Bilingualism: Language and Cognition

cambridge.org/bil

Dimension-specific roles of pitch and non-pitch processing in instructed L2 suprasegmental learning

Izumi Hosaka[1] and Kazuya Saito[1,2]

[1]Institute of Education, University College London, UK and [2]Graduate School of International Studies, Tohoku University, Japan

Research Article

Cite this article: Hosaka, I. and Saito, K. (2026). Dimension-specific roles of pitch and non-pitch processing in instructed L2 suprasegmental learning. *Bilingualism: Language and Cognition* 1–14. https://doi.org/10.1017/S1366728926101576

Received: 28 July 2025 UTC
Revised: 24 May 2026 UTC
Accepted: 25 May 2026 UTC

Keywords:
English word stress; L2 suprasegmentals; perceptual training; auditory processing; perceptual-cognitive aptitude

Corresponding author:
Izumi Hosaka;
Email: izumi.hosaka.23@ucl.ac.uk

This research article was awarded Open Data and Open Materials badges for transparent practices. See the Data Availability Statement for details.

Abstract

Speakers of pitch-accent languages, such as Japanese, are often overly sensitive to pitch cues, which may affect second language (L2) suprasegmental learning. The present study examined whether sensitivity to non-pitch cues (i.e., second formant, duration and amplitude rise time) is associated with more accurate perception of L2 English word stress following phonetic training. A total of 115 Japanese English-as-a-foreign-language learners were assessed on pitch and non-pitch dimensions of auditory processing and assigned to experimental or control groups. The experimental groups received perceptual phonetic training on English word stress, cued not only by pitch, but also by non-pitch information (duration, intensity and vowel quality). Results showed that brief training significantly enhanced prosodic perception of trained words. Importantly, auditory processing exhibited *dimension-specific* effects: more precise processing of non-pitch cues was associated with superior perception and greater training gains, whereas pitch sensitivity was unrelated to learning outcomes.

Highlights

- Perceptual training enhanced L2 English word stress perception for L1 Japanese learners.
- Training gains were specific to trained words, with limited generalization.
- Sensitivity to non-pitch cues predicted better stress perception and learning.
- Pitch sensitivity was unrelated to training gains in English word stress.
- L1 pitch-based strategies may hinder adaptation to L2-appropriate cue weighting.

1. Introduction

Effective second language (L2) communication depends not only on accurate perception and production of segmentals (e.g., vowels and consonants) but also on suprasegmentals such as intonation, rhythm and stress (Derwing et al., 2022). Although L2 pronunciation instruction has traditionally emphasized segmental accuracy, a growing body of empirical evidence points to the greater impact of suprasegmentals on intelligibility and comprehensibility especially when it comes to L2 English speech (Field, 2005; Isaacs & Trofimovich, 2012; Kang et al., 2010). This underscores the need to expand pedagogical focus toward these overarching prosodic elements for English-as-a-Foreign-Language (EFL) learners all over the world.

Among suprasegmental features, word stress plays a pivotal role in English. It profoundly influences both native and non-native listeners' ability to recognize words in real-time processing (Field, 2005) and comprehend and recall the content of speech (Hahn, 2004). English word stress is characterized by a complex combination of multiple acoustic cues, including the perception of fundamental frequencies (i.e., pitch) as well as non-pitch information such as duration, intensity and vowel quality[1] (Fry, 1955; Kochanski et al., 2005). Successful stress perception and production therefore require listeners not only to detect these acoustic cues but also to assign appropriate relative weight to each cue in accordance with its informativeness in the L2.

For many EFL learners, especially those whose first language (L1) prosodic systems fundamentally differ from English in structures and cue weightings, acquiring English word stress presents notable challenges. As languages vary in how strongly they rely on pitch, duration, intensity and vowel quality to signal prosody, speakers of different languages develop language-specific perceptual strategies that prioritize certain acoustic dimensions over others (e.g., Archibald, 1997; Peperkamp & Dupoux, 2002). As a result, L2 learners tend to over-rely on

[1]Here, pitch, duration, and intensity are treated as suprasegmental cues typically realized at the level of the syllable. Although vowel quality refers to the segmental-level spectral properties of sounds, it also serves as another important cue to stress *absence*. When vowels are not stressed, they are reduced, which is often characterized by more centralized formant structures (compared to stressed vowels).

© The Author(s), 2026. Published by Cambridge University Press. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (http://creativecommons.org/licenses/by/4.0), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

acoustic dimensions that are salient in their L1 while underutilizing cues that are more informative in the L2 (Idemaru et al., 2012; Kong & Edwards, 2016).

Empirical evidence consistently supports this cue-weighting account (e.g., Nguyễn et al., 2008; Yu & Andruski, 2010; Zhang & Francis, 2010). For example, Wang (2008) observed that tonal language users (Mandarin speakers) tend to over-rely on pitch while deemphasizing duration and vowel quality cues for English stress perception, wherein learners need to attend to both pitch and non-pitch information. Importantly, these challenges do not necessarily stem from the absence of stress or prominence in the L1 phonological system, but rather from cross-linguistic differences in the relative weighting of acoustic cues used to signal prominence.

Japanese learners of English encounter similar issues due to fundamental differences between the prosodic systems of Japanese and English (Cutler & Otake, 1999). Japanese is a pitch-accent language, where prominence is predominantly realized by variations in fundamental frequency (F0) patterns with limited use of intensity (Sugito, 1980; Vance, 2008). Its mora-timed rhythm, where each mora (a basic phonological unit such as a consonant–vowel unit or a long vowel) contributes relatively equally to overall timing (Port et al., 1987), yields relatively uniform mora duration and full vowel articulation (Beckman, 1982; Tsujimura, 1996). In contrast, English stress is a multidimensional phenomenon that involves a broader set of acoustic cues: pitch, duration, intensity, and most importantly, vowel reduction (Fry, 1955). Its stress-timed rhythm, where stressed syllables occur at regular intervals, leads to reduced vowels and compressed timing in unstressed syllables (Dauer, 1983). As vowel reduction and durational contrast—hallmarks of English stress—are largely absent in Japanese, Japanese learners of English often fail to perceive and produce non-pitch correlates of stress. Instead, they tend to produce all syllables with relatively uniform prominence and length (Fokes et al., 1984; Mochizuki-Sudo & Kiritani, 1991).

From a pedagogical perspective, the primary challenge therefore lies in facilitating the recalibration of L1-based perceptual strategies. For learners from pitch-accent language backgrounds (e.g., L1 Japanese listeners), successful L2 prosodic acquisition depends on their ability to reduce reliance on pitch and to shift attention toward non-pitch cues that are more informative for L2 prosody.

A growing body of research supports the effectiveness of instructional interventions focused on L2 English suprasegmentals. Studies show that both segmental- and suprasegmental-focused instruction enhance the global comprehensibility of L2 speech production (Derwing et al., 1998). For English word stress specifically, even short-term targeted instruction has shown measurable gains in L2 learners' perception and production (e.g., Tanner & Landon, 2009). These findings indicate that L2 suprasegmentals, including word stress, are teachable skills through well-designed instruction.

Over the past five decades, diverse instructional approaches have been developed and tested to support L2 pronunciation learning (Derwing et al., 2022), including High Variability Phonetic Training (HVPT), explicit instruction (Thomson & Derwing, 2015), awareness-raising (Pennington & Ellis, 2000) and combined approaches (Couper, 2006; Derwing & Rossiter, 2003; Wiener et al., 2020). Despite these advancements, pronunciation instruction remains underemphasized in many classroom settings, often due to insufficient teacher training, limited class time, or inflexible curricula (Derwing et al., 2022). In light of these constraints, computer-assisted perceptual training has emerged as a practical and effective alternative for self-directed or blended learning. Studies consistently demonstrate the benefits of perceptual training, especially when targeting a specific linguistic feature (Saito & Plonsky, 2019). It has shown promise across both segmental (e.g., Bradlow, 2008 for Japanese speakers' English /r/−/l/ acquisition) and suprasegmental levels (e.g., Carpenter, 2015 for French learners' English word stress acquisition). Moreover, scholars have suggested that directing learners' attention to specific target form through explicit instruction (Pederson & Guion-Anderson, 2010) and feedback (Lee & Lyster, 2016; Sakai & Moorman, 2018) significantly enhances perceptual training outcomes.

Recent meta-analyses reinforce these findings, showing medium-to-large effect sizes (r = 0.71–0.92) of perceptual training on L2 phonology acquisition (e.g., Sakai & Moorman, 2018; Uchihara, Karas, & Thomson, 2025; Yao et al., 2025). Uchihara, Karas, and Thomson (2025) investigated the effectiveness of perceptual training technique in both L2 speech perception and production accuracy, by analyzing a total of 79 studies. They reported mean perception gains of 14.12% for trained stimuli and 12.96% for untrained stimuli. A subsequent meta-analysis of a total of 31 studies focusing on production outcomes (Uchihara et al., 2024) reported mean gains of 10.56% for trained items and 4.50% for untrained items. These findings collectively suggest that perceptual training techniques are particularly effective in enhancing L2 perception accuracy as measured by identification and/or discrimination tasks with measurable generalization, while also yielding modest benefits for production accuracy.

Despite the well-established benefits of perceptual training and explicit instruction for L2 pronunciation broadly, their application to L2 prosody acquisition remains limited. Notably, several studies have applied perceptual training to one single instance of L2 prosody learning—i.e., the acquisition of Mandarin tones by non-tonal L1 speakers. The existing studies have equally demonstrated significant improvement in lexical tone perception and long-term retention (e.g., Perrachione et al., 2011; Sadakata & McQueen, 2014; Wang et al., 1999). However, the generalizability of such findings to learning of other L2 prosodic features has remained unclear; very little is known about the specific effect of perceptual training on English word stress despite its substantial impact on overall comprehensibility and intelligibility (Derwing et al., 2022).

Recently, Petrova (2025) took a first step toward investigating L1 Mandarin learners' English word stress acquisition via HVPT. Her perceptual training was designed using disyllabic nonwords produced by multiple talkers and in multiple phonetic contexts, incorporating two tasks: a forced-choice identification task, where participants identify the stress pattern of an exemplar nonword, and a category discrimination task, where participants choose which of the three tokens has a different stress pattern (i.e., the oddity task). After six training sessions, the experimental group improved by 12.30% in English word stress perception compared to the pretest, while the control group that received vocabulary training remained unchanged.

While perceptual training has generally proven to be an effective intervention for L2 phonological acquisition, it is important to note that the extent of these benefits varies considerably depending on a range of factors. This variability in training outcomes stems from both learner-external factors—such as task types (Carlet & Cebrian, 2019, 2022), the provision of explicit instruction (Alves & Luchini, 2017) and the specific phonological target (Cebrian & Carlet, 2014)—and learner-internal factors, such as L2 proficiency (Lee & Hwang, 2016) and perceptual-cognitive abilities (e.g., attention

https://doi.org/10.1017/S1366728926101576 Published online by Cambridge University Press

control: Mora-Plaza, Ortega, & Mora, 2022). Among these learner-internal factors, recent literature highlights the crucial role of domain-general auditory processing in shaping L2 speech acquisition, underscoring the importance of examining its interaction with instruction across diverse contexts (for a comprehensive review, see Mora, 2022).

Building on this line of L2 prosody research, the current study examines how perceptual training facilitates word stress acquisition in relation to participants' auditory processing profiles. Domain-general auditory processing is defined as one's ability to detect, discriminate and encode the fine-grained acoustic details of sounds across various kinds of sounds, including speech, music and environmental sounds (Saito & Tierney, 2025). Reflecting its domain-generality, auditory timing skills developed in one domain—such as musical training (e.g., rhythmic entrainment)—have been shown to transfer to linguistic and literacy outcomes, including phonological awareness, speech-in-noise perception and reading (Tierney & Kraus, 2014). These findings suggest that precise auditory processing plays a foundational role in multiple domain-specific phenomena, including L2 phonological acquisition (Mueller et al., 2012). This idea is formalized in the auditory precision hypothesis-L2, which proposes that precise auditory processing acts as a bottleneck for language learning. Empirical research increasingly supports this hypothesis: for example, Legeris and Hazan (2010) found that formant discrimination ability predicted improvement in the perception of English vowel contrasts (/i/−/e/ and /a/−/o/) among L1 Greek learners following HVPT. Similarly, Zhang et al. (2024) observed that perceptual acuity was associated with learning gains in Mandarin lexical tone perception following HVPT among L1 English learners.

In the context of English word stress acquisition, auditory processing may play a particularly critical role. Unlike segmental acquisition, which relies on relatively localized, stable cues (e.g., formant frequencies in vowels), English stress processing involves interpreting multiple, covarying acoustic signals across larger linguistic units. Thus, learners with more refined perceptual acuity may be better able to detect and interpret the nuanced acoustic cues associated with stress (e.g., pitch, formant structure, duration and intensity), thereby facilitating the accurate mental representations of English stress patterns.

Although limited in number, emerging research suggests that more precise auditory processing underpins better pre-instructional performance and learning gains in English word stress acquisition. For example, Saito et al. (2020) found that L1 Chinese learners with more precise formant discrimination demonstrated higher English word stress accuracy after eight months of immersion. In an instructional context, Petrova (2025) reported that pitch acuity predicted better word stress perception at pretest, although auditory abilities did not significantly moderate post-training improvements.

Taken together, existing studies point to a relatively robust link between auditory processing and baseline perceptual sensitivity and training outcomes in L2 phonological acquisition. However, a critical gap remains in the literature: while most evidence to date has focused on segmental contrasts and lexical tones, the extent to which the impact of auditory processing generalizes to other suprasegmental features—particularly English word stress—remains largely underexplored.

Addressing this gap, the present study investigates how individual differences in auditory processing influence the effectiveness of perceptual training on English word stress acquisition among Japanese EFL learners. Grounded in the revised Speech Learning Model (SLM-r) and supported by empirical findings for the perceptual training paradigm (e.g., Alves & Luchini, 2017; Pederson & Guion-Anderson, 2010), this study employed a training design that combined perceptual training with explicit instruction and feedback to direct learners' attention to target acoustic cues. By doing so, this study aims to advance our understanding of how to design and deliver effective suprasegmental instruction personalized to individual learners' auditory abilities.

2. Current study

Building upon the reviewed literature, the current study addressed the following research questions in the context of Japanese EFL learners:

1. What are the effects of perceptual training on learners' perception of English word stress?
2. To what extent do individual differences in auditory processing moderate these training effects?

Based on previous findings and theoretical considerations, two predictions were proposed:

2.1. Prediction 1: Perceptual phonetic training facilitates English word stress perception

We predict that perceptual training will significantly improve Japanese EFL learners' perception of English word stress. This prediction is supported by prior empirical research demonstrating that explicit, perception-based instruction effectively facilitates the acquisition of L2 suprasegmentals, including word stress (e.g., Carpenter, 2015; Tanner & Landon, 2009). In line with the SLM-r (Flege & Bohn, 2021), which underscores that successful phonological acquisition critically depends on the precise perception of phonetic details, we expect that targeted phonetic training such as HVPT will enhance learners' sensitivity to English word stress, particularly given that Japanese learners typically receive limited exposure in EFL settings.

2.2. Prediction 2: Auditory processing moderates training effects

We predict that individual differences in domain-general auditory processing will moderate the effectiveness of perceptual training on English word stress perception. This prediction is grounded in prior research showing that auditory processing abilities influence how effectively learners can utilize acoustic cues to make phonological decisions (Saito & Tierney, 2025). Consistent with this account, previous studies have also demonstrated that auditory processing moderates training gains in other areas of L2 phonology (e.g., English vowel contrasts, Legeris & Hazan, 2010; Mandarin lexical tone, Zhang et al., 2024), although its role in word stress acquisition remains underexplored.

Specifically, we predict that learners with greater sensitivity to acoustic cues other than pitch (F0)—namely formant (F2), amplitude rise time and duration cues—will show greater improvement in word stress perception following training. This enhanced sensitivity is expected to better support attention to the temporal and spectral cues that are more informative for English word stress than pitch cues. In contrast, precise pitch processing is not expected to positively predict training gains and may potentially interfere with performance, given that Japanese learners tend to rely on pitch cues while deemphasizing other cues relevant to English prosody due to their L1-tuned cue weightings.

https://doi.org/10.1017/S1366728926101576 Published online by Cambridge University Press

3. Method

This study employed a quasi-experimental pretest–posttest design. The design included a within-subject factor of Time (pretest, posttest) and a between-subject factor of Group (three experimental groups receiving phonetic training with different spacing schedules and one control group receiving explicit instruction only).

3.1. Participants

A total of 119 Japanese EFL learners (66 females, 53 males) were initially recruited from universities across Japan. Eligibility criteria were (a) native Japanese speaker, (b) currently enrolled in a Japanese university, and (c) no study-abroad or extended stay (>1 month) in English-speaking countries. These criteria ensured a homogenous sample with comparable linguistic and educational backgrounds.

Participants were recruited through an electronic flyer distributed to Japanese universities and posted on a website for psychological experiment participation. All participants were enrolled in Japanese universities, majoring in various subjects. They had learned English exclusively through EFL classrooms in Japan, predominantly through non-native teachers, with an average of 10.3 years of study (*Range* = 4–23 years). The mean age was 22 years (*Range* = 18–38). Their English proficiency levels, assessed by the LexTALE vocabulary test (Lemhöfer & Broersma, 2012), ranged from CEFR B1 to B2, with an average score of 58.8% (SD = 8.6; 95%CI [57.2, 60.3]). None reported hearing impairments or extended experience (more than one month in total) in English-speaking countries.

The data for this study were collected as part of a larger research project examining the perceptual-cognitive, experiential and linguistic profiles of Japanese EFL learners. In our prior study (Hosaka & Saito, 2026), we reported cross-sectional analyses of the data, showing that, after extensive EFL education, Japanese EFL learners' acquisition of word stress was primarily determined by individual differences in auditory processing. In the present project, we report the intervention component of the project, specifically the longitudinal effects of perceptual training on word stress perception.

Four participants were excluded from the final dataset because they failed to complete all required training sessions or the posttest within the designated timeframe, resulting in a final sample of 115 participants. These participants were randomly assigned to Experimental (n = 100) and Control (n = 15).[2] This imbalance in group sizes (n = 100 in the experimental condition versus n = 15 in the control condition) reflects the primary aim of the study. To examine the effects of individual differences in auditory processing among learners who received training, a larger sample was allocated to the experimental group to ensure sufficient statistical power for detecting individual differences. The control group was included primarily to monitor and account for potential test–retest effects; a direct comparison between the experimental and control group was not a central focus of this study, as the efficacy of perceptual training paradigms has been robustly demonstrated in prior research (e.g., Yao et al., 2025).

[2]The participants in the experimental group were subdivided into one of three training conditions: (1) Expanding spacing (n = 31), (2) Equal spacing (n = 31), or (3) Massed spacing (n = 38). They differed in the interval between training sessions (0–1–2-4-8 days, 3–3–3-3-3 days, and 0–0–0-0-0 days). No significant effects of spacing were observed (ps > .05), indicating that differences in session distribution did not influence the outcomes of the current study. More detailed results of the analyses will be reported elsewhere.

3.2. Power analysis

As shown in the previous meta-analyses, L2 phonetic training studies typically demonstrate *medium* sized effects (e.g., Saito & Plonsky, 2019). To attain such effect size (f^2 = 0.15), a power analysis using G*Power (Faul et al., 2009) indicated a minimum sample size of 109. A total of 115 participants were included in the final dataset for the current project and considered statistically sufficient to achieve statistical power of 0.80 (α = .05).

3.3. Data collection

Given the practical convenience for the researcher (based in the UK) and the participants located throughout Japan, all the data-collection procedures were conducted remotely using the online experimental platform Gorilla (Anwyl-Irvine et al., 2020). Study components including word stress perception pre- and posttests, auditory processing measures, perceptual training sessions and a background questionnaire were fully administered online.

Participants were instructed to complete the study on a desktop or laptop with a stable internet connection, in a quiet, private space. The use of earphones or headphones was required to ensure optimal audio quality during auditory tasks. To ensure data reliability in this fully online, self-paced study, a preliminary screening process was implemented. This aimed to familiarize participants with the online environment, minimize technical issues and verify access to required tool (PC, headphones, Google Chrome) in a distraction-free setting.

Potential participants first reviewed detailed information about the study procedures and performed sound checks. Next, they took the LexTALE vocabulary test (Lemhöfer & Broersma, 2012) to assess English proficiency. Only those scoring below 70% (i.e., low to intermediate proficiency) were included to avoid ceiling effects on word stress tasks. Overall, 45 applicants failed this screening and were not invited to the main study to reduce the risk of low-quality or disengaged data. These steps helped minimize technical and attentional issues during the online tasks.

Prior to the main study, participants were provided with a detailed information sheet in Japanese explaining all tasks and requirements and were encouraged to ask questions if any part of the procedure was unclear. Only those who passed the screening and provided informed consent received personalized access links for the main study via email.

On Day 1, participants first completed a pretest, a battery of auditory processing tests consisting of six subtests measuring different auditory dimensions and a background questionnaire. Subsequently, all participants received brief explicit instruction on English word stress. Experimental groups then completed the first training session on the same day. Afterward, email reminders prompted participants to complete subsequent training sessions according to their assigned spacing condition (expanding, equal, or massed). Participants were instructed to complete each session within 24 hours of notification (Day 2–6). Upon completing their final session (or after explicit instruction for the control group), participants took the posttest, which included the same word stress perception measure administered at pretest, with items presented in a randomized sequence.

3.4. Outcome measures (word stress tests)

3.4.1. Test items

The perception tests included 62 disyllabic English words selected from the first and second 1000-word families in the

https://doi.org/10.1017/S1366728926101576 Published online by Cambridge University Press

British National Corpus and the Corpus of Contemporary American English (BNC/COCA) word lists (Nation, 2017). Of these, 31 items were used as target items in the training sessions. These high-frequency words were chosen to ensure that participants were likely familiar with their meanings, thereby reducing potential confounds related to lexical unfamiliarity (see Supplementary Information S1).

Prioritizing high-frequency words served two key purposes: (1) enhancing practical relevance to real-world communication and typical FL classroom instruction and (2) minimizing lexical interference in pronunciation tasks. Previous research has shown that familiar words are usually pronounced more accurately than unfamiliar ones (Flege et al., 1998), likely because known lexical items impose lower cognitive demands, allowing learners to focus more directly on phonological features such as stress (Uchihara, Saito et al., 2025).

Given that Japanese learners of English often omit or misplace stress due to the lack of contrastive stress in their L1, using high-frequency words enabled a more accurate assessment of learners' word stress proficiency. It also facilitated identifying learners' specific needs for pronunciation instruction in commonly used vocabulary.

To maintain a clear focus on stress perception and avoid task-related confounds, the following word types were excluded from the item set: (a) noun–verb minimal pairs with stress shift (e.g., REcord versus reCORD, PERmit versus perMIT), which may cue learners to rely on grammatical rather than prosodic knowledge and (b) compound nouns (e.g., *toothbrush*, *airport*), to avoid assessing compounding rules rather than general word stress perception.

3.4.2. Test procedure

The word stress perception test assessed participants' ability to accurately identify the stressed syllable in spoken English words. The test, modeled after Wang et al. (1999), lasted approximately five minutes to complete and was delivered via the Gorilla platform (see Supplementary Information S2 for test materials).

Each trial began with the auditory presentation of a target word, recorded by a native speaker of American English. Following the audio, the orthographic form of the word appeared on screen, alongside numbered syllables (e.g., "1" and "2" for two-syllable words; "1," "2," and "3" for three-syllable words). Participants were instructed to select the number corresponding to the syllable they perceived as stressed.

The test included two practice trials to familiarize participants with the task format. Although no time limit was imposed, participants were encouraged to respond as quickly as possible to minimize reliance on conscious reflection and maintain a focus on perceptual processing. Scoring followed a binary accuracy system: 1 point for correct identification of the stressed syllable, and 0 points for incorrect identification were given. To ensure data quality and discourage inattentive or potentially dishonest responses, response times (RTs) were screened. A cut-off threshold was set based on the mean RT ± 1 SD. Trials with RTs falling outside this range were marked as incorrect. This decision aimed to control for unusually fast or slow responses that may reflect random clicking, distraction, or task disengagement.

3.5. Auditory processing measures

To assess participants' auditory sensitivity to acoustic cues relevant to English word stress, discrimination thresholds were measured across four acoustic dimensions: pitch (F0), second formant frequency (F2), amplitude rise time and duration. Stimuli consisted of synthesized complex tones presented in AXB discrimination tasks, in which only one acoustic dimension varied within each stimulus set while all other acoustic properties were held constant. The approach follows standard psychoacoustic methodology designed to isolate perceptual sensitivity to specific acoustic parameters (Moore, 2012). The stimuli and procedures used in the study were *directly* adapted from validated auditory processing tasks developed by Kachlicka et al. (2019) and Saito et al. (2020), available through the *SLA Speech Tools* platform (Mora-Plaza, Saito, et al., 2022; http://sla-speech-tools.com/research).

The dimension-specific discrimination tasks were administered in the following order: pitch, formant, amplitude rise time, and duration. The tasks were adopted to measure participants' sensitivity to pitch information (pitch discrimination task) and non-pitch information (formant, amplitude rise time and duration discrimination tasks) relevant to the acquisition of English word stress (the main focus of the current study). Native listeners of English rely on multiple acoustic cues, including pitch height, duration and amplitude, to detect the *presence* of lexical stress in each syllable.

Importantly, native listeners also use the relative spacing between formants to detect the *absence* of lexical stress. Formant structure is relevant to the English word stress because vowel quality is an important acoustic correlate of the distinction between stressed and unstressed syllables (Fear et al., 1995). In particular, unstressed vowels are typically characterized by reduced, *centralized* formant structures, whereas stressed vowels exhibit more peripheral formant patterns (Gay, 1978; Rosner & Pickering, 1994). Because these stress-related differences are reflected in the spectral structure of vowels, sensitivity to formant information may facilitate perception of word stress. In the current study, an F2 discrimination task was used as a proxy for broader formant sensitivity, based on the assumption that performance on F2 reflects sensitivity to overall formant structure required for English word stress perception.

During each trial, participants heard three complex tones separated by 500-ms intervals and were asked to decide whether the first or third stimulus differed from the second. Responses were given by selecting "1" or "3" on the screen. Thresholds were determined using an adaptive staircase procedure (Levitt, 1971), where task difficulty increased after three consecutive correct answers and decreased after a single incorrect response. The test began at a mid-difficulty level (Level 50), with stimulus levels ranging from 1 to 100, and step sizes decreasing after each reversal (i.e., a switch in direction of difficulty). Step size changes were as follows: 10 steps initially, 5 steps after the second reversal, 2 steps after the third, and finally 1 step. Each task ended after 70 trials or 8 reversals. The average threshold from the second reversal onward was recorded, with lower threshold scores indicating higher auditory sensitivity.

3.6. Background questionnaire

Participants completed a comprehensive background questionnaire to report their language learning and musical training experiences (see Supplementary Information S3). The questionnaire collected data on (a) the onset of L2 learning, (b) the total years of L2 instruction; (c) the total hours of English learning and use in the classroom per week; (d) the total hours of English learning and use outside the classroom per week; (e) the total minutes of self-study

https://doi.org/10.1017/S1366728926101576 Published online by Cambridge University Press

in English per week; (f) the total minutes of English use for entertainment (watching TV or videos in English); (g) the total minutes of language learning app use; (h) past experience of listening training; (i) past experience of speaking training; (j) past experience of any instructions on word stress; and (k) experience of musical training (if yes, how many years).

3.7. Perception training

Participants in the experimental groups completed six sessions of perceptual training on English word stress. Each session consists of explicit instruction on English word stress, followed by perceptual phonetic training with immediate feedback. The control group received only explicit instruction, with no perceptual training.

3.7.1. Training procedure

To maximize the effectiveness of the perceptual training, both explicit instruction and immediate feedback were incorporated, based on evidence from prior research in L2 phonological learning (e.g., Lee & Lyster, 2016; Pederson & Guion-Anderson, 2010; Sakai & Moorman, 2018).

Explicit instruction was provided at the outset of each training session to raise learners' awareness of the relevant acoustic cues to English word stress (e.g., pitch, duration, intensity and vowel quality). Instructional content included explanations of what word stress is, how it is phonetically realized and strategies for identifying stress patterns in multisyllabic words. These materials were delivered via on-screen slides accompanied by audio recordings of native speakers' examples. Participants were able to replay the audio as many times as needed to fully understand the concepts. This design was guided by findings from Wiener et al. (2020), who demonstrated that perceptual training alone did not yield significant improvement in L2 Mandarin tone production unless it was preceded by explicit instruction. This supports a training sequence in which focused instruction enhances perceptual readiness and enables learners to better process and benefit from input during subsequent training phases. After explicit instruction, the participants were divided into Experimental (receiving training) or Control (receiving no training) groups.

3.7.2. Experimental group

The main component of the perceptual training consisted of stress-identification tasks using 31 target items. In each trial, participants listened to a native speaker's pronunciation of a word and identified the stressed syllable by selecting the corresponding number on the screen (e.g., "1" for the first syllable, "2" for the second and so on). After each response, immediate feedback was provided, including the correct answer, the participant's current accuracy score and response speed (See Figure 1). The native speaker's example pronunciation was replayed to reinforce the accurate form. This decision was motivated by Lee and Lyster (2017), who found that perceptual training with auditory feedback—particularly when learners are exposed to either the correct target form or the incorrect form they had selected—facilitates learning and promotes transfer to production. In contrast, feedback including both target and nontarget forms, or no auditory reinforcement, was found to be less effective. Thus, the use of explicit instruction and targeted corrective feedback in the present study aimed to direct learners' attention to relevant acoustic cues and reinforce accurate phonological representations.

3.7.3. Control group

Those in the control group received only the explicit instruction component of the training. No perceptual practice was provided. This group was included to control for test–retest effects, as the word stress perception test was administrated at pretest and posttest.

4. Results

All the data, R script, testing and training materials are available on OSF: https://osf.io/enqux/?view_only=b25b0ab8a6264ac3997cca786c14c42c

Descriptive statistics for participants' word stress perception accuracy scores by Group (control versus experimental), Time (pretest versus posttest) and Item Type (trained versus untrained item) are summarized in Table 1 and visually presented in Figure 2. Perception test scores were operationalized as accuracy (i.e., the proportion of correct responses).

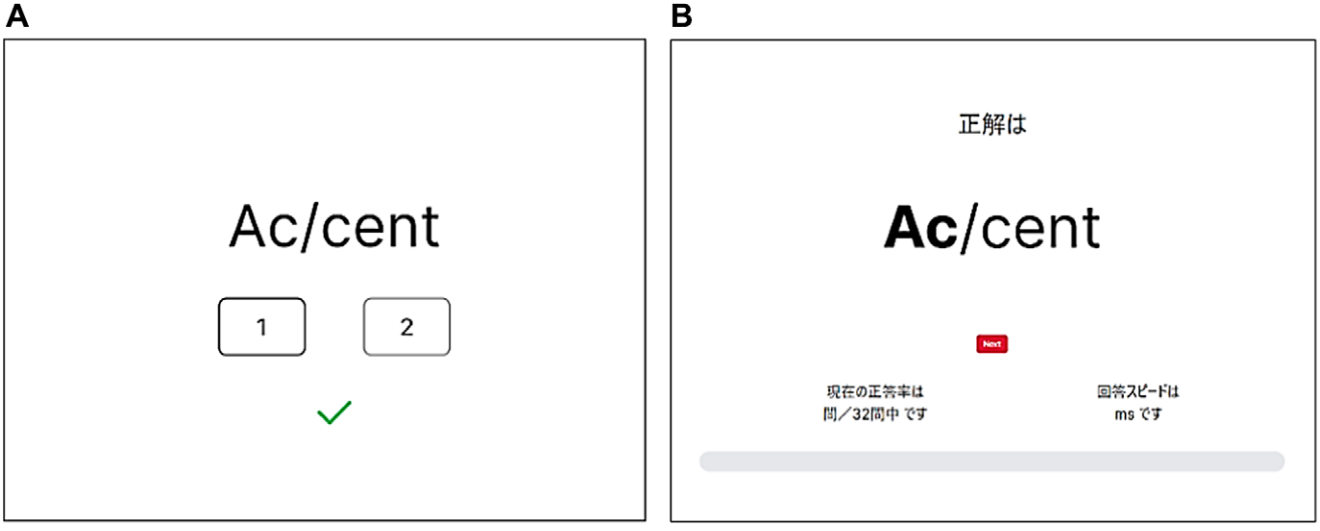


Figure 1. *Example screens from the perceptual training.* In the word stress identification task (1A), participants listen to a native speaker pronounce a target word, then select the stressed syllable by clicking its corresponding syllable number. Immediate feedback (1B) provides current accuracy scores and response speed for each response.

https://doi.org/10.1017/S1366728926101576 Published online by Cambridge University Press

Preliminary analyses indicated that participants generally demonstrated relatively advanced abilities in identifying stressed syllables in high-frequency English words (the BNC/COCA 1k–2k bands) at the outset, as reflected in relatively high mean accuracy and low standard deviation (control: M = 69.0–71.2%, SD = 0.45–0.46; experimental: M = 73.5–75.3%, SD = 0.43–0.44). Crucially, the experimental group showed substantial gains in accuracy from pretest to posttest across both trained and untrained items, with greater improvements observed for the trained items. On average, the experimental group's accuracy improved by 14.1% for trained items and by 6.8% for untrained items. In contrast, the control group showed smaller improvements (5.8–6.4%), given that all test items were untrained for this group.

To statistically evaluate the overall training effects, a generalized linear mixed-effects model was fitted using binomial logistic regression. Participants' binary accuracy for each stimulus (0 for incorrect, 1 for correct) served as the dependent variable. Fixed effects included Time (pretest versus posttest), Group (control versus experimental) and Item Type (untrained versus trained item), along with all two- and three-way interactions. Random intercepts were included for participants (ID) and items to account for variability across individuals and test items. A random-intercept structure was used, as including a by-participant random slope for Time did not improve model fit and risked overfitting given the modest sample size and complex fixed-effects structure. The model was implemented in R using the *glmer* function. We constructed the model as follows: Accuracy ~ Time * Group * Item Type + (1 | ID) + (1 | item).

As shown in Table 2, the model revealed a significant main effect of Time (b = 0.61, OR = 1.84, p = .002), indicating overall

Table 1. Descriptive summary of participants' improvement in perception accuracy of English word stress

	Pretest			Posttest		
Group	*M* (%)	SD	95% CI (%)	*M* (%)	SD	95% CI (%)
Control (n = 15)						
Target item	71.20	0.453	[67.10, 75.30]	77.60	0.417	[73.80, 81.40]
Untrained item	69.00	0.463	[64.80, 73.20]	74.80	0.434	[70.90, 78.80]
Experimental (n = 100)						
Target item	75.30	0.431	[73.80, 76.80]	89.40	0.308	[88.30, 90.50]
Untrained item	73.50	0.441	[72.00, 75.10]	80.30	0.398	[78.90, 81.70]

Note: Accuracy values represent proportion correct scores multiplied by 100 for presentation (M = mean percentage accuracy). Standard deviations (SD) are reported on the original proportion scale (0–1), derived from binary item-level accuracy data. Confidence intervals (CI) are reported as percentages.

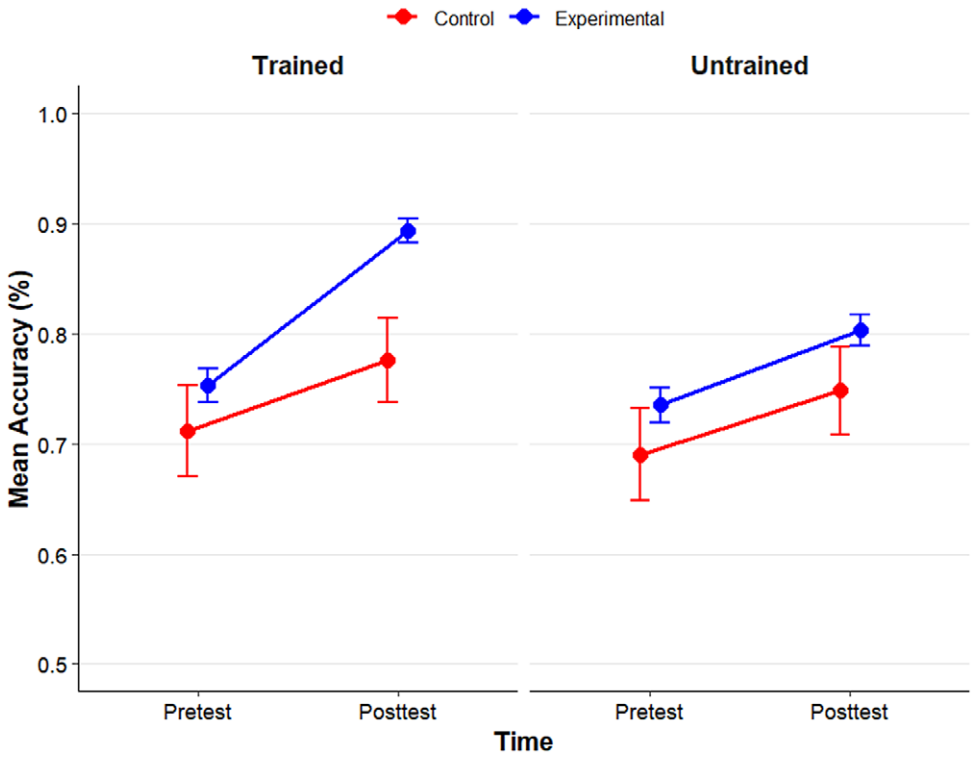


Figure 2. Mean accuracy with 95% CI by group, time and item type.

https://doi.org/10.1017/S1366728926101576 Published online by Cambridge University Press

Table 2. Summary of mixed effects modeling analyses of word stress perception accuracy

Fixed effects	*b*	*SE*	95% CI	*z*	*p*
Intercepts	1.359	0.530	[0.320, 2.397]	2.564	.010*
Time	0.609	0.200	[0.225, 1.002]	3.030	.002*
Group	0.180	0.430	[−0.662, 1.022]	0.419	.676
Item Type	−0.192	0.527	[−1.225, 0.841]	−0.364	.716
Time: Group	0.942	0.220	[0.510, 1.374]	4.278	< .001*
Time: Item Type	−0.062	0.286	[−0.622, 0.498]	−0.216	.829
Group: Item Type	−0.001	0.206	[−0.405, 0.403]	−0.005	.996
Time: Group: Item Type	−0.827	0.310	[−1.433, −0.220]	−2.671	.008*
Random effects	**Variance**	***SD***	$R^2_{conditional}$	$R^2_{marginal}$	
Participants	2.084	1.444	.654	.045	
Items	3.711	1.926			

Note: *Indicates $p < .05$. The reference categories in the model were pretest (Time) and trained items (Item Type).

improvement in word stress perception from pretest to posttest across all participants and item types. More importantly, a significant two-way Time × Group interaction ($b = 0.94$, $OR = 2.57$, $p < .001$) and a significant three-way Time × Group × Item Type interaction ($b = −0.83$, $OR = 0.44$, $p = .008$) were observed. These interactions indicate that changes in accuracy over time depended jointly on whether participants received phonetic training and whether items were trained.

To unpack this three-way interaction, pairwise comparisons of pretest and posttest performance were fitted across groups and item types, using estimated marginal means with Bonferroni correction. Between-group comparisons of pretest and posttest performance are summarized in Table 3.

At pretest, there were no significant differences between the experimental and control groups for either trained ($b = −0.18$, $OR = 0.84$, $p = .676$) or untrained items ($b = −0.18$, $OR = 0.84$, $p = .678$), confirming that the groups were comparable at baseline. At posttest, however, the experimental group outperformed the control group on trained items ($b = −1.12$, $OR = 0.33$, $p = .011$), suggesting that the control group had only 33% of the odds of a correct response compared to the experimental group. In other words, the experimental group was approximately three times more likely to correctly perceive trained word stress after training. In contrast, no significant group difference was found for untrained items at posttest ($b = −0.29$, $OR = 0.75$, $p = .501$), which indicates that training effectiveness was not generalizable from trained to untrained items.

Within-group comparisons of pretest and posttest performance are summarized in Table 4. The experimental group showed significant improvements from pretest to posttest on both trained items ($b = −1.55$, $OR = 0.21$, $p < .001$) and untrained items ($b = −0.66$, $OR = 0.52$, $p < .001$), with substantially larger gains for trained items. This indicates that the perceptual training substantially enhanced stress perception accuracy, with the odds of an incorrect response being approximately 4.76 times (calculated as 1/OR) lower on trained items after training. Notably, although the effect partially generalized to untrained lexical items, its size of improvement of untrained items ($b = −0.66$) was less than a half of that of trained items ($b = −1.55$). Indeed, the effect size was rather comparable to what the control group demonstrated (as an indicator of test–retest effects). The control group also demonstrated modest but statistically significant gains over time: for trained items ($b = −0.61$, $OR = 0.54$, $p = .002$) and untrained items ($b = −0.55$, $OR = 0.58$, $p = .007$). These improvements likely reflect general test practice effects and/or increased familiarity with the task. However, the gains observed in the experimental group—particularly for trained items—were substantially larger in magnitude than those observed in the control group, indicating that perceptual training yielded additional benefits beyond baseline improvement.

Taken together, these results demonstrate that the experimental group showed a significantly larger improvement specifically on trained items and was significantly more accurate than the control group after training. Importantly, while phonetic training facilitated accurate stress perception, this effect was primarily contingent

Table 3. Between-group comparisons at each time point (control − experimental)

Time	Item Type	Estimate (log-odds)	*SE*	*z*	*p*
Pretest	Trained	−0.180	0.430	−0.419	.676
Pretest	Untrained	−0.179	0.431	−0.415	.678
Posttest	Trained	−1.122	0.439	−2.558	.011*
Posttest	Untrained	−0.294	0.437	−0.674	.501

Note: *Indicates $p < .05$. Estimates are reported on the log-odds scale. Negative values indicate greater accuracy in the experimental group compared to the control group (control−experimental). *SE* = standard error.

Table 4. Time effects within each group (estimated marginal means)

Group	Item Type	Estimate (log-odds)	*SE*	*z*	*p*
Control	Trained	−0.609	0.201	−3.030	.002*
Control	Untrained	−0.547	0.204	−2.678	.007*
Experimental	Trained	−1.551	0.091	−17.076	< .001*
Experimental	Untrained	−0.662	0.079	−8.395	< .001*

Note: *Indicates $p < .05$. Estimates are reported on the log-odds scale. Negative values indicate increased accuracy at posttest compared to pretest (pretest−posttest contrast). *SE* = standard error.

https://doi.org/10.1017/S1366728926101576 Published online by Cambridge University Press

on trained lexical items. This provides evidence for item-specific phonetic learning, suggesting that extending these phonetic gains to untrained lexical contexts (i.e., lexical encoding) is a more demanding process that may be constrained by individual auditory processing abilities.

To address the second research question regarding whether individual differences in auditory processing moderated training effects, we first examined descriptive statistics for participants' raw scores across each auditory processing dimension (see Figure 3 and Supplementary Table S1). These scores were recorded on a 100-point scale, with smaller values indicating greater perceptual sensitivity. Descriptive statistics revealed that, on average, participants exhibited highest accuracy in pitch discrimination (M = 12.85 Hz, SD = 13.60). As shown in Figure 3, pitch discrimination scores demonstrated a strong positive skew, consistent with the view that Japanese learners—speakers of a pitch-accent language—are generally highly sensitive to pitch, in line with prior findings from speakers of other tonal or pitch-accent languages (e.g., Mandarin speakers; Wang, 2008). In contrast, participants showed lower accuracy in non-pitch dimensions, with mean discrimination thresholds of 42.17 Hz (SD = 17.08) for formant, 29.63 ms (SD = 22.77) for rise time and 22.38 ms (SD = 15.97) for duration.

To explore the dimension-specific relations between auditory processing and perceptual training outcomes, another mixed-effect regression model (Model 2) was fitted to the data of trained items from the experimental group, as only these participants received perceptual training, and training effect was primarily observed for trained items. The focus of the analyses was to assess whether and to what degree participants' training effects within the experimental group could be tied to two different types of their auditory processing profiles—pitch processing (pitch discrimination) and non-pitch processing (formant, duration and amplitude rise time discrimination). To ensure comparability across auditory processing measures, discrimination scores for pitch, formant, rise time and duration were first reverse-coded (where lower scores indicated higher discrimination accuracy), and then standardized into z-scores for subsequent analysis.

The dependent variable was binary accuracy on each item (0 = incorrect, 1 = correct). Fixed effects included Time (pretest versus posttest), the four standardized auditory processing dimensions (pitch, formant, rise time and duration) and their two-way interactions with Time. Participant ID and item were included as random intercepts to account for variability across individuals and test items. The model was implemented in R using the *glmer* function. We constructed the model as follows: Accuracy ~ Time * (Pitch + Formant + Rise time + Duration) + (1 | ID) + (1 | item).

To ensure the stability of the fixed effects, multicollinearity among the auditory processing predictors was checked using Variance Inflation Factors (VIFs). All VIF values were below 1.30 (pitch = 1.30, formant = 1.07, rise time = 1.27, duration = 1.20), confirming that the predictors were sufficiently independent.

As summarized in Table 5, a significant main effect emerged for Time (b = 1.06, OR = 2.90, p < .001), confirming the overall improvement in word stress perception from pretest to posttest within the experimental group. Significant main effects were also found for formant discrimination (b = 0.45, OR = 1.57, p < .001) and duration discrimination (b = 0.27, OR = 1.32, p = .049), suggesting that more precise formant and duration discriminations were

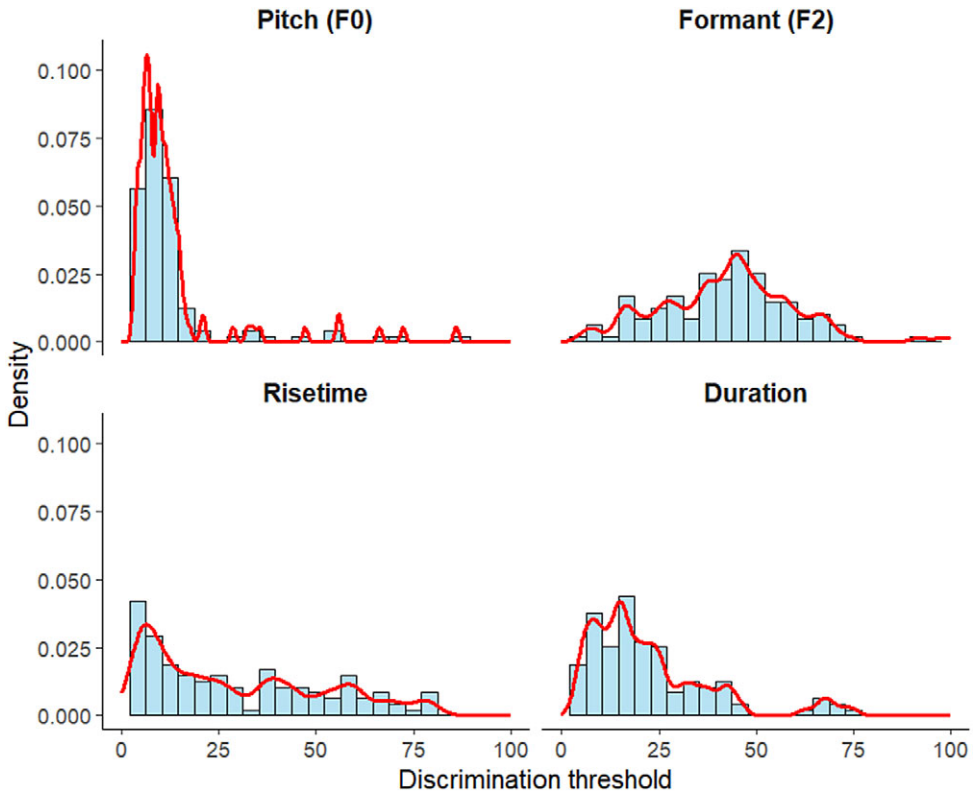


Figure 3. *Distribution of auditory processing scores (raw scores) by dimension.* Lower discrimination threshold values indicate greater perceptual sensitivity. Thresholds are expressed in Hz for pitch (F0) and formant (F2) and in milliseconds (ms) for rise time and duration. The y-axis represents probability density, such that the total area under each distribution equals 1.

https://doi.org/10.1017/S1366728926101576 Published online by Cambridge University Press

Table 5. Summary of mixed effects modeling analyses for all items (model 2)

Fixed effects	*b*	*SE*	95% CI	*z*	*p*
Intercept	1.416	0.276	[0.875, 1.957]	5.130	< .001*
Time	1.064	0.061	[0.945, 1.183]	17.497	< .001*
Pitch (F0)	−0.056	0.146	[−0.342, 0.230]	−0.384	.701
Formant (F2)	0.449	0.128	[0.198, 0.700]	3.505	< .001*
Rise time	0.168	0.142	[−0.111, 0.447]	1.180	.238
Duration	0.274	0.140	[0.001, 0.548]	1.967	.049*
Time: Pitch	−0.092	0.063	[−0.216, 0.032]	−1.453	.146
Time: Formant	−0.144	0.057	[−0.257, −0.032]	−2.509	.012*
Time: Rise time	0.164	0.065	[0.036, 0.292]	2.502	.012*
Time: Duration	0.014	0.060	[−0.104, 0.133]	0.236	.813
Random effects	**Variance**	***SD***	$R^2_{conditional}$	$R^2_{marginal}$	
Participants	1.467	1.211	.638	.072	
Items	3.678	1.918			

Note: * Indicates $p < .05$.

associated with better overall word stress perception at both pretests and posttests. This implies that learners who are more sensitive to subtle differences in formant frequency appear to be better able to detect vowel quality changes related to reduced vowels in unstressed syllables, and learners who are more sensitive to fine timing differences are better able to discern syllable length contrasts that cue English stress patterns. No significant main effects were found for other auditory dimensions.

Importantly, several significant two-way interactions emerged between Time and auditory processing dimensions, suggesting that specific auditory processing skills moderated the amount of training gains. A significant positive interaction was found for rise time discrimination (b = 0.19, OR = 1.21, p = .008), indicating that greater sensitivity to subtle timing changes in amplitude rise time was linked to greater improvement in stress perception following training. This suggests that learners with stronger temporal acuity may have successfully reoriented their focus on temporal information that signals English stress.

Conversely, a significant negative interaction emerged between Time and Formant discrimination ($b = -0.14$, $OR = 0.87$, $p = .012$), indicating that learners with higher formant sensitivity showed comparatively smaller training gains from pretest to posttest. Follow-up analyses using estimated marginal means revealed a pattern consistent with a ceiling effect. Individuals with high formant sensitivity (+1 *SD*) showed higher perception accuracy at both pretest (86.2%) and posttest (94.0%), whereas those with low formant sensitivity (−1 *SD*) showed lower accuracy at pretest (72.4%) and posttest (89.6%). This pattern suggests that learners with higher formant sensitivity began with relatively superior baseline performance, leaving less room for improvement. Finally, the predictive power of pitch processing did not reach statistical significance in any contexts ($p > .05$).

Together, these findings suggest that individual differences in auditory processing ability—specifically sensitivity to non-pitch cues—are associated with both baseline L2 prosody proficiency and the effectiveness of perceptual training. Learners with more precise formant and duration discrimination consistently demonstrated more accurate word stress perception, likely reflecting enhanced sensitivity to vowel quality changes and syllable duration contrasts. While formant discrimination was associated with overall proficiency rather than training-related gains, more precise rise time discrimination was uniquely linked to greater improvement following training. Overall, these results underscore the importance of sensitivity to acoustic cues other than pitch for English word stress perception, particularly in relation to the perceptual constraints faced by learners in this context.

5. Discussion

This study investigated the effects of perceptual training on English word stress perception among Japanese EFL learners and examined how individual auditory processing abilities moderated these effects. The findings provide robust empirical support for the effectiveness of perceptual training and underscore the critical role of auditory processing in shaping L2 suprasegmental learning.

Consistent with prior research on phonetic training for L2 prosody acquisition (e.g., Zhang et al., 2024), our results demonstrated that focused perceptual training significantly enhanced learners' ability to perceive English word stress. The experimental group showed substantial accuracy gains from pretest to posttest, particularly for trained items. The control group, which received only explicit instruction, also exhibited modest improvements (by approximately 6% for both trained and untrained items). These gains likely reflect test–retest effects or minor benefits from explicit instruction provided after the pretest. In contrast, the experimental group improved by approximately 14% on trained items, indicating that the perceptual training itself contributed an additional gain of roughly 7–8% beyond baseline improvement. Notably, the improvement for untrained items (7%) was comparable to the control group, suggesting that brief training (a total of 30 minutes) is insufficient for achieving robust generalization beyond item-specific learning.

These findings reinforce the efficacy of perception-based phonetic training for L2 suprasegmental learning, extending its effectiveness—previously demonstrated primarily in segmental learning (e.g., Uchihara, Karas, & Thomson, 2025)—to the acquisition of English word stress. Importantly, the results

https://doi.org/10.1017/S1366728926101576 Published online by Cambridge University Press

confirm that English word stress, which often resists incidental learning through exposure (Kondo, 2009), can be effectively taught through targeted and explicit phonetic training, even in relatively brief sessions.

Crucially, however, the training effects were largely restricted to trained lexical items. This item-specific learning pattern suggests that perceptual gains observed in this study primarily reflect improvements at phonetic levels, but not at lexical levels. When learners were tested on trained items, they were able to apply newly learned phonetic knowledge successfully; however, when tested on untrained lexical items, performance gains were more limited. This pattern aligns with prior research on lexical encoding of L2 phonological information, which shows that learners often first acquire new phonetic contrasts in a small set of lexical items before generalizing them to the broader lexicon (Saito, 2018). From this perspective, the present findings suggest that perceptual phonetic training can enhance attentions to stress-related acoustic cues, but integrating this knowledge into robust lexical representations likely requires more extensive exposure and training than the brief intervention (six 5-minute sessions).

Beyond confirming the effectiveness of perceptual phonetic training, the present study revealed that individual differences in auditory processing significantly influence L2 word stress acquisition, affecting both baseline performance and training gains. Specifically, our findings show that learners with enhanced sensitivity to formant and duration consistently demonstrated more accurate perception of English word stress. These advantages likely stem from their heightened ability to detect subtle changes in vowel quality (via F2 discrimination) and their greater attunement to rhythmic patterns (via duration discrimination), both of which are key phonetic characteristics of English stress (Dauer, 1983; Fry, 1955).

These findings expand on previous work by Saito et al. (2020), who reported similar benefits of precise formant discrimination for the production of English segmentals and word stress. The present study extends this line of research to perception, demonstrating that sensitivity to formant information also contributes to English word stress perception. Our results revealed that more precise formant discrimination consistently predicted more accurate perception of English word stress; however, it was not associated with greater training gains. Instead, amplitude rise time discrimination (which signals the perceptual salience of stressed syllables in English) predicted greater improvement from pretest to posttest. One interpretation of this pattern is a ceiling effect among learners with stronger formant discrimination. As indicated by the follow-up analyses, these learners started from a relatively high level of accuracy at pretest (86.2%), leaving limited scope for further improvement. Furthermore, the results suggest that L1 Japanese learners with more precise processing of temporal cues (rise time and duration) may have derived greater benefit from perceptual training than those with stronger sensitivity to spectral information (pitch and formant).

This interpretation may also explain why pitch discrimination did not significantly predict either baseline stress perception accuracy or training outcomes, consistent with our predictions. Although pitch is a salient prosodic cue in Japanese prosody, over-reliance on pitch may hinder acquisition of English word stress, which is also signaled by different acoustic features (duration, intensity and vowel quality). This pattern is consistent with previous research showing that Japanese speakers tend to prioritize pitch information in stress perception, even when it is not the most reliable cue in English (Mochizuki-Sudo & Kiritani, 1991). From a cue-weighting perspective, well-developed pitch discrimination may reflect entrenched L1-based perceptual strategies that are difficult to change and may interfere with the adaptation of a more L2-like perceptual strategy.

Taken together, these findings indicate that learners with more precise processing of non-pitch acoustic cues are better positioned to perceive stress contrasts, as they appear more able to reallocate attention away from L1-based reliance on pitch toward L2-relevant temporal cues through training. Accordingly, perceptual training may not be equally effective for all learners; rather, its effectiveness appears to depend partly on learners' underlying auditory processing profiles.

Building on these findings, we now consider their implications for L2 acquisition theory and pedagogy.

5.1. Theoretical and pedagogical implications

The present findings offer a meaningful contribution to L2 acquisition theory by informing the SLM-r (Flege & Bohn, 2021) beyond its traditional focus on L2 segmental learning to a suprasegmental context, specifically in the domain of word stress. Our results support the SLM-r's central claim that successful L2 acquisition is shaped by both the quality of input and learner-internal factors. In this study, focused phonetic training led to measurable improvements in English stress perception, with gains particularly evident among learners with stronger auditory processing of non-pitch dimensions (formant, duration and rise time).

Our findings further support the auditory precision hypothesis-L2 (Mueller et al., 2012). Adult L2 learners experience greater difficulty acquiring suprasegmental features than children because they must reallocate perceptual attention to L2-specific acoustic cues despite entrenched L1 cue-weighting strategies (McAllister et al., 2002). For speakers of pitch-accent languages, the challenge involves shifting attention away from an L1-salient cue (pitch) toward other L2-relevant cues (e.g., intensity) for English stress perception (i.e., L2-specific dimensional weighting strategies, Idemaru et al., 2012; Kong & Edwards, 2016). The observed training effects suggest that explicit, focused instruction can help learners recalibrate their L1-tuned perceptual strategies, provided they have sufficient perceptual sensitivity to non-pitch acoustic cues. In contrast, learners with lower auditory processing abilities demonstrated lower perception accuracy at pretest and more limited gains following training. Together, these results evidenced that auditory processing plays a critical role in determining the success of L2 suprasegmental learning.

From a pedagogical perspective, our findings advocate for incorporating perceptual training into L2 prosody instruction, especially for learners whose L1 places less weight on temporal cues in signaling prosodic prominence (e.g., L1 Japanese). While L1 Japanese speakers use duration information to distinguish phonemic mora length (e.g., *obasan* versus *obaasan*), duration is not typically a primary cue for lexical stress or prominence, with pitch playing a more central role (Mochizuki-Sudo & Kiritani, 1991). The demonstrated benefits of focused perceptual training present a practical approach to help these learners reweight their cues for accessing word stress. This instructional approach can complement both classroom and self-directed learning contexts. As the brief training used in this study (six 5-minute sessions) resulted primarily in item-specific improvements, longer or more varied phonetic training may be necessary to facilitate the encoding of phonological knowledge into a broader range of lexical representations beyond the trained items.

https://doi.org/10.1017/S1366728926101576 Published online by Cambridge University Press

Importantly, given the influence of individual auditory profiles, instruction may be most effective when tailored to learners' perceptual strengths and weaknesses. For example, learners with lower sensitivity to rise time or duration may benefit from explicit scaffolding that draws attention to these acoustic features. Such personalization can be implemented in cost-effective ways through targeted and technology-assisted training. Prior work has demonstrated that relatively brief, targeted training on specific auditory dimensions (e.g., formant discrimination) can lead to measurable improvements both in auditory processing and L2 phonetic learning (e.g., Saito, Petrova, et al., 2022). In addition, computer-assisted language learning (CALL) and mobile platforms can provide scalable solutions by using adaptive algorithms to adjust task difficulty or cue focus based on learner performance (e.g., Kim et al., 2026; Yang et al., 2024). Overall, these findings underscore the value of a personalized and dimension-specific approach to teaching L2 prosody.

5.2. Limitations and future directions

While the study provides robust evidence for the effectiveness of perceptual training, several limitations should be noted. First, the observed training effects were largely limited to trained items, indicating significant improvement at the phonetic level but only partial generalization to untrained lexical items. While this pattern is theoretically informative, it underscores the need for training paradigms that more directly target the refinement of phonolexical representations (e.g., Darcy et al., 2025). Future research should explore instructional approaches designed to bridge the gap between phonetic perception and lexical encoding of L2 suprasegmentals. Advancing this line of inquiry, however, will require the development of more sensitive outcome measures capable of capturing the robustness of suprasegmental lexical representations. In particular, commonly used task for segmental lexical encoding such as lexical decision task—which relies on binary word–nonword judgments—may be insufficient, as suprasegmental errors do not yield nonwords in the same way segmental errors do, despite potentially reducing comprehensibility.

Second, the present findings are based on speech stimuli produced by a single male speaker of American English. It remains an open question whether similar patterns would emerge if learners were exposed to a wide variety of speakers, including those who rely more heavily on pitch cues to signal stress. Although this study was grounded in core principles of perceptual phonetic training, expanding the design to a more comprehensive HVPT paradigm —by incorporating greater talker and phonetic context variability —would help clarify the robustness and generalizability of the observed cue-weighting effects. Furthermore, while single-talker training design was intended to reduce cognitive load for participants, future research using multiple talkers would provide a more robust examination of how dimension-specific auditory processing abilities interact with varying talker profiles.

Third, the current analysis focused on immediate-posttest results; complete data from the delayed posttest results, although partially collected, were not included. Further exploration is thus essential to fully explore the long-term retention of training gains and compare the relative effectiveness of different spacing schedules for sustained L2 prosody acquisition.

Finally, the present study focused exclusively on Japanese learners. While it provides valuable insights into learners with a pitch-accent L1 background, future research should examine whether these findings can generalize across other L1 groups, particularly speakers of other pitch-accent languages (e.g., Swedish, Norwegian), or non-pitch-accent languages (e.g., German, Spanish, Russian). In addition, including non-pitch-accent L1 groups would allow researchers to more directly isolate L1-specific influences on L2 word stress acquisition. Such cross-linguistic comparisons would provide a more comprehensive understanding of how L1 background interacts with auditory processing abilities to shape both preinstructional performance and training outcomes.

Supplementary material. The supplementary material for this article can be found at http://doi.org/10.1017/S1366728926101576.

Data availability statement. The data that supported the findings of this study are openly available in Open Science Framework at https://osf.io/enqux/?view_only=b25b0ab8a6264ac3997cca786c14c42c.

Training and testing materials used to conduct this study are openly available in Gorilla Open Science at https://app.gorilla.sc/openmaterials/1030671.

Acknowledgements. The study was derived from part of the first author's (IH) PhD dissertation submitted to University College London. The project was funded by a *Language Learning* Dissertation Grant awarded to IH and by a Leverhulme Trust Grant (No. RPG-2024-391) and a UK-ISPF Grant (No. 1185702223) awarded to KS. We gratefully acknowledge the participants, as well as the three anonymous reviewers and Editors of BLC for their insightful comments on an earlier version of the manuscript.

Competing interests. The authors declare none.

Ethical considerations. This study was approved by the IOE Research Ethics Committee of University College London (approval no. Z6364106/2024/03/28 social research) on April 8th, 2024.

Consent to participate. All participants provided written informed consents prior to data collection.

References

Alves, U. K., & **Luchini, P. L.** (2017). Effects of perceptual training on the identification and production of word-initial voiceless stops by Argentinean learners of English. *Ilha do Desterro*, **70**(3), 15–32. https://doi.org/10.5007/2175-8026.2017v70n3p15

Anwyl-Irvine, A. L., **Massonnié, J.**, **Flitton, A.**, **Kirkham, N.**, & **Evershed, J. K.** (2020). Gorilla in our midst: An online behavioral experiment builder. *Behavior research methods*, **52**(1), 388–407. https://doi.org/10.3758/s13428-019-01237-x

Archibald, J. (1997). The acquisition of English stress by speakers of nonaccentual languages: Lexical storage versus computation of stress. *Linguistics*, **35**(1), 167–182. https://doi.org/10.1515/ling.1997.35.1.167.

Beckman, M. (1982). Segment duration and the 'mora' in Japanese. *Phonetica*, **39**(2–3), 113–135. https://doi.org/10.1159/000261655.

Bradlow, A. R. (2008). 10. Training non-native language sound patterns: Lessons from training Japanese adults on the English /*/ - /l/ contrast. In J. G. H. Edwards & M. L. Zampini (Eds.), *Phonology and second language acquisition* (pp. 287–308). John Benjamins Publishing Company. https://doi.org/10.1075/sibil.36.14bra.

Carlet, A., & **Cebrian, J.** (2019). Assessing the effect of perceptual training on L2 vowel identification, generalization and long-term effects. In *A sound approach to language matters: In honor of Ocke-Schwen Bohn*. Aarhus Universitet. http://doi.org/10.7146/aul.322.218.

Carlet, A., & **Cebrian, J.** (2022). The roles of task, segment type, and attention in L2 perceptual training. *Applied PsychoLinguistics*, **43**(2), 271–299. http://doi.org/10.1017/S0142716421000515.

Carpenter, A. C. (2015). Phonetic training significantly mitigates the stress 'deafness' of French speakers. *International Journal of Linguistics*, **7**(3), 94–108. https://doi.org/10.5296/ijl.v7i3.7661.

Cebrian, J., & **Carlet, A.** (2014). Second-Language Learners' Identification of Target-Language Phonemes: A Short-Term Phonetic Training Study.

Canadian Modern Language Review, **70**(4), 474–499. https://doi.org/10.3138/cmlr.2318

Couper, G. (2006). The short and long-term effects of pronunciation instruction. *Prospect: an Australian Journal of TESOL*, **21**(1), 44–64.

Cutler, A., & **Otake, T.** (1999). Pitch accent in spoken-word recognition in Japanese. *The Journal of the Acoustical Society of America*, **105**(3), 1877–1888. https://doi.org/10.1121/1.426724.

Darcy, I., **Llompart, M.**, **Hayes-Harb, R.**, **Mora, J. C.**, **Adrian, M.**, **Cook, S.**, & **Ernestus, M.** (2025). Phonological processing and the L2 mental lexicon. *Studies in Second Language Acquisition*, **47**(1), 361–387. https://doi.org/10.1017/S0272263124000482.

Dauer, R. M. (1983). Stress-timing and syllable-timing reanalyzed. *Journal of Phonetics*, **11**(1), 51–62. https://doi.org/10.1016/S0095-4470(19)30776-4.

Derwing, T. M., **Munro, M. J.**, **Thomson, R. I.**, **Derwing, T. M.**, & **Munro, M. J.** (2022). Pronunciation learning and teaching. In *The Routledge handbook of second language acquisition and speaking* (pp. 147–159). Routledge. https://doi.org/10.4324/9781003022497-14.

Derwing, T. M., **Munro, M. J.**, & **Wiebe, G.** (1998). Evidence in favor of a broad framework for pronunciation instruction. *Language Learning*, **48**(3), 393–410. https://doi.org/10.1111/0023-8333.00047.

Derwing, T. M., & **Rossiter, M. J.** (2003). The effects of pronunciation instruction on the accuracy, fluency, and complexity of L2 accented speech. *Applied language learning*, **13**(1), 1–17.

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**(4), 1149–1160. http://doi.org/10.3758/BRM.41.4.1149.

Fear, B. D., **Cutler, A.**, & **Butterfield, S.** (1995). The strong/weak syllable distinction in English. *Journal of the Acoustical Society of America*, **97**(3), 1893–1904. https://doi.org/10.1121/1.412063.

Field, J. (2005). Intelligibility and the listener: The role of lexical stress. *TESOL Quarterly: A Journal for Teachers of English to Speakers of Other Languages and of Standard English as a Second Dialect.*, **39**(3), 399–423. https://doi.org/10.2307/3588487.

Flege, J. E., & **Bohn, O. S.** (2021). The revised speech learning model (SLM-r). In R. Wayland (Ed.), *Second language speech learning: Theoretical and empirical Progress* (pp. 3–83). Cambridge University Press. http://doi.org/10.1017/9781108886901.002.

Flege, J. E., **Frieda, E. M.**, **Walley, A. C.**, & **Randazza, L. A.** (1998). Lexical factors and segmental accuracy in second language speech production. *Studies in Second Language Acquisition*, **20**(2), 155–187. http://doi.org/10.1017/S0272263198002034.

Fokes, J., **Bond, Z. S.**, & **Steinberg, M.** (1984). Patterns of English word stress by native and non-native speakers. In *Proceedings of the tenth international congress of phonetic sciences* (pp. 682–686). De Gruyter. https://doi.org/10.1515/9783110884685-111.

Fry, D. B. (1955). Duration and intensity as physical correlates of linguistic stress. *Journal of the Acoustical Society of America*, **27**(4), 765–768. https://doi.org/10.1121/1.1908022.

Gay, T. (1978). Physiological and acoustic correlates of perceived stress. *Language and Speech*, **21**(4), 347–353. https://doi.org/10.1177/002383097802100409.

Hahn, L. D. (2004). Primary stress and intelligibility: Research to motivate the teaching of suprasegmentals. *TESOL Quarterly*, **38**(2), 201–223. https://doi.org/10.2307/3588378.

Hosaka, I., & **Saito, K.** (2026). What promotes success in L2 suprasegmental learning in foreign language classroom settings? The roles of perceptual, cognitive, and experiential individual differences in the acquisition of English word stress. *Applied PsychoLinguistics*, **47**(5), 1–29. http://doi.org/10.1017/S0142716426100460.

Idemaru, K., **Holt, L. L.**, & **Seltman, H.** (2012). Individual differences in cue weights are stable across time: The case of Japanese stop lengths. *The Journal of the Acoustical Society of America*, **132**(6), 3950–3964. https://doi.org/10.1121/1.4765076.

Isaacs, T., & **Trofimovich, P.** (2012). Deconstructing comprehensibility. *Studies in Second Language Acquisition*, **34**(3), 475–505. https://doi.org/10.1017/S0272263112000150.

Kachlicka, M., **Saito, K.**, & **Tierney, A.** (2019). Successful second language learning is tied to robust domain-general auditory processing and stable neural representation of sound. *Brain and Language*, **192**, 15–24. https://doi.org/10.1016/j.bandl.2019.02.004.

Kang, O., **Rubin, D.**, & **Pickering, L.** (2010). Suprasegmental measures of accentedness and judgments of language learner proficiency in oral English. *The Modern Language Journal*, **94**(4), 554–566. https://doi.org/10.1111/j.1540-4781.2010.01091.x.

Kim, Y., **Payant, C.**, **Skalicky, S.**, & **Namkung, Y.** (2026). Comparing the effectiveness of Duolingo, classroom instruction, and classroom + Duolingo instruction conditions on beginner-level French language development. *Studies in Second Language Acquisition*, 1–28. http://doi.org/10.1017/S0272263126101521.

Kochanski, G., **Grabe, E.**, **Coleman, J.**, & **Rosner, B.** (2005). Loudness predicts prominence: Fundamental frequency lends little. *The Journal of the Acoustical Society of America*, **118**(2), 1038–1054. https://doi.org/10.1121/1.1923349.

Kondo, M. (2009). Is acquisition of L2 phonemes difficult? Production of English stress by Japanese speakers. In *Proceedings of the 10th generative approaches to second language acquisition conference (GASLA 2009)* (Vol. **105**, p. 112). Cascadilla Proceedings Project.

Kong, E. J., & **Edwards, J.** (2016). Individual differences in categorical perception of speech: Cue weighting and executive function. *Journal of Phonetics*, **59**, 40–57. https://doi.org/10.1016/j.wocn.2016.08.006.

Lee, H.-Y., & **Hwang, H.** (2016). Gradient of learnability in teaching English pronunciation to Korean learners. *The Journal of the Acoustical Society of America*, **139**(4), 1859–1872. https://doi.org/10.1121/1.4945716.

Lee, A. H., & **Lyster, R.** (2016). The effects of corrective feedback on instructed L2 speech perception. *Studies in Second Language Acquisition*, **38**(1), 35–64. http://doi.org/10.1017/S0272263115000194.

Lee, A. H., & **Lyster, R.** (2017). Can corrective feedback on second language speech perception errors affect production accuracy? *Applied Psycholinguistics*, **38**(2), 371–393. http://doi.org/10.1017/S0142716416000254

Legeris, 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. *Journal of the Acoustical Society of America*, **128**(6), 3757–3768. https://doi.org/10.1121/1.3506351.

Lemhöfer, K., & **Broersma, M.** (2012). Introducing LexTALE: A quick and valid lexical test for advanced learners of English. *Behavior Research Methods*, **44**(2), 325–343. https://doi.org/10.3758/s13428-011-0146-0.

Levitt, H. C. C. H. (1971). Transformed up-down methods in psychoacoustics. *The Journal of the Acoustical Society of America*, **49**(2B), 467–477. https://doi.org/10.1121/1.1912375.

McAllister, R., **Flege, J. E.**, & **Piske, T.** (2002). The influence of L1 on the acquisition of Swedish quantity by native speakers of Spanish, English and Estonian. *English and Estonian. Journal of phonetics*, **30**(2), 229–258. https://doi.org/10.1006/jpho.2002.0174.

Mochizuki-Sudo, M., & **Kiritani, S.** (1991). Production and perception of stress-related durational patterns in Japanese learners of English. *Journal of Phonetics*, **19**(2), 231–248. https://doi.org/10.1016/S0095-4470(19)30219-0.

Mora, J. C. (2022). Aptitude and individual differences. In *The Routledge handbook of second language acquisition and speaking* (pp. 68–82). Routledge. https://doi.org/10.4324/9781003022497-7.

Mora-Plaza, I., **Ortega, M.**, & **Mora, J. C.** (2022). High-variability phonetic training under different conditions: Individual differences in auditory attention control. In M. Pawlak (Ed.), *Second language learning and teaching* (pp. 241–260). Cham: Springer. https://doi.org/10.1007/978-3-030-98218-8_14.

Mora-Plaza, I., **Saito, K.**, **Suzukida, Y.**, **Dewaele, J.-M.**, & **Tierney, A.** (2022). SLA Speech Tools - Tools for Second Language Speech Research and Teaching. http://sla-speech-tools.com.

Moore, B. C. (2012). *An introduction to the psychology of hearing*. Brill.

Mueller, J. L., **Friederici, A. D.**, & **Männel, C.** (2012). Auditory perception at the root of language learning. *Proceedings of the National Academy of Sciences*, **109**(39), 15953–15958. https://doi.org/10.1073/pnas.1204319109

Nation, I. S. P. (2017). *The BNC/COCA level 6 word family lists* (Version 1.0.0) [Data file]. Victoria University of Wellington. http://www.victoria.ac.nz/lals/staff/paul-nation.aspx

Nguyễn, T. A. T., **Ingram, C. J.**, & **Pensalfini, J. R.** (2008). Prosodic transfer in Vietnamese acquisition of English contrastive stress patterns. *Journal of Phonetics*, **36**(1), 158–190. https://doi.org/10.1016/j.wocn.2007.09.001.

https://doi.org/10.1017/S1366728926101576 Published online by Cambridge University Press

Pederson, E., & **Guion-Anderson, S.** (2010). Orienting attention during phonetic training facilitates learning. *The Journal of the Acoustical Society of America*, **127**(2), EL54–EL59. https://doi.org/10.1121/1.3292286.

Peperkamp, S., & **Dupoux, E.** (2002). A typological study of stress "deafness". In C. Gussenhoven & N. Warner (Eds.), *Laboratory phonology 7* (pp. 203–240). Mouton de Gruyter. https://doi.org/10.1515/9783110197105.

Perrachione, T. K., **Lee, J.**, **Ha, L. Y.**, & **Wong, P.** (2011). Learning a novel phonological contrast depends on interactions between individual differences and training paradigm design. *The Journal of the Acoustical Society of America*, **130**(1), 461–472. https://doi.org/10.1121/1.3593366.

Pennington, M. C., & **Ellis, N. C.** (2000). Cantonese Speakers' Memory for English Sentences with Prosodic Cues. *The Modern Language Journal*, **84**(3), 372–389. https://doi.org/10.1111/0026-7902.00075

Petrova, K. Y. (2025). *Mandarin speakers' English lexical stress acquisition: A cross-sectional and longitudinal perspective* (Doctoral dissertation). UCL (University College London).

Port, R. F., **Dalby, J.**, & **O'Dell, M.** (1987). Evidence for mora timing in Japanese. *The Journal of the Acoustical Society of America*, **81**(5), 1574–1585. https://doi.org/10.1121/1.394510.

Rosner, B. S., & **Pickering, J. B.** (1994). *Vowel perception and production*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198521389.001.0001.

Sadakata, M., & **McQueen, J. M.** (2014). Individual aptitude in mandarin lexical tone perception predicts effectiveness of high-variability training. *Frontiers in Psychology*, **5**, 1318. https://doi.org/10.3389/fpsyg.2014.01318.

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

Saito, K., **Petrova, K.**, **Suzukida, Y.**, **Kachlicka, M.**, & **Tierney, A.** (2022). Training auditory processing promotes second language speech acquisition. *Journal of Experimental Psychology: Human Perception and Performance*, **48**(12), 1410–1426. https://doi.org/10.1037/xhp0001042.

Saito, K., & **Plonsky, L.** (2019). Effects of second language pronunciation teaching revisited: A proposed measurement framework and meta-analysis. *Language Learning*, **69**(3), 652–708. https://doi.org/10.1111/lang.12345.

Saito, K., **Sun, H.**, & **Tierney, A.** (2020). Domain-general auditory processing determines success in second language pronunciation learning in adulthood: A longitudinal study. *Applied PsychoLinguistics*, **41**(5), 1083–1112. http://doi.org/10.1017/S0142716420000491.

Saito, K., & **Tierney, A.** (2025). Roles of domain-general auditory processing in second language speech learning revisited: What degree of precision makes a difference? *Studies in Second Language Acquisition*, **75**, 97–138. https://doi.org/10.1111/lang.12722.

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**(1), 187–224. https://doi.org/10.1017/S0142716417000418.

Sugito, M. (1980). Akusento, intonation no hikaku. [a comparison between accent and intonation]. In T. Kunihiro (Ed.), Nichi-ei hikakukooza [*Japanese-English contrastive studies*] (pp. 107–183). Taishuukan.

Tanner, M., & **Landon, M.** (2009). The effects of computer-assisted pronunciation readings on ESL learners' use of pausing, stress, intonation, and overall comprehensibility. *Language Learning & Technology*, **13**(3), 51–65. https://doi.org/10.64152/10125/44191.

Thomson, R. I., & **Derwing, T.M.** (2015) The effectiveness of L2 pronunciation instruction: A narrative review. *Applied Linguistics*, **36**(3), 326–344. https://doi.org/10.1093/applin/amu076

Tierney, A., & **Kraus, N.** (2014). Auditory-motor entrainment and phonological skills: Precise auditory timing hypothesis (PATH). *Frontiers in Human Neuroscience*, **8**, 949. https://doi.org/10.3389/fnhum.2014.00949.

Tsujimura, N. (1996). *An introduction to Japanese linguistics*. Blackwell.

Uchihara, T., **Karas, M.**, & **Thomson, R. I.** (2024). Does perceptual high variability phonetic training improve L2 speech production? A meta-analysis of perception-production connection. *Applied PsychoLinguistics*, **45**(4), 591–623. https://doi.org/10.1017/S0142716424000195.

Uchihara, T., **Karas, M.**, & **Thomson, R. I.** (2025). High variability phonetic training (HVPT): A meta-analysis of L2 perceptual training studies. *Studies in Second Language Acquisition*, **47**, 794–827. https://doi.org/10.1017/S0272263125100879.

Uchihara, T., **Saito, K.**, **Kurokawa, S.**, **Takizawa, K.**, & **Suzukida, Y.** (2025). Declarative and automatized phonological vocabulary knowledge: Recognition, recall, lexicosemantic judgment, and listening-focused employability of second language words. *Language Learning*, **75**(2), 458–492. https://doi.org/10.1111/lang.12668.

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

Wang, Q. (2008). *Perception of English stress by mandarin Chinese learners of English: An acoustic study* (Doctoral dissertation). University of Victoria.

Wang, Y., **Spence, M. M.**, **Jongman, A.**, & **Sereno, J. A.** (1999). Training American listeners to perceive mandarin tones. *The Journal of the Acoustical Society of America*, **106**(6), 3649–3658. https://doi.org/10.1121/1.428217.

Wiener, S., **Chan, M. K.**, & **Ito, K.** (2020). Do explicit instruction and high variability phonetic training improve nonnative speakers' mandarin tone productions? *The Modern Language Journal*, **104**(1), 152–168. https://doi.org/10.1111/modl.12619.

Yang, J., **Li, Y.**, **Wang, L.**, **Sun, B.**, **He, J.**, **Liang, Z.**, & **Wang, D.** (2024). Mobile application-based phonetic training facilitates Chinese-English learners' learning of L2. *Learning and Instruction*, **93**, Article 101967. https://doi.org/10.1016/j.learninstruc.2024.101967.

Yao, Y., **He, M.**, **Chen, F.**, & **Zhu, J.** (2025). A meta-analysis of second language phonetic training: Exploring overall effect and moderating factors. *Journal of Speech, Language, and Hearing Research*, **68**(4), 1784–1802. https://doi.org/10.1044/2024_JSLHR-24-00432.

Yu, V. Y., & **Andruski, J. E.** (2010). A cross-language study of perception of lexical stress in English. *Journal of Psycholinguistic Research*, **39**(4), 323–344. https://doi.org/10.1007/s10936-009-9142-2.

Zhang, Y., & **Francis, A.** (2010). The weighting of vowel quality in native and non-native listeners' perception of English lexical stress. *Journal of Phonetics*, **38**(2), 260–271. https://doi.org/10.1016/j.wocn.2009.11.002.

Zhang, W., **Liao, Y.**, & **Truong, H. T.** (2024). High variability phonetic training facilitates categorical perception of Mandarin lexical tones in L2 older adults: A link to auditory processing. *Language Teaching Research*. https://doi.org/10.1177/13621688241282259.

https://doi.org/10.1017/S1366728926101576 Published online by Cambridge University Press