R package version* 1(3.2).

Liu, N. H., Chiang, C. Y., & Chu, H. C. (2013). Recognizing the degree of human attention using EEG signals from mobile sensors. *Sensors*, 13(8), 10273–10286.

Liu, Y., Wang, M., Perfetti, C. A., Brubaker, B., Wu, S., & MacWhinney, B. (2011). Learning a tonal language by attending to the tone: An in vivo experiment. *Language Learning*, 61(4), 1119–1141.

Loui, P., Li, H. C., & Schlaug, G. (2011). White matter integrity in right hemisphere predicts pitch-related grammar learning. *Neuroimage*, 55(2), 500–507.

McCafferty, S. G. (2002). Gesture and creating zones of proximal development for second language learning. *The Modern Language Journal*, 86(2), 192–203.

McCafferty, S. G., & Stam, G. (Eds). (2009). *Gesture: Second language acquisition and classroom research*. Routledge.

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, 264, 746–748.

McNeill, D. (1985). So you think gestures are nonverbal? *Psychological Review*, 92(3), 350–371.

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, 63(2), 81.

Morett, L. M., & Chang, L. Y. (2015). Emphasising sound and meaning: Pitch gestures enhance Mandarin lexical tone acquisition. *Language, Cognition and Neuroscience*, 30(3), 347–353.

Morett, L. M., Feiler, J. B., & Getz, L. M. (2022). Elucidating the influences of embodiment and conceptual metaphor on lexical and non-speech tone learning. *Cognition*, 222, 105014.

https://doi.org/10.1017/langcog.2024.24 Published online by Cambridge University Press

Morett, L. M., Landi, N., Irwin, J., & McPartland, J. C. (2020). N400 amplitude, latency, and variability reflect temporal integration of beat gesture and pitch accent during language processing. *Brain Research*, 1747, 147059.

Motohashi-Siago, M., & Hardison, D. M. (2009). Acquisition of L2 Japanese geminates: Training with waveform displays. *Language Learning & Technology*, 13(2), 29–47.

Muradás-Taylor, B. (2022). Accuracy and stability in English speakers' production of Japanese pitch accent. *Language and Speech*, 65(2), 377–403.

Neufeld, G. (1978). On the acquisition of prosodic and articulatory features in adult language learning. *The Canadian Modern Language Review*, 34, 168–194.

Neufeld, G. (1988). Phonological asymmetry in second language learning and performance. *Language Learning*, 38, 531–559.

Ni, H. (2012). The effects of affective factors in SLA and pedagogical implications. *Theory and Practice in Language Studies*, 2(7), 1508–1513.

Noto, H. (1992). *Communicating in Japanese*. Sotakusha Publishing.

Novack, M. A., Congdon, E. L., Hemani-Lopez, N., & Goldin-Meadow, S. (2014). From action to abstraction: Using the hands to learn math. *Psychological Science*, 25(4), 903–910.

Oldfield, R. C. (1971). The assessment and analysis of handedness: the Edinburgh inventory. *Neuropsychologia*, 9(1), 97–113.

Pardo, D. B. (1995). Delay in oral production and pronunciation achievement in a foreign language. *Proceedings of the 13th International Congress of Phonetic Sciences*, 1, 270–273.

Peretz, I. & Zatorre, R. J. (2005). Brain organization for music processing. *Annual Review of Psychology*, 56, 89–114.

Peterson, L., & Peterson, M. J. (1959). Short-term retention of individual verbal items. *Journal of Experimental Psychology*, 58(3), 193–198.

Pi, Z., Zhu, F., Zhang, Y., & Yang, J. (2021). An instructor's beat gestures facilitate second language vocabulary learning from instructional videos: Behavioral and neural evidence. *Language Teaching Research*, 1–29.

Poeppel, D. (2003). The analysis of speech in different temporal integration windows: Cerebral lateralization as 'asymmetric sampling in time'. *Speech Communication*, 41(1), 245–255.

Postovsky, V. (1974). Effects of delay in oral practice at the beginning of second language learning. *Modern Language Journal*, 58, 229–239.

Postovsky, V. (1977). Why not start speaking later? In M. Burt, H. Dulay, & M. Finocchiaro (Eds.), *Viewpoints on English as a second language* (pp. 17–26). Regents.

Quandt, L. C., Marshall, P. J., Shipley, T. F., Beilock, S. L., & Goldin-Meadow, S. (2012). Sensitivity of alpha and beta oscillations to sensorimotor characteristics of action: An EEG study of action production and gesture observation. *Neuropsychologia*, 50(12), 2745–2751.

Sakamoto, E. (2011). *An investigation of factors behind foreign accent in the L2 acquisition of Japanese lexical pitch accent by English speakers.* [Doctoral Dissertation]. University of Edinburgh.

Sato, Y., Sogabe, Y., & Mazuka, R. (2010). Development of hemispheric specialization for lexical pitch–accent in Japanese infants. *Journal of Cognitive Neuroscience*, 22(11), 2503–2513.

Schlaug, G., Marchina, S., & Norton, A. (2009). Evidence for plasticity in white-matter tracts of patients with chronic Broca's aphasia undergoing intense intonation-based speech therapy. *Annals of the New York Academy of Sciences*, 1169(1), 385–394.

Sidtis, J. J. (1980). On the nature of the cortical function underlying right hemisphere auditory perception. *Neuropsychologia*, 18(3), 321–330.

Sime, D. (2006). What do learners make of teachers' gestures in the language classroom? *International Review of Applied Linguistics in Language Teaching*, 44(2), 211–230.

Sluijter, A. M. C. & van Heuven, V. J. (1996a). Spectral balance as an acoustic correlate of linguistics stress. *Journal of the Acoustical Society of America*, 100(4), 2471–2485.

Sluijter, A. M. C. & van Heuven, V. J. (1996b). Acoustic correlates of linguistics stress and accent in Dutch and American English. *Proceedings of the Fourth International Conference on Spoken Language Processing*, 2, 630–633.

Smotrova, T. (2017). Making pronunciation visible: Gesture in teaching pronunciation. *Tesol Quarterly*, 51 (1), 59–89.

Smotrova, T., & Lantolf, J. P. (2013). The function of gesture in lexically focused L2 instructional conversations. *The Modern Language Journal*, 97(2), 397–416.

https://doi.org/10.1017/langcog.2024.24 Published online by Cambridge University Press

Sueyoshi, A., & Hardison, D. M. (2005). The role of gestures and facial cues in second language listening comprehension. *Language Learning*, 55(4), 661–699.

Sweller, J. (2010). Element interactivity and intrinsic, extraneous, and germane cognitive load. *Educational Psychology Review*, 22(2), 123–138.

Sweller, J., Van Merrienboer, J. J., & Paas, F. G. (1998). Cognitive architecture and instructional design. *Educational Psychology Review*, 10(3), 251–296.

Tellier, M. (2008). The effect of gestures on second language memorisation by young children. *Gesture*, 8(2), 219–235.

Vance, T. J. (1987). *An introduction to Japanese phonology*. State University of New York Press.

Vance, T. J. (2008). *The sounds of Japanese*. Cambridge University Press.

Van Lancker, D., & Fromkin, V. A. (1973). Hemispheric specialization for pitch and "tone": Evidence from Thai. *Journal of Phonetics*, 1(2), 101–109.

Wang, Y., Behne, D. M., Jongman, A., & Sereno, J. A. (2004). The role of linguistic experience in the hemispheric processing of lexical tone. *Applied Psycholinguistics*, 25(3), 449–466.

Xi, X., Li, P., Bailly, F., & Prieto, P. (2020). Hand gestures facilitate the acquisition of novel phonemic contrasts when they appropriately mimic target phonetic features. *Journal of Speech, Language, and Hearing Research*, 63(11), 3571–3585.

Yoshioka, K., & Kellerman, E. (2006). Gestural introduction of Ground reference in L2 narrative discourse. *International Review of Applied Linguistics in Language Teaching*, 44(2), 173–195.

Zatorre, R. J., Belin, P., & Penhune, V. B. (2002). Structure and function of auditory cortex: Music and speech. *Trends in Cognitive Sciences*, 6(1), 37–46.

Zhang, Y., Bailly, F., & Prieto, P. (2020). Hand-clapping to the rhythm of newly learned words improves L2 pronunciation: Evidence from training Chinese adolescents with French words. *Language Teaching Research*, 24(5), 666–689.

Zhen, A., Van Hedger, S., Heald, S., Goldin-Meadow, S., & Tian, X. (2019). Manual directional gestures facilitate cross-modal perceptual learning. *Cognition*, 187, 178–187.

Zheng, A., Hirata, Y., & Kelly, S. D. (2018). Exploring the effects of imitating hand gestures and head nods on L1 and L2 Mandarin tone production. *Journal of Speech, Language, and Hearing Research*, 61(9), 2179–2195.

https://doi.org/10.1017/langcog.2024.24 Published online by Cambridge University Press

Appendix 1

1.1. Trained words in Exp. 1 and Exp. 2.

(Morae are separated by hyphens. All but asterisked items were also used in testing)

Type 0 (LHHH)	**Type 2 (LHLL)**
o–mi–ya–ge*	no–ne–zu–mi
ni–wa–to–ri	me–gu–su–ri
ma–bo–ro–shi	hi–ma–wa–ri
ja–ga–i–mo	ku–da–mo–no
na–ga–gu–tsu	a–sa–ga–o*
Type 1 (HLLL)	**Type 3 (LHHL)**
ka–ma–ki–ri	ta–ma–ne–gi
na–no–ha–na	ka–na–zu–chi*
mo–mi–no–ki	no–mi–mo–no
de–si–be–ru*	shi–ba–ka–ri
ma–ma–ta–chi	no–ko–gi–ri

1.2. Novel words that did not appear in training but only appeared in testing in Exp. 1

(Morae are separated by hyphens.)

Type 0 (LHHH)	**Type 2 (LHLL)**
to–tsu–ge–ki	go–mi–ba–ko[9]
u–ki–gu–mo	ka–ke–ji–ku
ma–chi–na–mi	ma–go–ko–ro
ra–ku–ga–ki	a–ze–mi–chi
wa–ka–mo–no	na–re–zu–shi
Type 1 (HLLL)	**Type 3 (LHHL)**
ka–mi–sa–ma	ba–ke–mo–no
ta–te–yo–ko	hi–ru–go–ro
ku–zu–no–ha	ku–gu–ri–do
pa–bu–ro–hu	mo–chi–mo–no
a–se–mi–zu	a–ya–to–ri

Appendix 2

Subjective assessment questionnaire in Exp. 2

The names of the training methods were different when we conducted Experiment 2. The terms in this article correspond with the following terms in this questionnaire:

Present paper terms	=	Terms used in this questionnaire
Baseline	=	Flat notation
Notation	=	Spatial notation
L–gesture	=	Left–hand gesture
R–gesture	=	Right–hand gesture

[9]Variations exist for the pitch pattern of this word: LHLL, LHHL and LHHH (Kindaichi, 1996; Broadcasting Culture Research Institute, 1998 and 2016; Hirayama, 1960). We went with LHLL following Kindaichi (1996) and following the speaker's most comfortable pronunciation so as to avoid any ambiguity in their signals due to their unfamiliar pronunciations.

https://doi.org/10.1017/langcog.2024.24 Published online by Cambridge University Press

Subjective Assessment Questionnaire

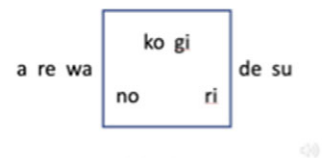
Please review the key below to remind yourself how we refer to each training condition you underwent. Then, answer the following questions about your perception of your experience in the different training conditions.

* Required

1. Subject number *

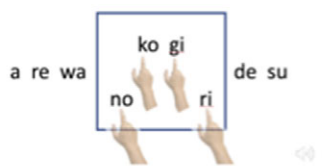
Flat notation

Spatial Notation



Left hand gesture

Right hand gesture



https://doi.org/10.1017/langcog.2024.24 Published online by Cambridge University Press

MENTAL EFFORT: Please answer the following question for training. Overall, this training condition made it _____ to perceive the pitch contrasts. 1= very easy, 2= easy, 3= neither easy or difficult, 4= difficult, 5=very difficult

	1	2	3	4	5	
very easy	○	○	○	○	○	very difficult

2. Chose the number that corresponds to your answer for each condition *

Mark only one oval per row.

	1	2	3	4	5
Flat notation	◯	◯	◯	◯	◯
Spatial notation	◯	◯	◯	◯	◯
Left hand gesture	◯	◯	◯	◯	◯
Right hand gesture	◯	◯	◯	◯	◯

3. ENJOYMENT: Please rank the trainings according to how much you enjoyed each one [1 = first choice (enjoyed most) , 4 = 4th choice (enjoyed least) *

Mark only one oval per row.

	1	2	3	4
Flat notation	◯	◯	◯	◯
Spatial notation	◯	◯	◯	◯
Left hand gesture	◯	◯	◯	◯
Right hand gesture	◯	◯	◯	◯

https://doi.org/10.1017/langcog.2024.24 Published online by Cambridge University Press

4. HOURS: If you were to participate in a follow up study on Japanese pitch accent learning, how many hours would you be willing to commit to each of these training conditions in one sitting? *

Mark only one oval per row.

	0.25	0.5	0.75	1	1.25	1.5	1.75	2	2.25
Flat notation	◯	◯	◯	◯	◯	◯	◯	◯	◯
Spatial notation	◯	◯	◯	◯	◯	◯	◯	◯	◯
Left hand gesture	◯	◯	◯	◯	◯	◯	◯	◯	◯
Right hand gesture	◯	◯	◯	◯	◯	◯	◯	◯	◯

5. DAYS: If you were to participate in a follow up study on Japanese pitch accent learning where you had to practice on your own, how many days per week would you want to practice?

Mark only one oval per row.

	1	2	3	4	5	6	7
Flat notation	◯	◯	◯	◯	◯	◯	◯
Spatial notation	◯	◯	◯	◯	◯	◯	◯
Left hand gesture	◯	◯	◯	◯	◯	◯	◯
Right hand gesture	◯	◯	◯	◯	◯	◯	◯

For the next 3 questions, you can choose the same answer if you like.

https://doi.org/10.1017/langcog.2024.24 Published online by Cambridge University Press

6. INTUITIVE: Which training condition did you find the most intuitive? *

Mark only one oval.

- ◯ Flat notation
- ◯ Spatial notation
- ◯ Left hand gesture
- ◯ Right hand gesture

7. HELPFULNESS: Which training condition did you find the most helpful? *

Mark only one oval.

- ◯ Flat notation
- ◯ Spatial notation
- ◯ Left hand gesture
- ◯ Right hand gesture

8. MOTIVATION: Which training condition made you feel most motivated to continue learning about Japanese pitch accent? *

Mark only one oval.

- ◯ Flat notation
- ◯ Spatial notation
- ◯ Left hand gesture
- ◯ Right hand gesture

https://doi.org/10.1017/langcog.2024.24 Published online by Cambridge University Press

Appendix 3

3.1. Exp. 1 fixed effect and variance estimates for logistic regression model of pitch pattern identification accuracy (observations = 9504)

Fixed effect	Coefficient	SE	z-Value	*p*
Intercept	−0.40	0.24	−1.68	0.093
Notation (vs baseline)	0.15	0.20	0.77	0.443
L–gesture (vs baseline)	0.04	0.20	0.18	0.858
Posttest (vs pretest)	0.52	0.12	4.44	<0.001***
Novel (vs trained)	0.10	0.28	0.35	0.726
Notation × posttest	−0.06	0.16	−0.39	0.696
L–gesture × posttest	−0.31	0.16	−1.87	0.061
Notation × novel	−0.24	0.16	−1.54	0.130
L–gesture × novel	−0.10	0.16	−0.61	0.543
Posttest × novel	−0.32	0.16	−2.01	0.044 *
Notation × posttest × novel	0.26	0.22	1.20	0.231
L–gesture × posttest × novel	0.47	0.22	2.11	0.035 *
Random effect				s^2
Participant				0.29
Word				0.76

3.2. Exp. 2 fixed effect and variance estimates for logistic regression model of pitch pattern identification accuracy (observations = 1920)

Fixed effect	Coefficient	SE	*z*-Value	*p*
Intercept	0.003	0.21	0.01	0.989
Notation (vs baseline)	−0.12	0.14	−0.86	0.390
L–gesture (vs baseline)	−0.19	0.14	−1.34	0.181
R–gesture (vs baseline)	−0.08	0.14	−0.60	0.550
Random effect				s^2
Participant				0.50
Word				0.47

3.3. Exp. 2 fixed effect and variance estimates for mixed effects model of EEG alpha power (observations = 172)

Fixed effect	Coefficient	SE	*df*	*t*-Value	*p*
Intercept	1.31	0.08	49.56	16.00	<0.001***
Notation (vs baseline)	0.04	0.04	126	1.18	0.239
L–gesture (vs baseline)	−0.02	0.04	126	−0.70	0.488
R–gesture (vs baseline)	0.03	0.04	126	0.87	0.388
Random effect					s^2
Participant					0.26

https://doi.org/10.1017/langcog.2024.24 Published online by Cambridge University Press

3.4. Exp. 2 fixed effect and variance estimates for mixed effects model of EEG theta power (observations = 140)

Fixed effect	Coefficient	SE	*df*	*t*-Value	*p*
Intercept	2.73	0.15	71.05	18.26	<0.001***
Notation (vs baseline)	−0.02	0.14	102	−0.11	0.915
L–gesture (vs baseline)	−0.06	0.14	102	−0.42	0.673
R–gesture (vs baseline)	−0.05	0.14	102	−0.33	0.741
Random effect					s^2
Participant					0.43

Cite this article: Hirata, Y., Friedman, E., Kaicher, C., & Kelly, S. D. (2024). Multimodal training on L2 Japanese pitch accent: learning outcomes, neural correlates and subjective assessments, *Language and Cognition* 16: 1718–1755. https://doi.org/10.1017/langcog.2024.24