LANGUAGE LEARNING
A Journal of Research in Language Studies

Language Learning ISSN 0023-8333

EMPIRICAL STUDY

Incidental and Multimodal High Variability Phonetic Training: Potential, Limits, and Future Directions

Kazuya Saito [a], Keiko Hanzawa,[b] Katya Petrova,[a] Magdalena Kachlicka [a], Yui Suzukida [a], and Adam Tierney[c]

[a]University College London, [b]Tokyo University of Science, and [c]Birkbeck, University of London

Abstract: Scholars have extensively investigated the effectiveness of high variability phonetic training (HVPT), that is, identification and discrimination of second language speech sounds produced by multiple speakers followed by trial-by-trial feedback. Building on the notion of incidental and multimodal learning in cognitive psychology (e.g., Lim & Holt, 2011), we developed a new, HVPT-based videogame paradigm in which participants aimed to shoot clay targets as fast as possible while being guided to learn sound cues as a by-product of planned learning. Focusing on the speech acquisition of 58 Japanese English-as-a-foreign-language learners, the current study set out to test the pedagogical potential and limits of the incidental HVPT approach. According to the results of statistical analyses, the effectiveness of incidental HVPT can be more clearly observed if it focuses on more learnable targets (e.g., acquisition of English [æ]–[ʌ] rather than [r]–[l] contrasts) with gains being more generalizable from trained to new speakers' voices and from perception to production dimensions.

We gratefully acknowledge insightful comments from anonymous *Language Learning* reviewers and Associate Editor Judit Kormos on earlier versions of the manuscript. We also thank Yuki Mori and Yoshiyuki Suzuki (ALC PRESS INC., JAPAN) for their assistance with the development of the mobile application for the purpose of the current project. This study was funded by a Leverhulme Trust Research Grant (RPG-2019-039), a Spencer Foundation Grant (202100074), and an ESRC Connection Grant (ES/S013024/1).

Correspondence concerning this article should be addressed to Kazuya Saito, University College London, Institute of Education, 20 Bedford Way, London, UK WC1H 0AL. Email: k.saito@ucl.ac.uk.

The handling editor for this article was Judit Kormos.

DOI: 10.1111/lang.12503

Keywords second language speech; high variability phonetic training; incidental learning; multimodal learning; videogaming

Introduction

To examine the mechanisms underlying postpubertal second language (L2) speech learning, scholars have investigated learners' behaviors when they receive various types of perception training in which the quantity, type, and timing of input are carefully controlled. Whereas different types of training methods (e.g., explicit vs. incidental; single-modal [language form only] vs. dual-modal [language and meaning]) have been examined and found to trigger different types of learning processes and outcomes in the fields of L2 grammar (Lyster & Saito, 2010) and L2 vocabulary (Uchihara, Webb, & Yanagisawa, 2019), L2 speech researchers have exclusively investigated a single-modal (language-focused), highly explicit training method, that is, high variability phonetic training (HVPT; Logan, Lively, & Pisoni, 1991). In this method, a learner is exposed to multiple voices during nonnative speech perception training. To develop stable, robust, and generalizable speech categories in a target language, learners are intensively exposed to new and/or partially acquired L2 sounds that are embedded in diverse phonetic, lexical, and speaker contexts. For each token, learners identify or discriminate target sounds in minimal pairs, with feedback presented after each trial.

When it comes to real-life L2 speech learning, however, L2 learners do not need only to attend to auditory but also to visuomotor aspects of sounds such as speakers' lip rounding, eye movement, and gestures (i.e., multimodal learning). If there is a mismatch between audio and visual cues, visual information can alter what an interlocutor hears (McGurk & MacDonald, 1976). Such speech perception learning likely takes place while L2 learners engage in conversations and multimedia exposure in which their primary focus lies in meaning conveyance (i.e., incidental learning) rather than in the phonetic analyses of speech sounds. To date, however, little has been known about how training can simulate the incidental and multimodal aspects of L2 speech learning and about how such training can facilitate the development of L2 speech perception and production (for meta-analyses, see Saito & Plonsky, 2019, and Sakai & Moorman, 2018).

Building on the notion of incidental learning in cognitive psychology (e.g., Lim & Holt, 2011), which aims to simulate the way auditory categories are learned in real-life settings, we developed incidental and multimodal HVPT. We asked 58 Japanese university students to play a videogame in which the goal was for them to shoot clay targets that flew across the screen from

various set points of origin. The more quickly that the students destroyed the clay targets, the more points they earned, prompting the students to attempt to anticipate the point of origin of each target as accurately as possible. This could be accomplished by making use of information from simultaneously presented speech sounds: different variants of target phonological contrasts (English consonants or vowels) that were presented before the release of each clay target. Because we did not give the participants any explicit instruction or feedback, they were thus forced to induce this strategy on their own through playing the game. By adopting a pretest and posttest design, we evaluated the potential and limits of the incidental HVPT approach in incidental L2 speech learning.

Background Literature

Psycholinguistic Model of Instructed Second Language Speech Learning

According to the psycholinguistic view of the development of L2 sound and word knowledge, learning can take place at multiple stages (see Saito, 2018, for an overview). Similar to L1 acquisition, learners first prioritize the acquisition of semantic information at the expense of linguistic accuracy when they encounter new words during L2 speech learning. Learners at this stage may know the meaning of L2 words but may be using L1 phonetic forms. As learners' L2 proficiency reaches a certain threshold for communicative success, they may move onto the stage of phonetic reattunement. For example, one such threshold was operationalized as vocabulary size of 6,000–7,000-word families that is believed to lead to adequate comprehension of various L2 oral discourses (Bundgaard-Nielsen, Best, & Tyler, 2011). At this level of proficiency, L2 learners progress to filling in phonetic details allowing them to perceive and produce more phonologically similar words such as minimal pairs. After ample input and output opportunities (e.g., conversations, multimedia exposure) with diverse interlocutors in various settings (e.g., home, work, and social), learners proceed to the automatization stage. At this level of proficiency, they can become increasingly capable of accessing L2 phonetic knowledge more accurately, promptly, and subconsciously regardless of talker conditions (e.g., males vs. females, familiar vs. unfamiliar voices) and task conditions (e.g., controlled vs. spontaneous; Saito, 2013a).

Following this line of thought, scholars have extensively examined how providing instruction can facilitate the different stages of L2 speech learning. It has been shown that learners can establish form-meaning connections of L2 words as a function of increased exposure to the target language (Trofimovich & Gatbonton, 2006). However, this receptive approach alone may not suffice to trigger phonetic-level attunement. Many L2 learners continue to use L1

phonological forms unless these interlanguage forms seriously hinder successful comprehension and communication (Major, 2001). Here, instruction and feedback play an important role in drawing learners' attention to the gaps in phonetic details between their current use of L2 words with L1 phonetic forms and the use of L2 words with target like L2 phonetic forms (Saito, 2013b). This noticing process is believed to lead to restructuring, updating, and refining the knowledge of the phonological form of words (Bundgaard-Nielsen et al., 2011). To help learners attend to the phonetic form rather than to the semantic aspects of words, many scholars have emphasized the importance of devising materials to enhance learners' awareness of sound-sized units of L2 speech. Such instruction can comprise explicit explanation of the acoustic properties and/or articulatory configurations of target sounds (Saito & Plonsky, 2019). To promote automatization, instruction can further help consolidate learners' partially acquired L2 speech knowledge. Through communicatively authentic (Saito, 2015) and systematically repetitive tasks (Suzuki & Hanzawa, 2021), learners can perceive and produce target forms provided by different talkers (Uchihara, Webb, Saito, & Trofimovich, 2021) and in different modalities (visual, audio, and kinesthetic; Tsunemoto, Lindberg, Trofimovich, & McDonough, 2021).

High Variability Phonetic Training

Over the past 30 years, ample evidence has been gathered that has shown that even postpubertal L2 learners can acquire new sounds when they receive HVPT (for comprehensive reviews, see Fraser, 2011; Sakai & Moorman, 2018; Thomson, 2018). From a theoretical standpoint, the effectiveness of the method has been ascribed to the variability of the input (Thomson, 2018). In this paradigm, participants engage in identification and discrimination tasks in which they choose which sounds they have heard from a small set of options followed by trial-by-trial feedback (e.g., *rock* vs. *lock*; see Shinohara & Iverson, 2018, for the relative effectiveness of identification vs. discrimination training). To expose learners to the high-variability nature of natural speech, target sounds are embedded in different lexical (e.g., *read* vs. *lead*) and/or phonetic contexts (e.g., *pray* vs. *play*) and produced by multiple speakers (e.g., males vs. females). For more methodological issues (e.g., talker variability, stimulus type), see Appendix S1 in the online Supporting Information.

The effectiveness of HVPT has been extensively examined in the context of Japanese learners' acquisition of English [r] and [l] and less so in the context of L2 English vowel acquisition. We have summarized in Tables 1 and 2 research that has shown that participants can demonstrate a moderate amount of

Table 1 Summary of five key high variability phonetic training (HVPT) studies on Japanese speakers' English [r] and [l] acquisition

Study	Participants	Hours of HVPT	Findings
Logan, Lively, & Pisoni (1991)	• 6 Japanese speakers in the United States (length of residence: 6 months to 3 years)	• 6 hours • 4,080 trials (272 trials × 15 sessions)	• 7.8% gains (78.1% → 85.9% for trained stimuli) • 83.7% (novel stimuli)
Lively, Logan, & Pisoni (1993)[a]	• 6 Japanese speakers in the United States (length of residence: 2 months)	• 6 hours • 4,080 trials (272 trials × 15 sessions)	• 5.6% gains (79.9% → 85.5% for trained stimuli) • 88% (novel stimuli)
Bradlow, Pisoni, Alahane-Yamada, & Tohkura (1997)	• 11 Japanese speakers in Japan (no experience overseas)	• 22.5 hours • 45 sessions • 12,240 trials (272 trials × 45 sessions)	• 16% gains (65% → 81% for trained stimuli) • 80–83% (novel stimuli)

(*Continued*)

Table 1 (Continued)

Study	Participants	Hours of HVPT	Findings
Iverson, Kuhl, Akahane-Yamada, Diesch, Tohkura, Kettermann, & Siebert (2003)	• 62 Japanese speakers (mixed English-as-a-foreign-language & English-as-a-second-language learners)	• 5 hours • 3,000 trials (300 trials × 10 sessions)	• 18% gains (50-60% → 70–80%)
Shinohara & Iverson (2018)	• 41 Japanese speakers (mixed English-as-a-foreign-language & English-as-a-second-language learners)	• 5 hours • 3,000 trials (300 trials × 10 sessions)	• 15% gains (70-80% → 90%)

[a] Experiment 1 from Lively et al. (1993).

Table 2 Summary of five key high variability phonetic training (HVPT) studies on L2 English vowel acquisition

Study	Participants	Hours of HVPT	Findings
Lambacher, Martens, Kakehi, Marasinghe, & Molholt (2005)	• 54 Japanese speakers (no experience overseas)	• 2 hours • 450 trials (75 trials × 6 sessions) • 5 vowels	• 18.6% gains (69.2% → 87.8%)
Lengeris & Hazan (2010)	• 28 Greek speakers (no experience overseas)	• 4 hours • 1125 trials (225 trials × 5 sessions) • 14 vowels	• 17.3% gains (48.6% → 65.68%) • 21.6% gains (48.6% → 78.2%)
Thomson (2012)	• 26 Chinese speakers in Canada (length of residence: 11.6 months)	• 2.5 hours • 1600 trials (200 trials × 8 sessions) • 10 vowels	• 10–20% gains

(*Continued*)

Table 2 (Continued)

Study	Participants	Hours of HVPT	Findings
Iverson, Pinet, & Evans (2012)	• 21 French speakers in France • 15 French speakers in the United Kingdom (length of residence: 18 months)	• 6 hours • 1800 trials (225 trials × 8 sessions) • 14 vowels	• 21% gains (40% → 61%) • 17% gains (59% → 76%)
Ortega, Mora, & Mora-Plaza (2019)	• 41 speakers of Catalan and Spanish in Spain	• 3 hours • 2,048 trials (512 trials × 4 sessions) • 2 vowels (æ, ʌ)	• Significant and generalizable production improvement

improvement in the context of English [r] and [l] (5–15% gains) and moderate-to-large amount of improvement in the context of L2 English vowel acquisition (15–20% gains). This relative difference in learning gains could be explained by the fact that the specific L1–L2 phonemic contrast (i.e., Japanese speakers' English [r] and [l] acquisition) is one of the most difficult instances in adult L2 speech acquisition (for more theoretical accounts, see the section Motivation for Current Study). In essence, the effectiveness of HVPT in both contexts generalizes to the target contrasts produced under novel lexical, phonetic, and speaker conditions (see all the primary studies in Tables 1 and 2). There is also some evidence that the effectiveness of HVPT can be robust and sustainable (see Bradlow, Akahane-Yamada, Pisoni, & Tohkura, 1999, for the results of the delayed posttests six months after the training).

Importantly, some scholars have argued that the effectiveness of HVPT has mainly been driven by the explicit nature of the training owing to the presence of trial-by-trial feedback (McClelland, Fiez, & McCandliss, 2002). During the treatment, participants are allowed to fully focus on one single task, that is, phonetic analyses of each stimulus, without allocating their cognitive resources to other tasks. Explicit HVPT may directly help learners become more aware of phonetic details (reattunement). Yet, explicit HVPT may not necessarily help access such knowledge at various levels of processing abilities (automatization). This is because the highly language-focused treatment deviates from the way language is used in real-life settings in which speakers encounter, access, and elaborate language via different modalities (auditory vs. visual vs. kinesthetic), levels of awareness (intentional vs. incidental), and degrees of engagement (single vs. dual tasks). Additionally, feedback in real-life settings can be indirectly provided via communication success as a form of positive evidence and via breakdowns as a form of negative evidence (Lyster & Saito, 2010).

Given the theoretical significance of HVPT, the highly explicit language-focused nature of the existing training method could be considered to be a drawback. In her review of HVPT, for example, Fraser (2011) referred to the method as "the deprecated 'drill and kill' training" (p. 12) because the mechanical repetition of simple language-focused tasks could be reminiscent of audio-lingual training methods. Decontextualized L2 training may not only be dissociated from the use and learning of language in real-life settings but also may negatively affect students' motivation, enjoyment, and attitude toward L2 learning—essential catalysts of successful and autonomous long-term L2 learning (Dewaele, Witney, Saito, & Dewaele, 2018).

Incidental and Multimodal Training

In response to the criticisms of the explicit HVPT method, the current study aimed to incorporate the line of research on incidental and multimodal speech learning proposed by Holt and her colleagues. Their work has highlighted both auditory and speech category learning (Gabay, Dick, Zevin, & Holt, 2015; Lim & Holt, 2011; Wade & Holt, 2005) and difficulties in acquiring new categories caused by language impairment (e.g., dyslexia; Gabay & Holt, 2015). Whereas intentional speech learning is characterized by overt and single focus on auditory category decisions due to explicit instruction and feedback (i.e., a single task condition), incidental speech learning involves visuomotor task completion as a primary goal without overt instruction and/or feedback on target phonological contrasts. For example, such main tasks include clicking on one of four boxes (e.g., the systematic multimodal associations reaction time task in Gabay et al., 2015) or on alien-looking creatures on a computer screen (e.g., the videogaming task in Lim & Holt, 2011).

In these tasks, participants attain rewards (more points) by clicking visuomotor targets as fast as possible. To assist in the successful completion of the main tasks, learners are induced to make use of sound cues as a by-product of planned learning. These stimuli are linked to a set of unique color, spatial, and audio cues. As the task progresses, other cues are gradually removed until only the sound cues remain available, so that players are guided to rely on them and strengthen the link between audio and visuomotor stimuli. As a function of increased experience with the task, learners can take action more swiftly, accurately, and automatically, which in turn leads to more successful outcomes in the game. Given that a dual focus of task completion is a key characteristic of incidental and multimodal training, learners can prioritize a main visuomotor task over a secondary language learning task or pay equal attention to working on both tasks. It should be noted that the notion of incidental and multimodal learning is different from implicit learning. Incidental and multimodal learning triggers some form of metalinguistic awareness during or after treatment, but implicit learning involves no awareness throughout (for more discussion on explicit, incidental, vs. implicit L2 speech learning, see Saito & Plonsky, 2019).

This incidental approach is hypothesized to facilitate L2 speech acquisition because it provides multimodal learning conditions. Multimodality is believed to play an important role in auditory categorization in real-life settings in which learners need to hear and see both audio and gestural aspects of speech categories while completing other nonlanguage tasks (Wade & Holt, 2005). In the incidental approach, there is no explicit instruction to attend to the auditory

stimuli. Yet, learners engage in goal-directed actions with a clear awareness of actions and rewards (e.g., earning points by clicking to shoot targets based on multisensory cues). This action-reward contingency has been found to be the source of successful outcomes in general learning behaviors (Tricomi, Delgado, & Fiez, 2004). In fact, there is evidence (e.g., Lim, Fiez, & Holt, 2014; Lim, Fiez, & Holt, 2019) that, while participants engage in incidental auditory category learning, the posterior striatum is functionally connected to the left posterior superior temporal cortex that mediates different aspects of reward processing and evaluation of outcomes. This suggests that participants actively engage in and take control over auditory categorization with positive emotion and motivation. On the other hand, the striatum is not activated during unsupervised, passive learning (Tricomi, Delgado, McCandliss, McClelland, & Fiez, 2006).

Research has demonstrated that incidental and intentional training could equally facilitate nonspeech auditory category learning (e.g., Seitz et al., 2010; Vlahou, Protopapas, & Seitz, 2012). In the context of artificial language learning, it has been shown that an incidental learning approach results in more gains than an intentional one, especially when target structures are nonsalient and difficult to explain with simple rules (Reber, 1989). Although incidental and multimodal training has occasionally been proposed, the findings have been mainly concerned with learning synthesized nonspeech sounds (e.g., Wade & Holt, 2005).

To our knowledge, Lim and Holt's (2011) study is the only study that has examined the role of incidental and multimodal learning in the context of L2 segmental learning.[1] A total of 13 Japanese residents in the United States moderately experienced with English as a L2 ($M_{\text{length of residence}} = 8.8$ months) received approximately two to three hours of incidental training through exposure to a video game requiring rapid responses to visuospatial information (15–30 minutes $\times$ 5 days). On a computer screen, their task was to capture two friendly aliens (via left mouse clicks) and destroy two enemy aliens (via right mouse clicks). Not only did each alien have a unique shape, color, and movement, but each alien also had a unique sound prompt (English [r] and [d] for the friendly ones; English [l] and [g] for the enemy ones). The sound samples preceded the appearance of the aliens. At the onset of the game, the aliens moved slowly so that participants could become familiar with the cues (shapes, color, movements, sounds). As the game progressed, however, the movements of the aliens sped up. As such, participants were induced to rely more on the sound cues which predicted alien type (friendly vs. enemy) and to keep up with the faster pace of the task. According to the results of pretests and posttests,

participants' English [r] and [l] accuracy improved by 8.5% (68.4 → 76.3%), although the group improvements did not reach statistical significance (p = .074). The findings were comparable to the amount of gains that Japanese speakers have typically demonstrated after receiving longer hours of explicit HVPT (e.g., see Logan et al., 1991, for 8% gains after six hours of training; see Table 1).

The Current Study

Though HVPT has been found to facilitate L2 speech learning (Sakai & Moorman, 2018), the mode of training thus far has been exclusively explicit, language-focused, and single-modal (Fraser, 2011). Given that there is emerging evidence that incidental and multimodal training facilitates speech learning (Lim & Holt, 2011), more research is strongly called for to further examine the potential and limitations of such training in adult L2 speech learning. Interfacing the proven method of HVPT and the notion of incidental and multimodal learning in Lim and Holt's study, we developed a new videogame that simulated clay target shooting.

Using 58 Japanese English-as-a-foreign-language speakers, the current study set out to test whether three hours of incidental training could impact their acquisition of the English [r] and [l] contrast and the English [æ] and [ʌ] contrast. The main task of the game was to shoot a clay target as soon as it appeared on the screen. The faster the participants shot the target, the more points they received. Although the participants were not explicitly informed about specific target-cue mappings, there were 16 clay targets that could be differentiated based on three types of cues. The targets varied in color—gold, yellow, purple, or red, followed one of the three unique movement patterns—curving left, straight, or curving right, and flew from one of three different cannons located at the bottom of the screen—left, mid, or right. Importantly, these clay targets were preceded by four different English sounds. Each sound was produced by four different speakers. As the game progressed, the visual cues disappeared, that is, all clay targets had the same colors, and the spatial (location) cues became randomized, that is, a clay target could emerge from any of the cannons. As such, the participants were inductively (without explicit instruction) guided to pay attention solely to the link between the sound cues and visuomotor actions (i.e., the movement pattern) to complete the task more promptly, accurately, and successfully.

To explore the acquisitional value of the incidental and multimodal version of HVPT in adult L2 speech learning, we chose four English sounds as the target sounds: [ra] and [la] for the consonant group and [hæ] and [hʌ] for the

vowel group. Not only do Japanese speakers have difficulty with the target contrasts, but the instances also differ in terms of the amount of inherent learning difficulty, that is, English [r] versus [l] is more difficult than English [æ] versus [ʌ].

According to the perceptual assimilation model-L2 (Best & Tyler, 2007), Japanese speakers' English [r]–[l] acquisition is a difficult instance (i.e., single category). The two English phones [r] and [l] are perceived as poor exemplars of the Japanese alveolar tap [ɾ] (Flege, Takagi, & Mann, 1995). The contrast differs acoustically in the formants F3 (1900–2000 Hz vs. 2400–2800 Hz) and F2 (1,200–1,400Hz vs. 1,600–2,000 Hz), as well as consonant duration (50–100 ms vs. 5–20 ms). Although native speakers use F3 as a primary cue to perceive the English [r] and [l] contrast, Japanese speakers mainly resort to F2 and duration to perceive the contrast (Ingvalson, McClelland, & Holt, 2011). This is because Japanese speakers typically rely on F2 and duration variation to differentiate vowel and to approximant contrasts (e.g., long-short, [ɾ]–[w]) and F3 variation is not fully used in their L1 phonetic system. Therefore, to acquire nativelike English [r]–[l] performance, Japanese speakers need to establish a new cue weighting strategy to attend to F3 variation (Iverson et al., 2003). However, acquiring a robust F3 representation is extremely difficult for Japanese L1 speakers because the acquisition of the F3 representation is resistant to the effects of short-term training (Ingvalson, Holt, & McClelland, 2011) and develops only after extensive immersion experience (e.g., 10+ years; Flege et al., 1995).

In contrast, Japanese speakers' acquisition of English [æ]–[ʌ] is a relatively easy instance (i.e., category goodness). Japanese speakers can assimilate the two English phones [æ] and [ʌ] to one L1 phone and one additional L2 phone without much difficulty. In other words, English [æ] can be perceived as a new sound that is sufficiently distinguishable from any neighboring L1 sounds (Japanese [e] and [a]), and English [ʌ] is merged to the Japanese central vowel [a] (Nishi, Strange, Akahane-Yamada, Kubo, & Trent-Brown, 2008). The contrasts differ in terms of F2 (1,200–1,400 Hz vs. 1,600–2,000 Hz; Deterding, 2006) and duration (80–100 ms as short vowels vs. 100–150 ms as long vowels; Crystal & House, 1988). Although neither [æ] nor [ʌ] is present in the L1 system, Japanese speakers use both cues for the distinctions of front–back vowels ([i] vs. [ɯ]) and short-long vowels ([a] vs. [a:]). Therefore, Japanese learners may not have much difficulty adjusting the relative weights of F2 and duration cues while creating a new category for English [æ] and mapping [ʌ] to the counterpart Japanese [a] (Nishi et al., 2008).

We formulated then the following research questions:

1. To what degree does incidental training improve Japanese speakers' perception of English [r] and [l]?
2. To what degree does incidental training improve Japanese speakers' perception of English [æ] and [ʌ]?
3. To what degree do gains in the perception domain transfer to the production domain?

As we pointed out above, few studies have examined the potential and limitations of incidental and multimodal L2 speech training (cf. Lim & Holt, 2011, as the only exception). Thus, the current investigation was exploratory rather than confirmatory, and we formulated no particular predictions. However, we expected that the size of training effectiveness, if any, could be larger for English [æ] and [ʌ] than for English [r] and [l] because we assumed the consonant contrast to be more difficult to learn than is the vowel contrast. We did not intend to examine the relative effectiveness of incidental versus explicit L2 speech training. Given that extensive research has already shown that explicit HVPT can significantly impact various areas of L2 speech perception development (e.g., Sakai & Moorman, 2018), the main objective of our investigation was to examine whether the novel training method of incidental and multimodal HVPT could help L2 learners reach similar gains as the explicit HVPT used in previous studies: 5–15% for English [r] and [l] and 15–20% for English vowels (see Tables 1 and 2).

Method

Participants

We initially recruited 70 Japanese English-as-a-foreign-language students at a university in Japan. They were freshman students majoring in engineering ($M_{age} = 19.2$ years, range: 18–21 years) who registered for a few hours of English lessons per week ($M = 2.5$ hr, range: 2–4 hr). They had previously spent five to 15 years learning English in a classroom setting in Japan ($M_{age\ of\ learning} = 10.1$ years, range: 3–16 years). Although most of the students had had no experience overseas, six reported having spent short stays abroad (approximately one month) in English-speaking countries. Based on the training log and exit questionnaire (for methodological details, see below), we eliminated 12 participants from the final analyses (< 20% attrition); for technological or personal reasons, they failed to complete pretest/posttests ($n = 5$) or training sessions ($n = 7$). We assigned the remaining 58 participants to two groups with different targets: (a) consonant group for English [r] and [l] contrast ($n = 33$) and (b) vowel group for English [æ] and [ʌ] ($n = 25$).

In comparison to previous studies ($N = 6$ to 62; see Tables 1 and 2), the sample size of our study ($N = 58$) was relatively large. Given that a previous meta-analysis had found the effect size of HVPT to be medium-to-large (Sakai & Moorman, 2018, for $d = 0.92$), we conducted a priori power analysis accordingly using G*Power (Faul, Erdfelder, Buchner, & Lang, 2009) with an estimation of a medium-to-large effect size ($f = .35$). The suggested number of participants for a research design with one between-subjects variable (group: consonant, vowel) and one within-subjects variable (time: pretests, posttests) to reach strong statistical power (.951) was 82. With the number of the participants whom we included in the final analyses ($N = 58$), the compromise power analysis generated slightly lower power of .910, which could be considered beyond the field-specific recommended threshold in instructed L2 acquisition research (i.e., .700–.800; Larson-Hall, 2015).

Design

Due to the COVID-19 pandemic, all the data collection took place online to ensure the health and safety of all the participants and researchers. On Day 1, all the participants attended a group orientation session with the second author via a video-conferencing tool. The participants were given a general overview of the project and received an instruction pack establishing the procedure for taking the pretests and posttests (how to log into the speech assessments), along with detailed training instructions (how to download, install, and play the videogame on their smartphone). To further strengthen the participants' understanding of the procedure and to ensure their completion of each task (pretest, training, posttest) in a timely manner, we assigned the individual participants to one of three personal tutors (the third, fourth, and fifth authors).

Once the group orientation had finished, the tutors individually contacted their tutees via email and started individual communication (20–25 emails per tutor). To check participants' technological setup—internet connection, sound system, and microphone, the tutors invited the participants to complete a demo version of the speech tasks (5 min). Upon confirmation that the participants' performance had recorded successfully, the tutors asked the participants to move on to the pretests (speech perception and production; Day 2) and then to proceed to the six sessions of incidental HVPT (Days 3–8).

Although the participants were allowed to play the game at their convenience, their assigned tutors automatically recorded and monitored their daily performance. We required the participants to send a screenshot of their final scores at the end of each session. On Day 10, one day after the final session of the training, we conducted posttests (speech perception and production) and

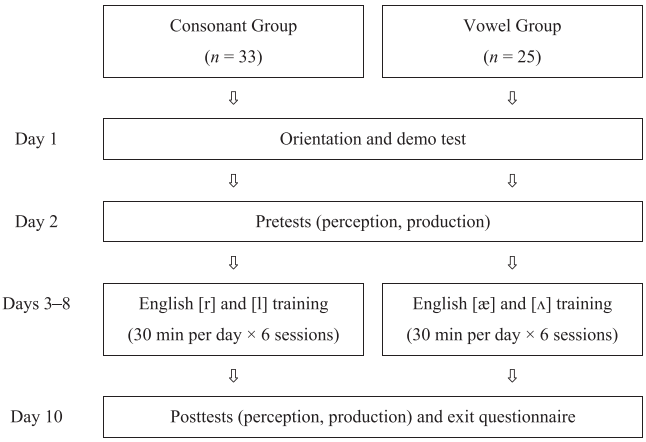

Figure 1 Summary of research design.

exit questionnaires in the same manner as we had done for the pretests. We conducted the demo session, pretests, and posttests through the Gorilla platform for online research (Anwyl-Irvine, Massonnié, Flitton, Kirkham, & Evershed, 2020). The participants received the training via a mobile phone application and engaged in the tests via their own computer.

Because we used the same test materials for pretests and posttests, we examined test-retest effects based on the English [æ] and [ʌ] performance of the consonant group who did not receive any training on the target vowels and the English [r] and [l] performance of the vowel group who did not receive any training on the target consonants.

We first analyzed the participants' perception test performance to see the extent to which the consonant and vowel groups enhanced their target of training. When there was a significant improvement in perception (with alpha set to .05), we further investigated the generalizability of perception gains to the production domains as a part of post hoc analyses. The exit questionnaire (Saito et al., 2022a) as well as the study materials (Saito et al., 2022b) are publicly available on IRIS (www.iris-database.org).

Incidental High Variability Phonetic Training

For the consonant group, the training comprised the experimental stimuli [rɒ] and [lɒ] as the target of instruction and the control stimuli [dɒ] and [gɒ] as distracters. With our focus on participants' performance on the experimental stimuli [rɒ] and [lɒ], the analyses mainly concerned how training helped

improve the participants' identification abilities of English [r] versus [l]. For the vowel group, we delivered the same format of treatment but with a focus on English [æ] and [ʌ]. The training comprised the experimental stimuli [hæ] and [hʌ] as the target of instruction and the control stimuli [hɪ] and [hɛ] as distracters. The main analyses focused on participants' performance on the experimental stimuli to examine how training helped enhance their ability to identify English [æ] and [ʌ].

Setup

The participants used their own smartphones (but not tablets) to download the training application available in Google Play and the Apple Store. Because we recorded all of the in-session gameplay in our in-house system via WiFi, we told the participants to play in a quiet room where they could access secure Internet connections. We assumed the quality of gameplay to be unrelated to the strength of the Internet connections because the application could run on both online and offline modes. To ensure high sound quality during gameplay, we also told the participants to use earphones, although we did not control for the type of earphone. Because we had assigned each participant to one of the research assistants as a tutor and had asked the participants to send a screenshot of their best daily score (see below), this allowed us to closely monitor participants' attendance and performance throughout the project. For more details about the experimental versus distracter stimuli, see Appendix S2 in the online Supporting Information.

Stimuli

According to recent HVPT research, greater improvement can be found when target sounds are embedded in nonword syllables rather than in real words because embedding in nonword syllables is believed to help learners attend to phonetic details of speech categories (Thomson & Derwing, 2016). Thus, we decided to use open syllables as training stimuli.

For the consonant group, four native speakers of British English (two males, M1 and M2, two females, F1 and F2) read "rock," "lock," "dot," and "got" in the carrier phrase, "the next word is ____." Following the procedure in Lim and Holt (2011), a researcher listened to each token carefully and excised the first open syllables, that is, [rɒ], [lɒ], [dɒ], and [gɒ], by putting a cursor on the beginning of stop closure [k] or [t] via Praat (Boersma & Weenink, 2019). For the vowel group, the same speakers read "hat," "hut," "hill," and "hell" using the carrier phrase, "the next word is ____." Using Praat, a researcher excised the first open syllables, that is, [hæ], [hʌ], [hɪ], and [hɛ], by putting a

cursor on the beginning of stop consonants [t] for "hat" and "hut" and on the beginning of antiresonance (the attenuation of formants) for "hill" and "hell." We normalized all the sound samples for amplitude and saved them as WAV files.

Training

Training involved high variability due to the presence of multiple talkers, multimodality due to the association of sounds with visual objects with unique movement, color, and spatial patterns, and incidental learning—task completion was based on visuomotor information as a primary goal with sound producing relevant supporting information.

The incidental HVPT training included six 30-minute sessions of videogame playing over six consecutive days. We asked the participants to shoot a clay target (by touching its location on the screen with their finger) as soon as it flew from one of the three cannons. We told them that the faster they shot clay targets, the more points they would earn.

We did not explicitly inform the participants about the number of the sounds or clay targets. There were 16 different clay targets representing four different sounds produced by four different speakers (M1, M2, F1, and F2). The sounds comprised [rɒ], [lɒ], [dɒ], and [gɒ] for the consonant training and [hæ], [hʌ], [hɪ], and [hɛ] for the vowel training. The clay targets could be differentiated based on visual cues: red, gold, yellow, or purple; spatial cues: flying from a left, central, or right cannon; and visuomotor cues: rightward, upward, or leftward trajectory (see Table 3 for a list of the different types of clay targets and available cues and Figure 2 for a screenshot of the gameplay).

For the consonant training, as in the precursor study by Lim and Holt (2011), targets with [rɒ] and [lɒ] were experimental tokens, and targets with [dɒ] and [gɒ] were control/distracter tokens. The participants' game performance ultimately depended on the extent to which they could accurately discriminate the contrast [rɒ] versus [lɒ] (not present in the L1 system). For the vowel training, targets with [hæ] and [hʌ] were experimental tokens, and targets with [hɪ] and [hɛ] were control/distracter tokens. Because we assumed that these Japanese speakers would not have difficulty perceiving English [hɪ] and English [hɛ], both of which are present in the L1 Japanese phonetic systems, the participants' ability to discriminate English [hæ] from [hʌ] determined their performance.

Each daily training session consisted of 30 one-minute rounds blocked into three levels (for a total of 30 min of training per day). We allowed the participants to take a short break upon their completing each level (i.e., after 10

Table 3 Summary of 16 different clay targets and their associated multimodal cues in consonant and vowel training

Cues	Clay targets 1–4[a]	Clay targets 5–8[a]	Clay targets 9–12[b]	Clay targets 13–16[b]
Consonant training				
Color	Red	Gold	Yellow	Purple
Cannons' location	Right	Center	Left	Left
Movements	Rightward	Upward	Leftward	Rightward
Preceding sound	English [rɒ]	English [lɒ]	English [dɒ]	English [gɒ]
Talkers	M1, M2, F1, & F2	M1, M2, F1, & F2	M1, M2, F1, & F2	M1, M2, F1, & F2
Vowel training				
Color	Red	Gold	Yellow	Purple
Cannons' location	Right	Center	Left	Left
Movements	Rightward	Upward	Leftward	Rightward
Preceding sound	English [hæ]	English [hʌ]	English [hɪ]	English [hɛ]
Talkers	M1, M2, F1, & F2	M1, M2, F1, & F2	M1, M2, F1, & F2	M1, M2, F1, & F2

Note. M1 = Male 1; M2 = Male 2; F1 = Female 1; F2 = Female 2.

[a] For experimental tokens: [rɒ, bɒ] in consonant training; [hæ, hʌ] in vowel training.

[b] For control/distracter tokens: [dɒ, gɒ] in consonant training; [hɪ, hɛ] in vowel training.

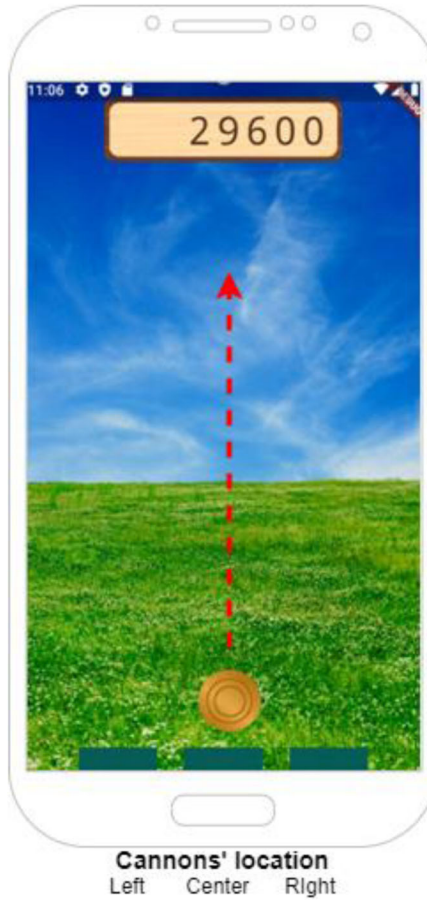

Figure 2 A screenshot of a single trial of the clay target shooting game. The speech category of the sound prompt (e.g., English [lɒ]) indicates a unique movement of the clay target before it appears on a screen (e.g., upward). The green rectangles at the bottom of the screen represent the cannons' location (left, center, and right), and the predicted movement trajectory (not visible to players) of a gold clay target is indicated via a red dashed line. [Color figure can be viewed at wileyonlinelibrary.com]

min). First a unique sound stimulus was played (1 s), and then a clay target was launched from one of three different canons at the bottom of the screen (1 s). All types of clay targets appeared for only one second, whether they moved rightward, leftward, or straightforward. Thus, we gave the participants a fixed amount of time to respond to the clay target regardless of the cannon location.

Table 4 Multimodality of one daily training session (30 rounds)

Level	Visual cues	Spatial cues	Visuomotor cues	Sound cues
Level 1 (1–10 rounds)	√	√	√	√
Level 2 (11–20 rounds)		√	√	√
Level 3 (21–30 rounds)			√	√

There were three different levels at each videogame session. At Level 1, all the visual (color), spatial (cannon location), visuomotor (clay target movements), and sound cues were available. At Level 2, color cues were removed, and all clay targets had identical colors. At Level 3, the remaining spatial cues were removed, so clay targets randomly flew from any of the cannons. Sound was the only cue that could aid the participants' prediction of the location and trajectory of the targets. As the game progressed (Levels 1 → 2 → 3), the participants were thus inductively guided to attend to the unique relationships between the sound and visuomotor cues (for the availability of the cues, see Table 4). As such, the participants became increasingly capable of predicting the movements of each clay target based on the target's preceding sound cues. It is unlikely that the participants had difficulty perceiving the control/distracter contrasts: [d]–[g] for consonant training; [ɪ]–[ɛ] for vowel training. What eventually determined the participants' success in the gameplay was their accurate and prompt identification of the experimental contrasts: [ɹ]–[l] for consonant training; [æ]–[ʌ] for vowel training.

Within each round, 16 different tokens were presented twice, resulting in a block length of 64 seconds (16 targets × 2 seconds × 2 utterances). Thus, the participants were exposed to 2,880 experimental exemplars: 16 targets ([ɹɒ, lɒ] for consonant training; [hæ, hʌ] for vowel training) × 30 rounds × 6 sessions. This number of encounters (2,880) was relatively small compared to those found in previous HVPT studies (e.g., 3,000–12,240 in Table 1). The length of the incidental training (3 hr) was comparable to that of Lim and Holt (2011; 2.5 hr).

To calculate the participants' scores, 120 frames were produced for each clay target in their one-second flight across the screen. Depending on how fast the participants hit the target, their scores for that target ranged from 0 points (failed to hit) to 120 points (hit within the first frame), resulting in a maximum possible score of 3,840 points (120 points × 32 clay targets) per round. Scores reset daily, though the participants' "High Score" was continually displayed at the top of their screen to motivate them to try to score ever higher.

Second Language Speech Perception Measures (Pretests and Posttests)

Stimuli

Given that we exposed the participants to nonword open syllables [rɒ] and [lɒ] for consonant training and [hæ] and [hʌ] for vowel training during the incidental HVPT training, we focused the analyses on the generalizability of the participants' gains to new untrained lexical items rather than to a new learning mode. We asked the participants to engage in a forced-choice identification test including 120 minimal pairs of monosyllabic words: 80 English [r]–[l] tokens and 40 English [æ]–[ʌ] tokens.

The [r]–[l] tokens included 20 English [r] and [l] minimal pairs produced by four native speakers of British English including two trained talkers (M1 and F1) and two untrained talkers (M3 and F3). The [æ]–[ʌ] tokens consisted of the 10 minimal pairs of English [æ] and [ʌ] produced by four different native speakers of English including two trained talkers (M1 and F1) and two untrained talkers (M3 and F3).

Word frequency has been found to influence L2 speech perception. According to Flege, Takagi, and Mann (1996), for example, Japanese speakers' speech perception was worse when English [r] and [l] phonemes appeared in less frequent words. To minimize the influence of lexical status on L2 speech perception, we carefully chose the target words from a list of the most frequent 3,000-word families according to the Coverage Calculator on the Compleat Lexical Tutor site (Cobb, 2020). The position of the vowels following the English [r]–[l] token was equally distributed in terms of height and backness. The consonants preceding the English [æ]–[ʌ] tokens were carefully matched for place of articulation: eight for labial, four for alveolar, and eight for velar consonants. Similarly, the following consonants were equally distributed: four voiceless stops, four voiced stops, and two nasals. For a list of target words, lexical status, and phonetic contexts, see Appendix S3 in the online Supporting Information. We recorded all the stimuli at a 40-kHz sampling rate and normalized them for peak intensity. We used the same materials for pretests and posttests.

Procedure

Each participant took the pretests and posttests individually using a unique ID. We instructed them to complete the tasks in a quiet room at home. A total of 120 stimuli played in a randomized order. Upon hearing a stimulus, the participants chose one of two minimal pair options. These options were presented via orthography (e.g., "read" vs. "lead"; "mad" vs. "mud"). Each trial was untimed. The entire task lasted 15 minutes. The results of Cronbach alpha

analyses showed that the participants' item-by-item performance demonstrated a high level of internal reliability at pretest ($\alpha = .92$) and at posttest ($\alpha = .91$).

Second Language Speech Production Measures (Pretests and Posttests)

Given that explicit HVPT (i.e., perception-based training) has also been found to help students transfer their gains from the perception to the production domain (Sakai & Moorman, 2018), we further examined the extent to which incidental HVPT impacted the participants' production skills. In the L2 speech literature, adult L2 speakers can demonstrate more nativelike pronunciation when tested via controlled tasks, such as word and sentence reading, than when tested via spontaneous speech tasks, such as picture description and interview (see, for example, Nagle, 2021). This could arguably be because the controlled task format allows speakers to carefully monitor their accurate production (Major, 2001). Thus, many scholars have emphasized the importance of measuring the robustness of L2 learners' abilities to access new sounds in more naturalistic, communicative, and dual-task settings (for a synthesis and meta-analysis, see Saito & Plonsky, 2019).

To tap into the participants' spontaneous, rather than controlled, production proficiency, following the procedure in Saito (2013a), we adopted a timed picture description. In this task, the participants described a series of pictures within a time limit (5 s for planning time and 30 s for picture description). To help the low-proficiency participants produce spontaneous speech of sufficient duration and to elicit target words, we gave them three-word prompts that they had to use.

We delivered the production task during the pretest and posttest sessions via the Gorilla online platform (Anwyl-Irvine et al., 2020). To keep the participants from excessively focusing on the target sounds, we asked them to undertake the production task first followed by the perception task. To facilitate the participants' understanding of the procedure, their tutor gave them instructions via email and a brochure detailing the task format. All the instructions were delivered in Japanese. Once they had logged in, the participants were asked to check their microphone settings by recording and listening to their own voice. Next, the participants recorded a practice trial to become familiar with the task procedure. Finally, they proceeded to the main task, that is, 20 picture descriptions.

We designed 10 stimuli to elicit the participants' English [r] and [l] production and another 10 stimuli to elicit their English [æ] and [ʌ] production (for a full list of target words, see Appendix S3 in the online Supporting

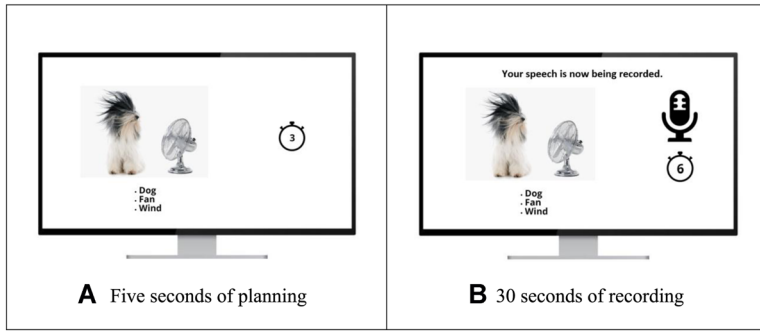

Figure 3 Screenshots of online timed picture description task. [Color figure can be viewed at wileyonlinelibrary.com]

Information). For each stimulus, one of the word prompts featured the same target words as the identification task, in which the target sounds English [r], [l], [æ], and [ʌ] appeared in word-initial positions, with preceding consonant and word frequency levels controlled for. In Figure 5, for example, "fan" is a target word. To indicate the amount of time left for planning (5 s; Figure 3, Panel A) and task completion (30 s; Figure 3, Panel B), a timer was displayed on the righthand corner of the screen throughout task completion.

Expert Rater Judgements

As we have detailed in the Results section, we found significant improvement only for the vowel group but not for the consonant group. Thus, we decided to investigate the transfer from the vowel group's perception performance to the production dimension of English [æ] and [ʌ]. Following the recommendations in Saito and Plonsky's (2019) framework of L2 speech analyses, we analyzed the quality of production data via expert raters' judgments. Given that the evaluation of the English [æ] and [ʌ] exemplars had already taken several hours, we asked the expert raters to focus on the analyses of only the vowel data. Although we had collected the participants' English [r] and [l] production, we did not submit them to rater analyses to avoid any unwanted effect of rater fatigue.

Building on the rating procedure for spontaneous L2 vowel production (Piske et al., 2011), two linguistically trained coders (L1 Japanese, near-nativelike L2 English proficiency) participated in the expert judgment sessions. In line with Flege et al.'s (1995) and Saito's (2013a) studies, we developed a 5-point scale to assess the quality of L2 pronunciation of English [æ] and [ʌ]:

- 0 points for L1 Japanese [a] substitutions,
- 1 point for the use of interlanguage forms,

- 2 points for probably English[æ]/[ʌ],
- 3 points for good English[æ]/[ʌ], and
- 4 points for nativelike English[æ]/[ʌ].

For more detailed justifications of the production analyses, see Appendix S4 in the online Supporting Information.

The coders first coded 10 out of 58 participants' English [æ] and [ʌ] productions (approximately 20% of the main dataset) to establish interrater reliability. The interrater agreement was relatively high, Cronbach $\alpha = .95$; ICC(3, 1) = .91 (consistency agreement). The first coder completed the rest of the analyses (48 participants), and we used her scores for the final analyses.

Finally, following the analysis procedure in Flege et al.'s (1995) study, we averaged the participants' performance for the 10 target words, deriving two pronunciation proficiency scores, one at pretest and another at posttest. This provided an index of change in participants' vowel production abilities in various phonetic contexts before and after training.

Exit Questionnaire

At the end of posttests, we requested that all the participants in the experimental group complete an exit questionnaire asking about their primary focus while playing the game. The participants' responses to the question of where they had placed their primary focus while playing the game varied. Whereas most of the participants reported having primarily attended to shooting clay targets ($n = 19$ for consonant; $n = 13$ for vowel), some reported that they had focused on the accurate identification of English [æ] and [ʌ] ($n = 2$ for consonant; $n = 5$ for vowel). Other participants reported that they had shared their attention between both gaming and phonological accuracy ($n = 12$ for consonant; $n = 7$ for vowel).

In our investigation, we considered the participants' retrospective and/or simultaneous focus on metalinguistic form as a characteristic of incidental and multimodal training. Thus, we speculated that the nature of the training could be considered incidental and multimodal for most of the participants (i.e., those who reported focusing on shooting and/or sharing their attention: 94% for the consonant group; 80% for the vowel group) and intentional for a minority of the students (6% for the consonant group; 20% for the vowel group). More precisely, we must acknowledge that a certain degree of explicit metalinguistic awareness was inevitably present throughout our training given that the participants may have been prompted to notice the phonological targets of instruction when they took the pretests. At the same time, however, it is important to point

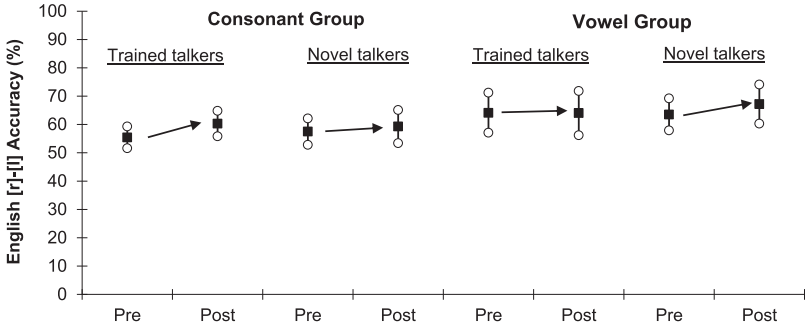

Figure 4 Mean accuracy scores (English [r] and [l]) and 95% confidence intervals of consonant and vowel groups at pretests and posttests.

out that (a) how to define and promote incidental learning for the timing and amount of awareness (for a more general notion of incidental L2 learning, see Uchihara et al., 2019) and (b) how to define and elicit learners' awareness (for a series of awareness measures, see Saito, 2019) have remained controversial (see Rebuschat, 2013). However, the results here confirmed the multimodal nature of the training because most of the participants reported their focus went beyond language-focused tasks.

Results

As Figure 1 illustrates, we used the English [r] and [l] pretest and posttest performance of the vowel group, who had not received any training on English [r] and [l], as a baseline to evaluate the consonant group's improvement patterns. Similarly, we used the consonant group's English [æ] and [ʌ] performance as a baseline for the analyses of the vowel group's improvement between pretests and posttests.

English [r] and [l] Analyses

Because five participants (one from consonant group, four from vowel group) reached 90% accuracy on the pretests, we eliminated them from the analyses (for a similar decision, see Iverson et al., 2003). Figure 4 visually displays the effectiveness of incidental HVPT on participants' English [r] and [l] proficiency, and Appendix S5 in the online Supporting Information summarizes the effectiveness of incidental HVPT on their English [r] and [l] proficiency. The descriptive results suggested that the consonant training group demonstrated about 5% gains in the trained talker condition.

Table 5 Summary of mixed effects modeling analyses of incidental high variability phonetic training on English [r] and [l]: Group gains

Fixed effects	*b*	*SE*	95% CI	*z*	*p*
Intercepts	0.697	0.170	[0.363, 1.030]	4.094	< .001
Time	0.006	0.109	[–0.208, 0.220]	0.055	.956
Group	–0.437	0.191	[–0.813, –0.061]	–2.283	.022
Talker	–0.056	0.157	[–0.365, 0.253]	–0.357	.721
Time × Group	0.219	0.138	[–0.053, 0.491]	1.586	.112
Group × Talker	0.133	0.136	[–0.135, 0.401]	0.978	.328
Time × Talker	0.175	0.153	[–0.126, 0.476]	1.140	.254
Time × Group × Talker	–0.320	0.194	[–0.701, 0.061]	–1.848	.079
Random effects	Variance	*SD*	R^2		
Participants	0.344	0.586	.163		
Items	0.262	0.512			

To examine the extent to which incidental training could impact the Japanese speakers' English [r] and [l] perception when we controlled for the random effects of their performance per item and per participant, we performed a mixed-effects binomial logistic regression analysis using the lmer function from the lme package (Version 1.1-21; Bates, Maechler, Bolker, & Walker, 2015) in R (R Core Team, 2018). We used the participants' binary accuracy scores for each stimulus (0 for incorrect, 1 for correct) as a dependent variable. We variable-coded the fixed effects in terms of group (1 for vowel, 2 for consonant), time (1 for pretests, 2 for posttests), and talker (1 for familiar voices used in training [M1 and F1] vs. 2 for untrained talkers [M3 and F3]). Random effects comprised participant ID (1–53) and stimulus ID (1–80). Due to the small sample size, we included random intercepts but not slopes. We used the following R command:

$$\begin{aligned}\text{MODEL} \neg \text{glmer}\ (&\text{Response} \sim \text{Time} + \text{Group} + \text{Talker} + \text{Time} : \text{Group} \\ &+ \text{Group} : \text{Talker} + \text{Time} : \text{Talker} + \text{Time} : \text{Group} : \text{Talker} \\ &+ (1|\text{Participants}) + (1|\text{Item}), \text{data} = \text{data}, \text{family} = \text{binomial})\end{aligned}$$

As Table 5 shows, the model yielded a significant main effect of group, indicating that the consonant group's English [r] and [l] perception accuracy ratio was significantly lower than that of the vowel group. We performed a set of post hoc multiple comparison analyses on the participants' logit scores (i.e., the

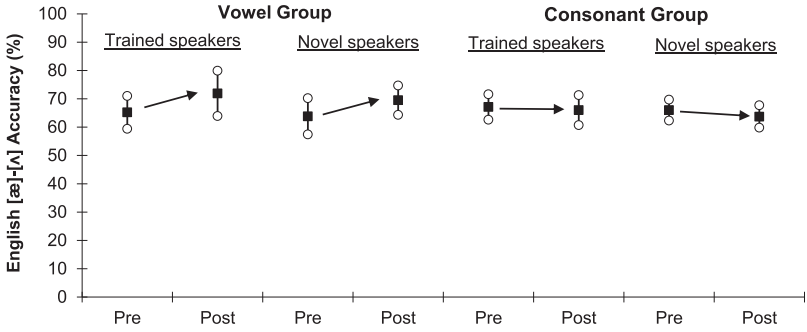

Figure 5 Mean accuracy scores and 95% confidence intervals of vowel and consonant groups at pretests and posttests.

probability of getting correct responses coded as 0 for incorrect and 1 for correct) by group using the estimated marginal means function from the emmeans package (Lenth et al., 2021) in R. The results showed that the vowel group generally demonstrated a significantly higher accuracy ratio ($M_{\text{log odds ratio}} = 0.716$, $SE = 0.146$, 95% CI [0.431, 1.002]) than did the consonant group ($M_{\text{log odds ratio}} = 0.375$, $SE = 0.122$, 95% CI [0.136, 0.615]) at a $p < .05$, $b = 0.341$, $SE = 0.172$, 95% CI [0.051, 0.672], $z = 1.982$, $p = .047$. However, none of the other main and interaction effects of time reached statistical significance.

English [æ] and [ʌ] Analyses

Because four participants ($n = 4$ from comparison) reached 90% accuracy at the pretest, we eliminated them from the analyses below. Figure 5 visually displays the descriptive statistics of participants' accuracy scores, and Appendix S5 in the online Supporting Information summarizes their descriptive statistics. The descriptive results showed that the vowel training group achieved 5–10% gains regardless of the talker conditions.

To investigate the effectiveness of the incidental training on English [æ] and [ʌ] perception while accounting for response variability across items and participants, we performed a mixed-effects binomial logistic regression analysis as in the English [r] versus [l] analyses. We constructed the model with participants' binary accuracy scores for each stimulus (0 for incorrect, 1 for correct) as dependent variables with group (vowel vs. consonant), time (pretest vs. posttest), and talker (trained vs. untrained talkers) as fixed effects predictors. We entered participant ID (1–54) and stimulus ID (1–40) as random effects. We again used the aforementioned R command. As Table 6 shows, the model identified a significant main effect of time and a significant two-way Time ×

Table 6 Summary of mixed effects modeling analyses of incidental high variability phonetic training: Group gains

Fixed effects	*b*	*SE*	95% CI	*z*	*p*
Intercepts	0.714	0.240	[0.243, 1.185]	2.973	.002
Time	0.351	0.158	[0.041, 0.662]	2.222	.026
Group	0.090	0.179	[–0.262, 0.443]	0.504	.614
Talker	0.005	0.317	[–0.617, 0.628]	0.018	.985
Time × Group	–0.405	0.200	[–0.800, -0.012]	–2.019	.043
Group × Talker	0.028	0.201	[–0.366, 0.424]	0.143	.886
Time × Talker	–0.032	0.227	[–0.478, 0.413]	–0.142	.887
Time × Group × Talker	–0.037	0.281	[–0.603, 0.528]	–0.130	.896
Random effects	Variance	*SD*	R^2		
Participants	0.1637	0.4046	.222		
Items	0.7588	0.8711			

Group interaction effect. However, a three-way Group × Time × Talker interaction effect did not reach statistical significance. The results suggested that the experimental and comparison groups may have differentially improved their L2 identification abilities over time regardless of talker conditions.

To further examine the significant Group × Time interaction effect, we performed follow-up multiple comparisons of participants' Group × Time logit scores on the model using the estimated marginal means function. We set an alpha level of .025 for these comparisons (i.e., Bonferroni corrected for two pairwise comparisons). The results showed that the probability of the accuracy ratio in the vowel group significantly improved from pretest ($M_{\text{log odds ratio}} = 0.717$, $SE = 0.182$, 95% CI [0.361, 1.070]) to posttest ($M_{\text{log odds ratio}} = 1.053$, $SE = 0.184$, 95% CI [0.693, 1.410]), $b = –0.336$, $SE = 0.113$, 95% CI [–0.664, –0.008], $z = –2.961$, $p = .003$. Yet, the gain scores of the consonant group ($M_{\text{log odds ratio}} = 0.821/0.734$, $SE = 0.168/0.167$, 95% CI [0.493/0.406, 1.150/1.060]) did not reach statistical significance, $b = 0.985$, $SE = 0.088$, 95% CI [,], $z = 0.985$, $p = .324$. Overall, the findings suggested that three hours of incidental HVPT could facilitate the participants' ability to identify English [æ] and [ʌ] and help generalize their improvement to real words across different speakers (6.2% gains).

Transfer From Perception to Production

According to the findings of the perception data, whereas the impact of incidental HVPT on English [r] and [l] acquisition was limited, we observed

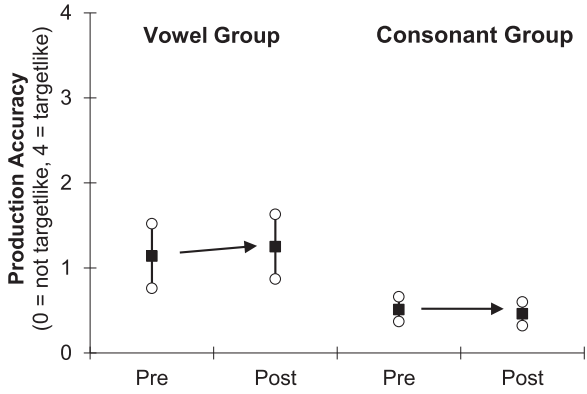

Figure 6 Mean accuracy scores (0 = not target-like, 4 = target-like) and 95% confidence intervals of vowel and consonant groups at pretests and posttests.

learning in both trained and untrained talkers in English [æ] and [ʌ] acquisition. Because the effectiveness of HVPT on the perception of English [r] and [l] appeared to be limited, we did not further pursue its transfer to the production domain.

Figure 6 visually displays the descriptive statistics of participants' averaged L2 vowel production scores, and Appendix S5 in the online Supporting Information summarizes these descriptive statistics. According to the results of Kolmogorov-Smirnov normality tests, the vowel and consonant groups' scores did not significantly differ from the normal distribution for the pretests, Ds = .183 and .195, ps = .189 and .260, and the posttests, Ds = .159 and .162, ps = .336 and .474. We submitted these production scores to a two-way ANOVA with group (vowel, consonant) as a between-subjects variable and time (pretest, posttest) as a within-subjects variable. We calculated the effect size using partial eta squared and interpreted it using Cohen's (1988) benchmarks: η_p^2 = .01 for small, .06 for medium, and .14 for large. The analysis of variance confirmed that there was a significant main effect for group, $F(1, 56) = 16.253$, $p = < .001$, $\eta_p^2 = .225$, and a significant Group × Time interaction effect, $F(1, 56) = 8.357$, $p = .005$, $\eta_p^2 = .130$. The main effects of time did not reach statistical significance, $F(1, 56) = 0.935$, $p = .338$, $\eta_p^2 = .016$. To further explore the significant Time × Group interaction effect, we performed post hoc multiple comparison analyses with the alpha level set to .025 (Bonferroni corrected). The results demonstrated that the vowel group significantly enhanced the target-likeness of their L2 vowel production proficiency with

medium-to-large effects, $F(1, 56) = 6.539$, $p = .013$, $\eta_p^2 = .105$. In contrast, the comparison group' performance did not demonstrate any significant change, $F(1, 56) = 2.147$, $p = .148$, $\eta_p^2 = .037$.

Discussion

In the field of L2 speech research, explicit, language-focused HVPT is a well-researched and proven method of L2 speech training (Sakai & Moorman, 2018) because it features a key aspect of L2 speech learning—intensive exposure to multiple voices. In cognitive psychology, the incidental training approach has been devised to simulate how auditory categories develop in real-life settings where language is acquired as a by-product of everyday communication without explicit instruction or feedback (Wade & Holt, 2005). Whereas the value of the incidental approach has been tested mainly in nonverbal sound learning paradigms, Lim and Holt's (2011) exploratory work showed some evidence that short-term incidental training can impact adult L2 speech learning to some degree. Extending this line of thought, our study developed an incidental HVPT training method and evaluated its potential and limits for 58 Japanese university-level students' acquisition of English [r] and [l], English [æ] and [ʌ], and the transferability of the training to production. While participants spent three hours engaged in a clay target shooting videogame on their smartphones, they were induced to make use of target English contrasts produced by multiple talkers to enhance their success in the game.

Overall, our findings echoed those of Lim and Holt (2011) in that the size of improvement on perception resulting from incidental HVPT was rather limited—5% gains for English [r] and [l]; 5–10% gains for English [æ] and [ʌ]—relative to the effectiveness of explicit HVPT—5–15% for English [r] and [l]; 15–20% for English vowels (see Tables 1 and 2). This could be ascribed to the fundamental nature of training. Explicit HVPT training enabled the participants to focus fully on the accurate identification and discrimination of the target L2 contrast—English [r] and [l], that is, single task conditions—while they received immediate explicit feedback. There is some evidence that provision of explicit error correction (rather than high variability) determines the effectiveness of L2 speech training (e.g., McClelland et al., 2002). Thus, it is possible that the moderate effects of incidental HVPT could be tied to the indirect, brief, and unsupervised nature of the training and that our findings relating to L2 phonology could mirror those in L2 grammar (Lyster & Saito, 2010) and L2 vocabulary (Uchihara et al., 2019) that demonstrated the superiority of explicit training over incidental learning.

Our study demonstrates that the instructional benefits of the incidental approach can vary across speech sound contrasts. Although the impact of incidental HVPT on Japanese speakers' English [r]–[l] acquisition remained unclear, the training significantly facilitated their English [æ]–[ʌ] development. One source of such variation could be related to the differential amount of learning difficulty—English [r]–[l] being more difficult than English [æ]–[ʌ]. For Japanese learners, the attainment of advanced L2 performance on English [r] and [l] appears to be extremely difficult. This involves changes in their reliance on perceptual cues, whereby cues that are salient in the L1 (F2, duration) must be downweighted, but acoustic cues that are normally underattended in the L1 must be upweighted for successful categorization of the contrast (F3 variation; Iverson et al., 2003). In such a relatively difficult L2 speech instance (i.e., single category; Best & Tyler, 2007), our findings showed that the effects of incidental HVPT were limited. In the case of the English [æ]–[ʌ] contrast, Japanese speakers exhibited more substantial learning, and their gains generalized not only to the perception of untrained talkers but also to production. This performance could be due to perceptual adjustments and the creation of new categories using cues already existing in Japanese speakers' vowel perception repertoire (F2, duration; Nishi et al., 2008). In this relatively easy L2 speech instance (i.e., category goodness), the effects of HVPT could be more robust if learners receive focused training with awareness of the target structures (Lambacher, Martens, Kakehi, Marasinghe, & Molholt, 2005).[2]

Limitations and Future Directions

Given that we took the first step toward examining the potential and limitations of the incidental HVPT approach, future work should be directed toward the conceptualization, development, and refinement of optimal L2 speech training methods along the following proposed lines of research. First, our findings should be replicated with more participants with diverse levels of L2 proficiency (e.g., Huensch & Nagle, 2021, showing that speaker proficiency affects pronunciation) and immersion experience (e.g., Iverson, Pinet, & Evans, 2012, for classroom vs. immersion learners), and targeting different L1–L2 pairings (e.g., Wang, Jongman, & Sereno, 2003, for L1 English speakers' acquisition of L2 Chinese tone). As such, future studies will allow researchers to confirm the validity and generalizability of the findings with strong statistical power in various contexts of L2 speech learning.

Importantly, previous studies on approaches to teaching L2 lexicogrammar have suggested that the effectiveness of incidental training can be observed under certain methodological conditions. For example, whereas the

effectiveness of short-term incidental training remains unclear, learning gains can be observed especially when learners receive an extensive amount of incidental training (e.g., see Uchihara et al., 2019, for a meta-analysis of incidental vocabulary acquisition through extensive reading) and when such gains can be observed in delayed rather than immediate posttests (Li, 2010). Future comparison studies should test training retention after providing different intensity and lengths of training (e.g., see Iverson et al., 2003, for 5 hr vs. Bradlow, Pisoni, Akahane-Yamada, Tohkura, 1997, for 30 hr) and assessing its effectiveness at multiple test timings (e.g., see Bradlow et al., 1999, for 6-month delayed tests). Such studies will shed light on the extent to which explicit and/or incidental HVPT can facilitate L2 speech acquisition in the long run.

Furthermore, the timing of incidental HVPT may need to be reconsidered in relation to participants' proficiency levels. The participants in this study were English-as-a-foreign-language students in Japan with limited conversation and immersion experience. According to existing literature in instructed L2 speech acquisition, there is some evidence showing that inexperienced learners, such as those in our study, are likely to have difficulty in noticing, encoding, and integrating the perceptual characteristics of new sounds when these sounds are embedded within communicatively-oriented instruction (e.g., see Saito, 2015, for phonological recasts and inexperienced L2 speakers; cf. Saito & Lyster, 2012).

In the context of word priming experiments, few learners have been found to access the phonological properties of new words accurately and fluently while engaging and prioritizing meaning in the target language (e.g., Trofimovich & Gatbonton, 2006). Comparatively, it has been revealed that more experienced L2 learners who are likely to have more advanced L2 phonetic representations benefit more from meaning-oriented approaches because they may simultaneously impact various dimensions of L2 speech proficiency (see Lee & Lyster, 2016, for speech perception; Saito, Suzukida, Oyama, & Akiyama, 2021, for speech production). Therefore, one promising future direction is to test the combination of explicit and incidental HVPT to promote effective and efficient learning of L2. At the initial stage of L2 speech learning, inexperienced learners may first need to be equipped with some form of L2 phonetic knowledge via explicit training and feedback. To this end, explicit HVPT could play a critical role in helping learners to notice and to attend to the perceptual characteristics of target sounds. Once auditory representations have been partially established, learners might then be more likely to be able to simultaneously engage in incidental and multimodal HVPT in a complementary manner so that they can practice new sounds under dual-task conditions to

proceduralize and automatize their newly developed phonetic knowledge (for a skill acquisition account of instructed L2 speech learning, see Saito & Plonsky, 2019).

Finally, we would like to stress that the main objective of our investigation was to conduct a first exploration of the potential and limits of incidental and multimodal L2 speech training. Given that the existing literature on explicit HVPT has shown that the training led to 5–20% gains (shown in Tables 1 and 2), the study set out to examine whether incidental HVPT can lead to comparable improvement. However, we did not design the study to make a direct comparison of different types of training within the same study. As for the relative effectiveness of incidental versus explicit training, although we call for such type-of-training research, much caution needs to be exercised. One major obstacle is to overcome a range of methodological differences during material development. After we conducted the current study, we found that the length of incidental HVPT could be considerably shorter than that of explicit HVPT. This is because trial-by-trial feedback was unnecessary in the incidental training but would be mandatory in explicit training. As a result, learners could receive a larger amount of input during a 30-min session (e.g., 480 trials for incidental HVPT in the current study vs. 200 trials for explicit HVPT in the previous studies in Tables 1 and 2). If researchers design a comparison study, it is crucial to control for treatment length and intensity because these variables differ between the explicit and incidental training paradigms, potentially resulting in different amounts of learning.

Conclusion

In our investigation, we proposed incidental and multimodal HVPT to provide L2 learners with opportunities to improve their perception of target sounds by working in both the auditory and visuospatial domains while their primary attention was directed to playing a clay target shooting game. In light of the psycholinguistic model of instructed L2 speech learning (as adopted by, for example, Saito & Plonsky, 2019), this approach is believed to have promoted not only phonetic reattunement but also seemed to result in generalizable gains thanks to its provision of multimodally enriched learning opportunities. At this level of proficiency, learners' improvement was relatively stable even when their performance was tested across different talkers (trained vs. untrained) and task dimensions (perception vs. spontaneous production). Our study indicated both the potential and limitations of the incidental and multimodal HVPT. On the one hand, the results of pretests and posttests demonstrated that incidental HVPT could facilitate L2 speech acquisition at various processing levels

(perception to spontaneous production). On the other hand, we observed such learning gains only when the treatment focused on the relatively easy aspects of L2 speech acquisition (e.g., English [æ]–[ʌ]), suggesting that more elaborate strategies may be needed for the relatively difficult aspects of L2 speech acquisition (e.g., English [r]–[l]).

Final revised version accepted 5 March 2022

Open Research Badges

This article has earned an Open Materials badge for making publicly available the components of the research methods needed to reproduce the reported procedure. All materials that the authors have used and have the right to share are available at http://www.iris-database.org. All proprietary materials have been precisely identified in the manuscript.

Notes

1 Some scholars have examined the use of videogaming as an implicit speech learning interface although their focus lay in assessment rather than training (e.g., Duran, Lewandowski, & Schweitzer, 2016).

2 A reviewer pointed out that another potential source of variation could be related to the inherent auditory salience of the target sounds. Given that vowels can be considered more salient than consonants (Cutler, Sebastián-Gallés, Soler-Vilageliu, & Van Ooijen, 2000), the participants may have had less difficulty noticing and integrating target sounds when they were vowels rather than consonants and when they were embedded in incidental and multimodal rather than language-focused training.

References

Anwyl-Irvine, A. L., Massonnié, J., Flitton, A., Kirkham, N., & Evershed, J. (2020). Gorilla in our midst: An online behavioral experiment builder. *Behaviour Research Methods*, *52*, 388–407. https://doi.org/10.3758/s13428-019-01237-x

Bates, D., Maechler, M., Bolker, B., & Walker, S. C. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Best, C. T., & Tyler, M. D. (2007). Nonnative and second-language speech perception. In O. S. Bohn & M. J. Munro (Eds.), *Language experience in second language speech learning: In honor of James Emil Flege* (pp. 13–34). Amsterdam, The Netherlands: John Benjamins.

Boersma, P., & Weenink, D. (2019). Praat: Doing phonetics by computer (Version 6.0.46). Retrieved from http://www.praat.org/

Bradlow, A. R., Pisoni, D. B., Akahane-Yamada, R., & Tohkura, Y. (1997). Training Japanese listeners to identify English /r/and /l/: IV. Some effects of perceptual learning on speech production. *The Journal of the Acoustical Society of America*, *101*, 2299–2310. https://doi.org/10.1121/1.418276

Bradlow, A. R., Akahane-Yamada, R., Pisoni, D. B., & Tohkura, Y. I. (1999). Training Japanese listeners to identify English/r/and/l/: Long-term retention of learning in perception and production. *Perception & Psychophysics*, *61*, 977–985. https://doi.org/10.3758/BF03206911

Bundgaard-Nielsen, R. L., Best, C. T., & Tyler, M. D. (2011). Vocabulary size is associated with second-language vowel perception performance in adult learners. *Studies in Second Language Acquisition*, *33*, 433–461. http://doi.org/10.1017/s0272263111000040

Cobb, T. (2020). Coverage Calculator (Version 1.2) [Computer program]. Retrieved from https://www.lextutor.ca/cover/

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.

Crystal, T. H., & House, A. S. (1988). The duration of American-English vowels: An overview. *Journal of Phonetics*, *16*, 263–284. https://doi.org/10.1016/s0095-4470(19)30500-5

Cutler, A., Sebastián-Gallés, N., Soler-Vilageliu, O., & Van Ooijen, B. (2000). Constraints of vowels and consonants on lexical selection: Cross-linguistic comparisons. *Memory & Cognition*, *28*, 746–755. https://doi.org/10.3758/BF03198409

Deterding, D. (2006). The pronunciation of English by speakers from China. *A Journal of Varieties of English World-Wide*, *27*, 175–198. https://doi.org/10.1075/eww.27.2.04det

Dewaele, J. M., Witney, J., Saito, K., & Dewaele, L. (2018). Foreign language enjoyment and anxiety: The effect of teacher and learner variables. *Language Teaching Research*, *22*, 676–697. https://doi.org/10.1177/1362168817692161

Doughty, C. J. (2019). Cognitive language aptitude. *Language Learning*, *69*, 101–126. https://doi.org/10.1111/lang.12322

Duran, D. Lewandowski, N., & Schweitzer, A. (2016). A 3D computer game for testing perception of acoustic detail in speech. *Proceedings of Meetings on Acoustics 22ICA*, *28*, 060004. https://doi.org/10.1121/2.0000493

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, *41*, 1149–1160. https://doi.org/10.3758/BRM.41.4.1149

Flege, J. E., Takagi, N., & Mann, V. (1995). Japanese adults can learn to produce English /ɹ/and /l/ accurately. *Language and Speech*, *38*, 25–55. https://doi.org/10.1177/002383099503800102

Flege, J. E., Takagi, N., & Mann, V. (1996). Lexical familiarity and English-language experience affect Japanese adults' perception of /ɹ/and /l/. *The Journal of the Acoustical Society of America*, *99*, 1161–1173. https://doi.org/10.1121/1.414884

Fraser, H. (2011). Teaching teachers to teach /r/and /l/ to Japanese learners of English : An integrated approach. In *Proceedings of PTLC 2011* (pp. 11–15). Retrieved from https://www.phon.ucl.ac.uk/ptlc/ptlc-2011-proceedings/ptlc2011_fraser-002.pdf

Gabay, Y., Dick, F. K., Zevin, J. D., & Holt, L. L. (2015). Incidental auditory category learning. *Journal of Experimental Psychology: Human Perception and Performance*, *41*, 1124–1138. https://doi.org/10.1037/xhp0000073

Gabay, Y., & Holt, L. L. (2015). Incidental learning of sound categories is impaired in developmental dyslexia. *Cortex*, *73*, 131–143. https://doi.org/10.1016/j.cortex.2015.08.008

Huensch, A. & Nagle, C. (2021). The effect of speaker proficiency on intelligibility, comprehensibility, and accentedness in L2 Spanish: A conceptual replication and extension of Munro and Derwing (1995a). *Language Learning*, *71*, 626–668. https://doi.org/10.1111/lang.12451

Ingvalson, E. M., Holt, L. L., & McClelland, J. L. (2011). Can native Japanese listeners learn to differentiate /r–l/on the basis of F3 onset frequency? *Bilingualism: Language and Cognition*, *15*, 255–274. https://doi.org/10.1017/S1366728911000447

Ingvalson, E. M., McClelland, J. L., & Holt, L. L. (2011). Predicting native English-like performance by native Japanese speakers. *Journal of Phonetics*, *39*, 571–584. https://doi.org/10.1016/j.wocn.2011.03.003

Iverson, P. Kuhl, P. K., Akahane-Yamada, R., Diesch, E., Tohkura, Y., Kettermann, A., & Siebert, C. (2003). A perceptual interference account of acquisition difficulties for non-native phonemes. *Cognition*, *87*(1), B47–B57. https://doi.org/10.1016/S0010-0277(02)00198-1

Iverson, P., Pinet, M., & Evans, B. G. (2012). Auditory training for experienced and inexperienced second-language learners: Native French speakers learning English vowels. *Applied Psycholinguistics*, *33*, 145–160. https://doi.org/10.1017/S0142716411000300

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

Larson-Hall, J. (2015). *A guide to doing statistics in second language research using SPSS and R*. New York, NY: Routledge.

Lee, A. H., & Lyster, R. (2016). The effects of corrective feedback on instructed L2 speech perception. *Studies in Second Language Acquisition*, *38*, 35–64. https://doi.org/10.1017/S0272263115000194

Lengeris, A., & Hazan, V. (2010). The effect of native vowel processing ability and frequency discrimination acuity on the phonetic training of English vowels for native speakers of Greek. *The Journal of the Acoustical Society of America*, *128*, 3757–3768. https://doi.org/10.1121/1.3506351

Lenth, R. V., Buerkner, P., Herve, M., Love, J., Miguez, F., Riebl, H., & Singmann, H. (2021). Estimated marginal means aka least-square means (Version 1.7.3) [Computer software]. Retrieved from https://cran.r-project.org/web/packages/emmeans/index.html

Li, S. (2010). The effectiveness of corrective feedback in SLA: A meta-analysis. *Language Learning*, *60*, 309–365. https://doi.org/10.1111/j.1467-9922.2010.00561.x

Lim, S. J., Fiez, J. A., & Holt, L. L. (2014). How may the basal ganglia contribute to auditory categorization and speech perception? *Frontiers in Neuroscience*, *8*, 1–18. https://doi.org/10.3389/fnins.2014.00230

Lim, S. J., Fiez, J. A., & Holt, L. L. (2019). Role of the striatum in incidental learning of sound categories. *Proceedings of the National Academy of Sciences of the United States of America*, *116*, 4671–4680. https://doi.org/10.1073/pnas.1811992116

Lim, S. J., & Holt, L. L. (2011). Learning foreign sounds in an alien world: Videogame training improves non-native speech categorization. *Cognitive Science*, *35*, 1390–1405. https://doi.org/10.1111/j.1551-6709.2011.01192.x

Lively, S. E., Logan, J. S., & Pisoni, D. B. (1993). Training Japanese listeners to identify English /r/and /l/: II. The role of phonetic environment and talker variability in learning new perceptual categories. *Journal of the Acoustical Society of America*, *94*, 1242–1255. https://doi.org/10.1121/1.408177

Logan, J. S., Lively, S. E., & Pisoni, D. B. (1991). Training Japanese listeners to identify English /r/and /1/: A first report. *Journal of the Acoustical Society of America*, *89*, 874–886. https://doi.org/10.1121/1.1894649

Lyster, R., & Saito, K. (2010). Oral feedback in classroom SLA: A meta-analysis. *Studies in Second Language Acquisition*, *32*, 265–302. http://doi.org/10.1017/S0272263109990520

Major, R. C. (2001). *Foreign accent: The ontogeny and phylogeny of second language phonology*. New York, NY: Routledge.

McClelland, J. L., Fiez, J. A., & McCandliss, B. D. (2002). Teaching the /r/-/l/discrimination to Japanese adults: Behavioral and neural aspects. *Physiology and Behavior*, *77*, 657–662. https://doi.org/10.1016/S0031-9384(02)00916-2

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, *264*(5588), 746–748. https://doi.org/10.1038/264746a0

Nagle, C. L. (2021). Revisiting perception–production relationships: Exploring a new approach to investigate perception as a time-varying predictor. *Language Learning*, *71*, 243–279. https://doi.org/10.1111/lang.12431

Nishi, K., Strange, W., Akahane-Yamada, R., Kubo, R., & Trent-Brown, S. A. (2008). Acoustic and perceptual similarity of Japanese and American English vowels. *The Journal of the Acoustical Society of America*, *124*, 576–588. https://doi.org/10.1121/1.2931949

Ortega, M., Mora, J. C., & Mora-Plaza, I. (2019, May). *The role of visual monitoring in training L2 vowels*. Paper presented at the 6th International Conference on English Pronunciation: Issues and Practices (EPIP 6), Skopje, North Macedonia.

Piske, T., Flege, J., MacKay, & Meador, D. (2011). Investigating native and non-native vowels produced in conversational speech. In M. Wrembel, M. Kul, & K. Dziubalska-Kołaczyk (Eds.), *Achievements and perspectives in the acquisition of second language speech: New Sounds* 2010 (pp. 195–205). Frankfurt am Main: Peter Lang.

R Core Team (2018). R: A language and environment for statistical computing (Version 4.0.5) [Computer software]. Vienna, Austria: R Foundation for Statistical Computing. www.r-project.org

Reber, A. S. (1989). Implicit learning and tacit knowledge. *Journal of Experimental Psychology: General*, *118*, 219–235. https://doi.org/10.1037/0096-3445.118.3.219

Rebuschat, P. (2013). Measuring implicit and explicit knowledge in second language research. *Language Learning*, *63*, 595–626. https://doi.org/10.1111/lang.12010

Saito, K. (2013a). Age effects on late bilingualism: The production development of /ɹ/by high-proficiency Japanese learners of English. *Journal of Memory and Language*, *69*, 546–562. https://doi.org/10.1016/j.jml.2013.07.003

Saito, K. (2013b). Reexamining effects of form-focused instruction on L2 pronunciation development: The role of explicit phonetic information. *Studies in Second Language Acquisition*, *35*, 1–29. https://doi.org/10.1017/S0272263112000666

Saito, K. (2015). Communicative focus on second language phonetic form: Teaching Japanese learners to perceive and produce English /ɹ/without explicit instruction. *Applied Psycholinguistics*, *36*, 377–409. https://doi.org/10.1017/S0142716413000271

Saito, K. (2018). Advanced segmental and suprasegmental acquisition. In P. Malovrh & A. Benati (Eds.), *The handbook of advanced proficiency in second language acquisition* (pp. 282–303). Hoboken, NJ: Wiley Blackwell. https://doi.org/10.1002/9781119261650.ch15

Saito, K. (2019). Individual differences in second language speech learning in classroom settings: Roles of awareness in the longitudinal development of Japanese learners' English /ɹ/pronunciation. *Second Language Research*, *35*, 149–172. https://doi.org/10.1177/0267658318768342

Saito, K., Hanzawa, K., Petrova, K., Kachlicka, M., Suzukida, Y., & Tierney, A. (2022a). *Questionnaire. Materials from "Incidental and multimodal high variability phonetic training: Potential, limits, and future directions"* [Questionnaire]. IRIS Database, University of York, UK. https://doi.org/10.48316/qb2m-tb03

Saito, K., Hanzawa, K., Petrova, K., Kachlicka, M., Suzukida, Y., & Tierney, A. (2022b). *Study materials. Materials from "Incidental and multimodal high variability phonetic training: Potential, limits, and future directions"* [Collection: Language task]. IRIS Database, University of York, UK. https://doi.org/10.48316/rmfn-dw32

Saito, K., & Lyster, R. (2012). Effects of form-focused instruction and corrective feedback on L2 pronunciation development of /ɹ/by Japanese learners of English. *Language Learning*, *62*, 595–633. http://doi.org/10.1111/j.1467-9922.2011.00639.x

Saito, K., & Plonsky, L. (2019). Effects of second language pronunciation teaching revisited: A proposed measurement framework and meta-analysis. *Language Learning*, *69*, 652–708. https://doi.org/10.1111/lang.12345

Saito, K., Suzukida, S., Oyama, T., & Akiyama, Y. (2021). How does longitudinal interaction promote second language speech learning? Roles of learner experience and proficiency levels. *Second Language Research*, *37*, 547–571. https://doi.org/10.1177/0267658319884981

Sakai, M., & Moorman, C. (2018). Can perception training improve the production of second language phonemes? A meta-analytic review of 25 years of perception training research. *Applied Psycholinguistics*, *39*, 187–224. https://doi.org/10.1017/S0142716417000418

Seitz, A. R., Protopapas, A., Tsushima, Y., Vlahou, E. L., Gori, S., Grossberg, S., & Watanabe, T. (2010). Unattended exposure to components of speech sounds yields same benefits as explicit auditory training. *Cognition*, *115*, 435–443. https://doi.org/10.1016/j.cognition.2010.03.004

Shinohara, Y., & Iverson, P. (2018). High variability identification and discrimination training for Japanese speakers learning English /r/–/l/. *Journal of Phonetics*, *66*, 242–251. https://doi.org/10.1016/j.wocn.2017.11.002

Suzuki, Y., & Hanzawa, K. (2021). Massed task repetition is a double-edged sword for fluency development. *Studies in Second Language Acquisition*, 1–26. https://doi.org/10.1017/S0272263121000358

Thomson, R. I. (2012). Improving L2 listeners' perception of English vowels: A computer-mediated approach. *Language Learning*, *62*, 1231–1258. https://doi.org/10.1111/j.1467-9922.2012.00724.x

Thomson, R. I. (2018). High variability [pronunciation] training (HVPT): A proven technique about which every language teacher and learner ought to know. *Journal of Second Language Pronunciation*, *4*, 208–231. https://doi.org/10.1075/jslp.17038.tho

Thomson, R. I., & Derwing, T. M. (2016). Is phonemic training using nonsense or real words more effective? In J. Levis, H. Le, I. Lucic, E. Simpson, & S. Vo (Eds.), *Proceedings of the 7th Pronunciation in Second Language Learning and Teaching Conference* (pp. 88–97). Ames, IA: Iowa State University.

Tricomi, E., Delgado, M. R., & Fiez, J. A. (2004). Modulation of caudate activity by action contingency. *Neuron*, *41*, 281–292. https://doi.org/10.1016/S0896-6273(03)00848-1

Tricomi, E., Delgado, M. R., McCandliss, B. D., McClelland, J. L., & Fiez, J. A. (2006). Performance feedback drives caudate activation in a phonological learning task. *Journal of Cognitive Neuroscience*, *18*, 1029–1043. https://doi.org/10.1162/jocn.2006.18.6.1029

Trofimovich, P., & Gatbonton, E. (2006). Repetition and focus on form in processing L2 Spanish words: Implications for pronunciation instruction. *The Modern Language Journal*, *90*, 519–535. https://doi.org/10.1111/j.1540-4781.2006.00464.x

Tsunemoto, A., Lindberg, R., Trofimovich, P., & McDonough, K. (2021). Visual cues and rater perceptions of second language comprehensibility, accentedness, and fluency. *Studies in Second Language Acquisition*, 1–26. https://doi.org/10.1017/S0272263121000425

Uchihara, T., Webb, S., Saito, K., & Trofimovich, P. (2021). The effects of talker variability and frequency of exposure on the acquisition of spoken word knowledge. *Studies in Second Language Acquisition*, 1–24. https://doi.org/10.1017/S0272263121000218

Uchihara, T., Webb, S., & Yanagisawa, A. (2019), The effects of repetition on incidental vocabulary learning: A meta-analysis of correlational studies. *Language Learning*, *69*, 559–599. https://doi.org/10.1111/lang.12343

Vlahou, E. L., Protopapas, A., & Seitz, A. R. (2012). Implicit training of nonnative speech stimuli. *Journal of Experimental Psychology: General*, *141*, 363–381. https://doi.org/10.1037/a0025014

Wade, T., & Holt, L. L. (2005). Incidental categorization of spectrally complex non-invariant auditory stimuli in a computer game task. *The Journal of the Acoustical Society of America*, *118*, 2618–2633. https://doi.org/10.1121/1.2011156

Wang, Y., Jongman, A., & Sereno, J. A. (2003). Acoustic and perceptual evaluation of Mandarin tone productions before and after perceptual training. *The Journal of the Acoustical Society of America*, *113*, 1033–1043. https://doi.org/10.1121/1.1531176

Supporting Information

Additional Supporting Information may be found in the online version of this article at the publisher's website:

Appendix S1. Methodological Issues in High Variability Phonetic Training Research.
Appendix S2. Experimental Versus Distracter Stimuli.
Appendix S3. Minimal Pairs in Perception and Production Tasks.
Appendix S4. Justification of Methods.
Appendix S5. Descriptive Statistics of Pretest and Posttest Data.

Appendix: Accessible Summary (also publicly available at https://oasis-database.org)

Learning new second language sounds as a by-product of playing a videogame: Potential and limitations

What This Research Was About and Why It Is Important

In the field of second language (L2) speech, researchers have extensively investigated a language-focused, highly explicit training method, that is, high variability phonetic training (HVPT). However, many have claimed that the

nature of such instruction has heavily relied on decontextualized practice that might not develop learners' communicative competence. Recently, scholars have shown that learning new sounds as a by-product of another activity in a multimodal context, in other words, incidental and multimodal training, could be more effective than the exclusive use of explicit and language-focused training. To test the pedagogical potential and limits of this approach, we conducted an intervention study and designed a web-based shooting game focusing on Japanese speakers' English phonetic acquisition. Participants were told that the faster they shot targets, the more points they would earn. Unknown to the participants, each target was preceded by unique English consonants and vowels. As such, L2 learners were incidentally guided to use phonological cues and acquire a series of novel foreign sounds as a by-product of playing the videogame.

What the Researchers Did

- We recruited 58 Japanese learners of English. They were divided into two groups, consonant training ($n = 33$) and vowel training ($n = 25$). They used their smartphones to play a clay shooting game, for a total of 3 hours over 6 days. As soon as a clay target flew on the screen, participants shot it by touching its location on the screen with their finger.
- Two different phonological contrasts in English were used as the target of training. In consonant training, participants focused on the discrimination of English [r] and [l] (e.g., "rock" vs. "lock"). In vowel training, participants worked on the discrimination of English [æ] and [ʌ] (e.g., "hat" vs. "hut").
- Although participants were not told, there were four targets with unique colors (red, gold, yellow, and purple) and movements (rightward, upward, and leftward). Right before each target appeared on the screen, participants heard unique English sounds that appeared predictably before specific movements. As such, participants could, without having been informed, predict each clay's movements based on the preceding sound cues for the movements.

What the Researchers Found

- We found not only that incidental training significantly improved Japanese participants' L2 speech perception but also that gains in the perception domain successfully transferred to the production domain.
- Learning gains were observed in the acquisition of English [æ] and [ʌ], but not of English [r] and [l].

- One source of this variation in learning could be due to the differential amount of learning difficulty (with English [r]-[l] being more difficult than English [æ]-[ʌ]). Although few Japanese speakers have been found to master nativelike English [r] and [l] performance, many seem to achieve advanced proficiency in pronouncing English [æ] and [ʌ] sounds.

Things to Consider

- The findings suggest a potential value of an incidental and multimodal approach to L2 speech learning, that is, learning both auditory, visuospatial, and motor domains of new sounds as a by-product of gameplay.
- The findings suggest that this approach can be beneficial, at least when the treatment focuses on a relatively easy aspect of L2 speech acquisition (e.g., Japanese speakers' acquisition of English [æ]–[ʌ] sounds).
- However, more elaborate strategies may be needed when training focuses on relatively difficult aspects of L2 speech acquisition (e.g., Japanese speakers' acquisition of English [r]–[l] sounds).
- Gains were not as large as those found by other studies that used more explicit types of pronunciation training.

Materials, data, open access article: Materials are publicly available at https://www.iris-database.org.

How to cite this summary: Saito, K. (2022). Learning new second language sounds as a by-product of playing a videogame: Potential and limitations. *OASIS Summary* of Saito, Hanzawa, Petrova et al. (2022) in *Language Learning*. https://oasis-database.org