Training Auditory Processing Promotes Second Language Speech Acquisition

Kazuya Saito[1], Katya Petrova[1], Yui Suzukida[1], Magdalena Kachlicka[2], and Adam Tierney[2]
[1]University College London
[2]Birkbeck, University of London

Abstract[1]

Recent evidence suggests that domain-general auditory processing (sensitivity to the spectro-temporal characteristics of sounds) helps determine individual differences in L2 speech acquisition outcomes. The current study examined the extent to which focused training could enhance auditory processing ability, and whether this had a concomitant impact on L2 vowel proficiency. A total of 98 Japanese learners of English were divided into four groups: (1) Auditory-Only (F2 discrimination training); (2) Phonetic-Only (English [æ] and [ʌ] identification training); (3) Auditory-Phonetic (a combination of auditory and phonetic training); and (4) Control training. The results showed that the Phonetic-Only group improved only their English[æ] and [ʌ] identification, while the Auditory-Only and Auditory-Phonetic groups enhanced both auditory and phonetic skills. The results suggest that a learner's auditory acuity to key, domain-general acoustic cues (F2 = 1200-1600 Hz) promotes the acquisition of knowledge about speech categories (English [æ] vs. [ʌ]).

Key words: Auditory processing; second language acquisition; second language speech; phonetic training; auditory training

[1] We are grateful to all the participants in the current project and the following colleagues for their assistance with data collection: Nobuhiro Kamiya, Hiroko Nakamura, and Noriko Nakanishi. The project was funded by Leverhulme Trust (RPG-2019-039) and Spencer Foundation (202100074).
Correspondence concerning this article should be addressed to Kazuya Saito, University College London, 20 Bedford Way, London, WC1H0AL, United Kingdom Email: k.saito@ucl.ac.uk

Public Significance

Investigating 98 adult Japanese speakers' acquisition of English [æ] and [ʌ], the current study examined whether domain-general auditory processing (i.e., precise representation of sounds) can be improved via focused online training, and whether auditory training can enhance speech learning. Our study shows that improving a learner's auditory processing promotes the acquisition of knowledge about speech categories (English [æ] vs. [ʌ]) establishing auditory processing as a causal factor in language learning. These findings suggest that auditory training could help remediate difficulties with L2 speech learning in some individuals with auditory deficits, and that auditory testing could help predict which individuals are capable of proficient L2 learning. The current study could be considered as the very first attempt to answer one of the most well-researched topics (i.e., the mechanisms underlying language learning) by interfacing psycholinguistics, education, and hearing research perspectives in an interdisciplinary manner.

Introduction

In the field of cognitive psychology, much attention has been given to examining the perceptual and cognitive systems which govern first language acquisition. One influential theoretical account proposes that domain-general auditory processing (representation of acoustic dimensions) is a critical determinant of language learning throughout the lifespan, as it facilitates phonetic, lexical, and morphosyntactic analysis of language (Goswami, 2015). A paradigm has emerged which states that auditory processing plays an even more critical role in *second language* (L2) acquisition (Mueller et al., 2012). Studies conducted within this paradigm have provided (a) cross-sectional evidence showing that high-level L2 speech proficiency is linked to both language experience and precise auditory processing abilities (e.g., Kachlicka et al., 2019); and (b) longitudinal evidence showing that individuals with precise auditory sensitivity demonstrate larger L2 gains when they use the target language in both naturalistic and classroom settings (e.g., Sun et al., 2021).

Currently, more research is needed to establish the directionality of the relationship between auditory processing and learning outcomes. Specifically, although it is clear that auditory processing and language learning are linked, it is difficult to determine whether precise encoding of domain-general auditory information *drives* language learning, or whether enhanced auditory precision is a *by-product* of language learning. To this end, one exploratory question concerns to what degree auditory processing can be enhanced via focused training, and whether it would subsequently impact speech learning. In the current investigation, we examined a total of 98 Japanese speakers' acquisition of English [æ] and [ʌ] contrasts with a pre- and post-test design. Using paradigms developed in L2 phonetics (e.g., Logan et al., 1991 for High Variability Phonetic Training), and L1 hearing research (e.g., Merzenich et al., 1996 for auditory discrimination training), we developed two different types of focused training: (a) the identification of target contrasts using multi-talker speech stimuli (i.e., phonetic training) and (b) the discrimination of the relevant acoustic cues (second formant frequency in the range 1200-1600 Hz) within nonverbal sounds (i.e., auditory training). To unravel the relative effectiveness of the methods, we aimed to compare three different types of training (Auditory-Only, Phonetic-Only, and Auditory-Phonetic) against a comparison group who received phonetic training on different targets (English [r] and [l]).

Background

Auditory Processing in L1 Acquisition

Auditory processing is defined as the ability to encode and proceduralize spectral and temporal characteristics of sounds. This domain-general ability has been proposed to be a bottleneck for spoken language acquisition: spectral and temporal details convey phonemic, phonological, and prosodic categories (Werker, 2018) while pitch, amplitude, and duration cues also contain information relevant to detection of word-, collocation-, sentence-, and phrase boundaries (Cutler & Butterfield, 1992), suffixes, inflection, and articles (Joanisse & Seidenberg, 1998) and word order (Penner et al., 2001). Individuals widely differ in the precision of their auditory discrimination (e.g., Kidd et al., 2007), and auditory skills are associated with a range of language outcomes (e.g., speech-in-noise perception, vocabulary use, literacy, and phonological awareness; Anvari et al., 2002; Bavin et al., 2010; Boets et al., 2008; Douglas & Willatts, 1994; Lamb & Gregory, 1993; Talcott et al., 2000; Tierney et al., 2021; for longitudinal evidence on the link between auditory processing and L1 vocabulary development over the first 3 years of life, see Kalashnikova et al., 2019).

Futhermore, some toddlers with auditory processing deficits experience delays in phonological, lexical, and morphosyntactic learning, and these delays can lead to a range of language problems, including dyslexia (Hornickel & Kraus, 2013) and other disorders (e.g., Russo et al., 2008 for Autism Spectrum Disorders). A meta-analysis conducted by Hämäläinen et al., (2013) found that differences in the auditory processing abilties of dyslexic and normal hearing individuals are relatively medium-to-large (d = 0.9 for duration; d = 0.8 for risetime; d = 0.7 for pitch; cf. Witton et al., 2020 for the results of the moderator analyses). It is noteworthy, however, that not all dyslexic children have auditory deficits (about 40% according to Ramus, 2003), and that those with auditory deficits may have problems in other areas of cognitive functioning (e.g., Snowling et al., 2018 for attentional control). As a result, individual differences in auditory processing may help drive variability in language learning success, but are not the only potential contributor to language impairments such as dyslexia. Recently, some scholars have begun to examine the role of audiotory processing in the context of post-pubertal L2 speech learning—a unique testing ground for the life-long role of auditory processing in langauge learning (Mueller et al., 2012).

Auditory Processing in L2 Speech Learning

Post-pubertal learners demonstrate a great deal of individual variation in their L2 speech outcomes, with some achieving nativelike proficiency and others retaining a strong foreign accent. Much of the research on the ultimate attainment of L2 speech has focused on demographic factors as predictors of learning success (e.g., length of immersion, Flege et al., 1995; timing of immersion, Abrahamsson & Hyltenstam, 2009; frequency of L2 use, Derwing & Munro, 2013; native vs. non-native interlocutors, Flege & Liu, 2001; and classroom vs. immersion learning context, Mora & Valls-Ferrer, 2012). Other scholars have investigated learner-internal perceptual-cognitive predictors (e.g., Darcy et al., 2015 for working memory; Linck et al., 2013 for implicit and statistical learning abilities; (Mora-Plaza et al., 2021) for attention and switching; Ghaffarvand Mokari & Werner, 2019 for inhibitory control; Hu et al., 2013 for phonemic coding). One *perceptual* ability which has attracted increasing research attention is auditory processing ability (e.g., Mueller et al., 2012). Some scholars have argued that having precise auditory processing could be even *more* consequetial in adult L2 speech acquistion compared to L1 speech acquistion (Saito et al., 2020b)because the former takes place in the same linguistic space as the fully developed and automatized L1 system. This implies that

L2 speech learning is susceptible to the influence of L1 phonetic structures. When adult learners are exposed to new sounds, they analyze acoustic signals using existing L1 cue strategies, at least in the initial stages of learning (e.g., Japanese speakers using F2 variation and largely ignoring F3 variation for perceiving English [r] and [l]). Auditory deficits could directly impact the extent to which learners are able to retune their cue weighting patterns, which could in turn lead to learning difficulties and delays. Furthermore, unlike L1 acquisition, where children generally enjoy ample opportunies for language input, the quantity and quality of L2 conversational opportunites are substantially limited, even in immersive settings (Flege & Liu, 2001). Thus, the slightest auditory disadvantage may hinder learners from making the most of precious opportunities for the robust, precise, and prompt analysis of acoustic input.

To date, research has shown that individuals with precise auditory processing abilities are better able to perceive foreign sounds that they have never encountered (e.g., Kempe et al., 2015), and are able to attain highly advanced L2 proficiency after years of immersion experience (e.g., Kachlicka et al., 2019 for grammar; Saito, et al., 2020 for speech production; Omote et al., 2017 for speech perception; for the longitudinal relationship between auditory processing and one year of L2 speech learning, see Saito et al., 2020b; Sun et al., 2021). Moreover, gains from explicit phonetic training in various areas of L2 speech learning have been linked to L2 learners' auditory processing profiles (e.g., Lengeris & Hazan, 2010 for formant acuity vs. English vowels; Qin et al., 2021 for pitch acuity vs. Cantonese lexical tones; Leong et al., 2018 for amplitude acuity vs. L2 segmental acquisition under adverse conditions).

Although the existing literature strongly supports the association between individual differences in basic auditory perception skills and post-pubertal L2 speech learning, the *directionality* of this audition-language link remains unclear. Specifically, we have yet to know (a) whether one's auditory precision predicts the rate of success in subsequent language learning; (b) whether one's specific experience with language acquisition shapes the degree of auditory acuity; or (c) whether the development of auditory and linguistic abilities co-evolve. In addition, it is possible that a third factor (i.e., other cognitive abilities) may be associated with both auditory processing and L2 speech learning, and thus be responsible for driving acquisition.

One way to approach this topic is to investigate how auditory processing training impacts the development of auditory *and* linguistic abilities. This would have ample theoretical and practical implications. For theory building, it would provide some of the first empirical evidence on the causal link between perception and acquisition (i.e., that enhanced auditory processing promotes L2 speech learning even without any phonetic training). In terms of practical relevance, the study could have important implications for how to stimulate the language development of L2 learners with relatively low auditory processing abilities. Research on the development, validation, and provision of auditory training is a high priority, to determine whether such training could help all L2 learners notice, learn, and master new sounds regardless of their auditory processing profiles. The current study took a first step towards pursuing this scholarly enquiry by exploring the effects of auditory processing training on L2 speech acquisition.

Training Auditory Processing

In the field of first language acquisition, auditory processing training has been explored as a remedy to language problems that arise from deficient auditory processing skills (Goswami, 2015; Merzenich et al., 1996). In such a paradigm, learners are asked to discriminate among and between non-verbal and/or speech sounds which vary along a particular acoustic dimension (e.g., formant, pitch, duration, or amplitude). Doing so induces learners to selectively focus on

analyzing spectro-temporal information while paying less attention to other dimensions of sounds. To promote the transfer of auditory learning gains to language acquisition, some scholars have proposed combining auditory processing training with language learning training (e.g., phonemic discrimination, reading training; for auditory training studies examining the effects of phonemic discrimination under noise conditions on adults with hearing loss, see Henshaw & Ferguson, 2013; for music training studies examining the effects of music practice on auditory and language abilities among children, see Tierney et al., 2015; for elderly adults with age-related hearing loss, see Dubinsky et al., 2019).

Table 1

Summary of 10 Key Auditory Training Studies

Study	Participants	Focus, length, and stimuli of training	Findings
Merzenich, et al. (1996)	• *n* = 7 language-based learning impairment LI children (Study 1) • 11 LLI children (Study 2)	• Focus: Auditory skills (temporal processing) • Stimuli: Nonverbal and speech sounds • Length: 20 minutes over 19-28 sessions (6-9 hours)	• Training helped children with auditory deficits improve both auditory processing (temporal reproduction) and language functions (phonemic identification).
Hayes, et al. (2003)	• *n* = 27 children with learning problems (e.g., attention deficit) for Experimental • *n* = 22 children (with/without learning problems) for Control	• Focus: Global auditory *and* language skills via phoneme discrimination, auditory memory, auditory sequencing, auditory attention, rhyming and sound blending skills (using Earobics) • Stimuli: Nonverbal and speech sounds • Length: 1-hour × 25-40 sessions over 8 weeks (25-40 hours)	• Children with language learning problems improved composite auditory processing scores, but not cognitive and academic abilities. • Alternation was observed in the cortical representation of speech in quiet and in noise (gateway to sensory input). • No changes were observed in brainstem responses.
Schaffler, et al. (2004)	• *n* = 41 children with dyslexia	• Focus: Auditory skills (pitch and intensity discrimination) • Stimuli: Nonverbal sounds • Length: 10-15 minutes × 10 sessions over 10 days (1.5-2.5 hours)	• Children's auditory processing skills improved. • Gains were transferred to phonological and spelling skills.
Strehlow, et al. (2006)	• *n* = 44 children with dyslexia	• Focus: Reading training plus auditory temporal processing or phoneme processing (duration) • Stimuli: Nonverbal sounds for auditory training; speech sounds for phoneme processing training • Length: 2 hours per day × 10-12 weeks (100-120 hours)	• Auditory processing and phonemic skills were enhanced; improvements lasted for six months. • The generalizability of the gains to reading and spelling was unclear (although some advantage was observed for auditory processing training group in the delayed post-tests).
Micheyl, et al. (2006)	• *n* = 64 (30 musicians, 38 non-musicians)	• Focus: Pitch discrimination task without feedback (pure/harmonic tones; ± noise) • Length: 1.5-2.5 hours × 5 sessions over 5 days (7.5-12.5 hours) • Stimuli: Nonverbal sounds (pure and harmonic tones)	• Musicians (10 hours of training) demonstrated more precise pitch discrimination than non-musicians. • Four to eight hours of training helped non-musicians obtain pitch acuity that was comparable to musicians.
McArthur, et al. (2008)	• *n* = 28 children with specific language	• Focus: Auditory skills (frequency discrimination, rapid auditory processing)	• Training improved auditory processing skills on immediate and delayed post-tests (1 week and 3 months after training).

	impairment or specific reading disability	vs. phonemic skills (vowel, consonant-vowel discrimination) • Stimuli: Nonverbal sounds for auditory training; speech sounds for phoneme processing training • Length: 30 minutes × 24 sessions over 6 weeks (120 hours)	• Although language development was observed (reading, language, spelling), it could have been ascribed to test-retest effects.
Whitton, et al. (2017)	• *n* = 24 elderly hearing-impaired participants (*M* = 70 years)	• Focus: Audio-motor integration skills (discrimination and motor action) or auditory working memory • Stimuli: Nonverbal sounds (different in intensity, pitch, formant or noise) • Length: 3.5 hours × 8 weeks (27 hours)	• Auditory motor training enhanced both auditory processing (auditory discrimination; digit and word identification in noise) and language skills (speech-in-noise perception). • Training gains were not sustainable on the delayed post-tests (6 weeks after the treatment). • Working memory training enhanced relevant cognitive abilities (working memory capacities) but not language skills.
Whiteford & Oxenham (2018)	• *n* = 20 adults with congenital amusia and 20 age-matched controls	• Focus: Pitch discrimination task or comparison task (interaural level difference training) • Stimuli: Pure tones (500 Hz) • Length: 1-2 hours × 4 sessions	• Both training groups improved pitch discrimination abilities (i.e., test-retest effects). • 11 out of 20 participants attained melody discrimination beyond the global diagnostic criteria for amusia (measured via the Montreal Battery of Evaluation of Amusia training).
Pires & Schochat (2019)	• *n* = 25 children with spelling disorders	• Focus: Auditory temporal training (using four games adapted from the Fast ForWord software) • Stimuli: Nonverbal and modified speech sounds • Length: 30 minutes × 8 sessions (4 hours)	• Auditory training group not only improved auditory temporal ordering skills, but also reduced the frequency of phonological-based orthographic errors.
Fostick, et al. (2020)	• *n* = 86 elderly normal hearing adults (60-83 years)	• Focus: Auditory temporal processing training vs. intensity discrimination training • Stimuli: Nonverbal pure tones (differing in interstimulus intervals for temporal training and intensity for intensity training) • Length: 12 sessions	• Auditory training positively impacted relevant abilities (temporal training for temporal thresholds, intensity training for intensity thresholds) • Auditory temporal training (but not intensity discrimination training) helped enhance aging adults' speech perception (word identification under quiet and noise conditions) at immediate and delayed post-tests (1 day and 3 months after training)

As summarized in Table 1, findings from the existing literature suggest that children and adults can enhance both auditory and language abilities when they receive both auditory and language training in a complementary fashion (e.g., Pires & Schochat, 2019). Such training could be very brief (e.g., Whiteford & Oxenham, 2018 for about 4 hours). There is some evidence that the effectiveness of training can last several months (e.g., Strehlow et al., 2006 for the results of six-month delayed post-tests). When it comes to the transferability of auditory training to language domains (e.g., speech-in-noise perception; word recognition; pseudoword repetition; spelling accuracy), however, the findings are mixed. For example, positive transfer is likely to happen when auditory training involves both nonverbal and speech sounds (e.g., Merzenich et al., 1996 for phonemic identification; Whitton et al., 2017 for speech-in-noise perception; Pires & Schochat, 2019 for spelling error reduction). Some research has shown that the use of nonverbal stimuli in auditory processing training can still lead to robust gains in language (e.g., Schäffler et al., 2004 for phonological and spelling skills). However, skill specificity has also been reported. In some studies, auditory training benefits have been limited to auditory processing enhancement, but without any significant impact on other dimensions of cognitive, language, and academic abilities (e.g., Hayes et al., 2003 for the lack of cognitive and academic benefits; McArthur et al., 2008 for the lack of reading, language, and spelling benefits). In terms of methods, whereas most of the literature used similar tasks (nonverbal sound or/and speech discrimination), the length and intensity of training varied to a great degree (e.g., Schäffler et al., 2004 for 1.5-2.5 hours; Strehlow et al., 2006 for 120 hours). Finally, the effectiveness of auditory processing training (measured via auditory development and language transfer) has been examined in a range of populations (e.g., normal-hearing children and adults; children with auditory deficits and dyslexia; elderly adults with hearing loss).

Motivation for Current Study

The current study took a first step towards examining the extent to which the provision of auditory processing training can enhance adult L2 speech learning. Unlike the existing literature, which looked at *broad* relationships between auditory processing training and global areas of language (e.g., sound and word identification in noise), a novelty of the current study is our scrutiny of a specific L2 speech acquisition instance—Japanese speakers' acquisition of English [æ] and [ʌ]. These two vocalic sounds mainly differ in terms of F2 (1100-1300 Hz vs. 1400-1600 Hz; Hawkins & Midgley, 2005). Because neither of the phones exist in Japanese, inexperienced Japanese listeners use two main strategies to perceive them—(a) perceiving English [æ] as a new sound (which is sufficiently distinguishable from any neighboring L1 sounds, Japanese [e] and [a]); and (b) assimilating English [ʌ] to the Japanese central vowel [a] (Nishi et al., 2008). Some studies have shown that brief phonetic training on these features (2-4 hours) can lead to learning gains in the range of 10-20% (e.g., Lambacher et al., 2005; Nishi & Kewley-Port, 2008; for similar findings, Ortega et al., 2019 with Spanish learners; and Thomson, 2012 with Chinese learners).

A total of 98 Japanese learners participated in the current investigation. They were divided into four treatment conditions—(a) Auditory-Only (3 hours of non-verbal auditory processing training on the primary acoustic correlate of the English [æ] and [ʌ] contrast, F2 = 1200-1600 Hz), (b) Phonetic-Only (3 hours of phonetic training on English [æ] and [ʌ]), (c) Auditory-Phonetic (1.5 hours of non-verbal auditory and phonetic training, respectively), and (c) Control (3 hours of phonetic training on English [r] and [l]). To examine the impact of training on the development of auditory processing *and* speech proficiency, the participants took speech

perception (English [æ] and [ʌ] identification) *and* auditory processing tests (F2 discrimination) before and after the treatment. The following research question and predictions were formulated:

- **Research Question**: To what degree can three different types of training (Auditory-Only, Phonetic-Only, Phonetic-Auditory) impact participants' F2 processing abilities and English [æ] and [ʌ] identification abilities?
- **Predictions**: Echoing the assumptions underlying the auditory precision hypothesis in L1 (Goswami, 2015) and L2 contexts (Kachlicka et al., 2019), we hypothesized that auditory processing would promote L2 speech acquisition. Thus, an asymmetric relationship was predicted: (a) Participants in the Phonetic-Only group will improve their English [æ] and [ʌ] perception skills (e.g., gains in the range of 10-20%; Lambacher et al., 2005) but not their F2 discrimination skills; (b) participants in the Auditory-Only group will improve their F2 discrimination skills, and this will exert a positive influence on their English [æ] and [ʌ] perception accuracy. Given some positive outcomes in the literature (e.g., Merzenich et al., 1996), we predicted that the combination approach (Auditory-Phonetic) would be expected to promote both auditory and phonetic learning. However, we did not have particular predictions as to whether participants in the Auditory-Phonetic group would *outperform* those in the Auditory-Only or Phonetic-Only groups.

Method

Participants

A total of 98 Japanese speakers of English participated in the current study. Due to the constraints of the global COVID-19 pandemic, recruitment and data collection were conducted online. A digital flyer was created to recruit volunteer participants who were interested in improving their L2 English listening and speaking proficiency via research-based training methods. The flyer was disseminated across major social media networks (e.g., Facebook, Twitter), on the investigators' university websites, and through mailing lists consisting of more than 500 Japanese researchers and teachers inside and outside of Japan. Initially, 108 interested participants contacted the researchers, and received detailed instructions about the scholarly purpose and nature of the project. During the orientation session (via email; for details, see below), the participants were explicitly told that they had to secure eight consecutive days during which they would be able to complete all of the treatment sessions (as shown in Figure 1). Participants who confirmed their commitment to the project were invited to join. For a range of technological and logistical reasons (for the details of the screening procedure, see the Results section), the data of 10 participants were not used in the final analyses (10% attrition). In total, 98 participants completed the project (32 males, 66 females).

Their biographical backgrounds varied in terms of chronological age (M_{age} = 23.8 years, *SD* = 8.1, *Range* = 18-59), starting age of English education ($M_{age\ of\ learning}$ = 11.6 years, *SD* = 4.0, *Range* = 3-15 years), and experience with pronunciation (*n* = 24 for yes, *n* = 54 for no) and music training (*n* = 35 for yes, *n* = 63 for no). Although all the participants were based in Japan at the time of the project, some reported having prior immersion experience (48 out of 98 reported more than three years of study- and living-abroad).[2] All participants were randomly assigned to the four treatment conditions: Auditory-Only (*n* = 22), Phonetic-Only (*n* = 22), Auditory-Phonetic (*n* = 21), and Control (*n* = 33). The purpose of the control group was to check the presence of test-retest effects; to ensure robust statistical power, we intended to recruit a sufficiently large number of participants relative to the experimental groups.

According to the results of the previous meta-analyses on L2 speech training, medium-to-large effect sizes were reported (e.g., *d* = .92 in Sakai & Moorman, 2018). Using G*Power (Faul et al., 2007), therefore, a priori power analysis was performed with an estimation of the medium-to-large effect size (*f* = .35). *N* = 112 was suggested as an adequate number of participants for the research design with one between-subjects factor (Group: Auditory-Only, Phonetic-Only, Auditory-Phonetic, and Control) and one within-subjects factor (Time: pre, post-tests) to reach strong statistical power (.96). In light of the number of participants which we actually included in the final analyses (*N* = 98), the compromise power analysis generated slightly lower power, .94. We considered the figure (*N* = 98 for power of .94) to be adequate to guarantee the validity of the current analyses and findings because it was substantially beyond the field-specific threshold in instructed L2 speech research (i.e., power of .70-.80; Larson-Hall, 2015).

[2] Using a dummy code for participants' prior immersion experience backgrounds (0 = little experience abroad, 1 = more than one year of immersion experience), a set of follow-up analyses were conducted

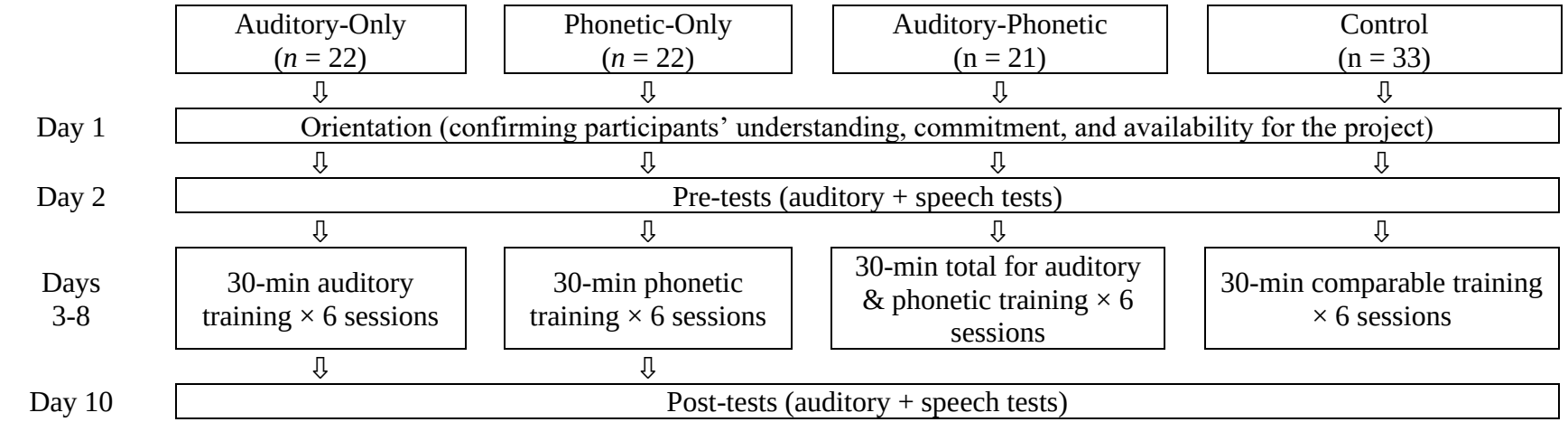


Figure 1

Summary of Research Design

Orientation Session

The pre-/post-tests and training sessions were conducted online via the online experiment platform, Gorilla (Anwyl-Irvine et al., 2020). Throughout the project, participants engaged in each activity using their own computers with internet access. To monitor participants' performance at each stage of the project (pre-test, training, and post-test) and to provide individualized support, each participant was assigned to one of the three investigators as a "tutee". To ensure that participants accessed the equipment and facilities necessary for the online testing and training sessions (computer or laptop, headsets/earphones, and a quiet room with a stable Internet connection), and that they familiarized themselves with the online platform, individual orientation sessions were set up between tutors and tutees via email or via a videoconferencing tool. In these sessions, participants were briefed on the purpose and content of the project (improving L2 English phonological skills via different types of research-based training methods) and were asked to complete a short practice session. The purpose of this session was to acquaint them with the test and training procedures through the completion of a few practice trials comprising nonverbal and speech sounds that were not used in the main study. Only after participants confirmed that the materials worked without any technological problems, were they allowed to proceed to the main part of the project.

Auditory-Only Training

Participants in the Auditory-Only group received a total of six 30-minute auditory processing training sessions. The goal of the auditory processing training was to help enhance participants' capacity to discriminate between certain ranges of frequency variation in second formant frequency (1200-1600 Hz). As described in detail below, the sounds were *artificial*, as they comprised flat F0 and formant contours (within trials), a flat harmonic spectrum, and only two formants. F2 was manipulated to vary from 1200 to 1600 Hz. Our hypothesis was that exposure to such artificial sounds could help participants enhance their domain-general spectral sensitivity (particularly in the range of 1200-1600 Hz), and that this would subsequently impact their domain-specific speech perception abilities of the English [æ] and [ʌ]. However, participants were not exposed to English [æ] and [ʌ] tokens throughout the training.

The unique methodological feature of auditory processing training was that the target acoustic parameter was manipulated in the context of simple, monotonous, and non-speechlike stimuli. The intention was to guide L2 learners to solely focus on the target parameter (F2 variation) in an acoustically abstract space, where they have never established any perceptual strategies. From a conceptual and methodological standpoint, the operationalization of auditory training here differs significantly from phonetic training, which involves distinguishing natural speech sounds based on multiple cues (the height, transition and distance of F1, F2, and F3 for English [æ] and [ʌ]). Auditory processing training also departs from synthesized speech training, wherein learners engage in repeated exposure to the same phonemes that differ in a particular acoustic parameter (e.g., Iverson et al., 2005 for enhanced F3 variation in English [r] and [l] tokens). Although learners are supposed to work on the target acoustic information by adjusting their existing cue weighting strategies, they may encounter difficulty doing so because the acoustic manipulation is embedded in the context of phonemes with varied pitch height and contour, and with multiple formants. Thus, these cues may not be perceptually salient. With such acoustically complex, rich signals, L2 learners may be induced to use interlanguage strategies that they have already established to perceive phonemes without using the target acoustic information (e.g., Japanese listeners using F2 and duration variation to perceive English [r] and [l]; Ingvalson et al., 2011).

As in McArthur et al. (2008), the auditory training took the form of an A x B discrimination task. As shown in Figure 2, participants were asked to choose whether the first or third sound (target stimuli) was different from the second stimulus (standard stimulus). In each trial, the target and standard stimuli differed only in terms of the primary correlate of English [æ] and [ʌ] (F2 variation = 1200-1600 Hz). Participants were asked to complete each training session in one sitting. 24 hours were required between sessions. Their daily performance was recorded and monitored via the Gorilla platform. Once the tutor checked the successful recording of the data, participants received a confirmation email and a reminder for the next session. If participants did not complete a session within 24 hours, they would be sent a reminder and given another 24 hours to complete the sessions. If they still did not complete the session, their data were excluded from the dataset.

Stimulus. The stimuli for the Auditory-Only training were synthesized using a custom MATLAB script (The MathWorks, Inc., Natick, Massachusetts). The tones were 40-harmonic complex tones with a 15-ms on-off cosine ramp, equal amplitude across harmonics, and three formants imposed using a parallel formant filter bank (F1 = 500 Hz, F2 variable (as outlined below), F3 = 2500 Hz; Smith, 2007). The resulting sound was clearly *artificial* and did not sound like natural speech (neither English [æ] nor [ʌ]). The second formant frequency, pitch and duration were varied in the following way. The F2 values ranged in 200 equal mel-scale steps from 1200 to 1600 Hz (the target parameter). The selected range was based on the average F2 values of Hawkins and Midgley's (2005) acoustic analyses of male and female British English speakers' English [æ] and [ʌ] production. As shown in Table 1 in Hawkins and Midgley, individuals likely used diverse F2 thresholds somewhere between 1200 and 1600 Hz to distinguish English [æ] and [ʌ] due to their anatomical differences (e.g., length of vocal tract). Our assumption here is that this F2 range (1200-1600 Hz) roughly corresponds to the variation of English [æ] and [ʌ] produced by native speakers in real life contexts. To train participants to focus on the relevant dimension (F2) while ignoring irrelevant dimensions, the stimuli also varied along two distracter parameters: pitch (in 10 equal mel-scale steps from 70 Hz to 150 Hz) and duration (in 10 equal steps from 80 ms to 220 ms). In total, there were 20,000 stimuli (200 F2 steps × 10 pitch steps ×10 duration steps). Stimulus presentation was scripted such that variability in F0 and duration was present only between trials. Within-trial variability was confined to changes in F2 (1200-1600 Hz), with standard and target tones differing only along the F2 dimension while all other acoustic parameters were held constant.

Procedure. There were a total of 200 trials for each session, divided into four blocks (50 trials per block). Three stimuli were presented at each trial. The second stimulus was always the standard, and the target stimulus was always either the first or third tone. The F2 frequency of the standard stimulus was pseudo-randomly selected from the target range (1200-1600 Hz) using an adaptive staircase procedure (Levitt, 1971), whereby stimuli presentation was adapted to participants' performance. Namely, for the initial trials, the standard stimuli were drawn from the distributional ends of the generated acoustic space. For each target dimension (200 steps), the distracter parameters were randomly selected (i.e.,10 pitch steps [70-150 Hz] and 10 duration steps [80-220 ms]). Regardless of the setting of the distracter parameters, the focus of the training was always accurate discrimination of the sounds in terms of the target parameter (i.e., F2 = 1200-1600 Hz).

As participants progressed through training and the task difficulty increased more standard stimuli were drawn nearer to the centre of the distribution. As outlined in the previous section, variability along the distracter parameters (pitch and duration dimensions) was limited to

between-trial stimuli presentation. Pitch and duration values did not vary between standard vs. target stimuli within each trial. In practical terms, successful trial-by-trial performance required participants to keep track of differences between the standard and target tones that were defined only along the dimension of interest (F2). As such, the training was designed to find how small of a difference in the target F2 parameter participants could hear (while ignoring the variation in the distracter parameters). To promote learning, participants were provided with trial-by-trial feedback (correct or incorrect). Further, the training platform tracked, recorded, and displayed participants' F2 discrimination abilities on a 100-point scale. Larger values indicated more precise auditory processing abilities. Since the target acoustic dimension of the training stimuli varied in 200 equal mel-scale steps (F2: 1200-1600 Hz), the distance between the standard and target stimuli was first recorded on a 200-point scale, and then reversed and converted to a 100-point scale (where higher numbers indicated better performance). Finally, they were given their average accuracy score for the training session. Participants were also provided with their daily average accuracy scores at the end of each training session. Each session was designed to last between 25-35 minutes.

Figure 2
Screenshot of Auditory Processing Training. The task instruction was provided at the upper end of the screen ("Which sound was different? Select 1 or 3"). Trial-by-trial written feedback was provided ("correct" in green or "incorrect" in red color). Participants' current auditory sensitivity level (larger values indicate greater auditory sensitivity on a 100-point scale) was displayed at the right upper corner of the screen ("Level 15").

Phonetic-Only Training

Using a format similar to high variability phonetic training (e.g., Ortega et al., 2019; Thomson, 2012), participants in the Phonetic-Only group received a total of three hours of identification training on natural English [æ] and [ʌ] tokens produced by multiple talkers. Upon hearing each stimulus, they were asked to choose whether it corresponded to [hæ] from "hat" or [hʌ] from "hut." As shown in Figure 3A, each target word was presented with a corresponding visual, with the target vocalic sounds being underlined. Both of the target words ("hat, hut") fell within the first 3000-word families (Cobb, 2021). Thus, participants were assumed to be already familiar with these words, thus minimizing the effects of lexical frequency on L2 speech perception (i.e., L2 listeners tend to show more difficulty in perceiving new sounds when they are embedded in infrequent and unfamiliar words; Flege et al., 1996).

This kind of training was hypothesized to facilitate L2 speech learning for the following reasons. First, by exposing participants to English [æ] and [ʌ] exemplars produced by multiple talkers, it was assumed that they would learn how to attend to between-category variation (e.g., F2 at around 1200 Hz for English [æ] vs. 1600 Hz for English [ʌ]) while ignoring within-category variation (e.g., English [æ] produced by males vs. females). This fine-tuning process is believed to help learners develop more robust and generalizable phonetic representations (see Barriuso & Hayes-Harb, 2018). Second, by exposing participants to target sound syllables repeatedly ([hæ] and [hʌ] out of lexical contexts), it was assumed that their attention would be explicitly drawn to target phonemic accuracies which would otherwise be difficult to notice. As the semantic (rather than phonological) aspects of words are generally prioritized in speech perception, this kind of intervention is helpful to push learners to attend to phonetic units of word knowledge, and improve their phonetic representations (see Thomson, 2012 for the use of open speech syllables).

As with the auditory processing training, all the training materials were incorporated into the Gorilla platform (Anwyl-Irvine et al., 2020) so that participants could complete the training

using their own equipment. They were given 24 hours to complete each session with a maximum 24-hour interval (Days 3-8). Since their progress and in-session performance (accuracy, reaction time) was recorded in the Gorilla platform, participants received an email reminder about the next section upon completion of each session. If participants did not complete a session within 48 hours (24 hours plus another 24-hour extension), their data were eliminated from the analyses.

Stimuli. A total of six native speakers of British English (3 males [M1, M2, M3], 3 females [F1, F2, F3]) read "hat" and "hut" ten times. Once the best exemplars were chosen, the first open syllables ([hæ], [hʌ]) were excised, normalized for amplitude, and saved as WAV files.

Procedure. There were a total of 192 trials (6 talkers × 2 target phonemes [æ, ʌ] × 16 repetitions) for each 30-minute training session. In each trial, participants were asked to identify which word they had heard ("hat" vs. "hat"). Trial-by-trial feedback was provided for both accuracy (i.e., correct or incorrect; see Figure 3A) and reaction time (i.e., how fast they provided responses; see Figure 3B). At the end of each session, participants were shown average accuracy and reaction time. Accuracy scores were used as the index of in-session performance for the statistical analyses. Each session was designed to last between 25-35 minutes.

3A: Accuracy feedback (per trial)

3B: Response time feedback (per trial)

Figure 3

Screenshot of Phonetic Training. In the first screen (3A), the task instruction at the top asked participants to indicate which sound was played ("Which sound did you hear? Select 1 or 3"). Trial-by-trial feedback was provided (check mark in green for correct response) together with averaged total accuracy scores (% between the two visuals). In the second screen (3B), participants received feedback on their response time ("your response speed was x ms").

Auditory & Phonetic Training

It has been suggested that a combination of phonetic and auditory training can help participants transfer gains from auditory processing training to language outcomes (e.g., Fast ForWord, Earobics; Merzenich et al., 1996). Thus, the participants in this group spent the first half of each session on auditory training (15 minutes for 100/200 trials) and the other half on phonetic training (15 minutes for 100/200 trials). Following the methods in previous literature (e.g., Hayes et al., 2003), and given that enhanced auditory abilities were hypothesized to transfer to language development (Merzenich et al., 1996), participants took the auditory training, followed by the phonetic training. This allowed us to examine the extent to which the combined method could maximize the effectiveness of two different types of training (auditory vs. phonetic training).

Control Training

Participants in this group engaged in comparable phonetic training using a smartphone application, but with a focus on a different phonological contrast (English [r] and [l]). They spent 30 minutes on this phonological contrast (without any focus on English [æ] and [ʌ] throughout the training). The stimuli in the control training comprised four open speech syllables (English [ra], [la], [da], [ga]) produced by four different talkers. Following the incidental and multimodal L2 speech learning paradigm (Lim & Holt, 2011), the training was operationalized as a clay target shooting game. Participants were instructed to shoot a range of flying objects as fast as possible to earn more points. Each object entailed a unique color (e.g., red, blue, yellow, blue), movement (e.g., upward, rightward, leftward), and sound (e.g., English [ra], [la], [da], [ga]). While playing the game, they were guided to learn the combination of phonological and visuomotor cues.

In a different venue (Saito et al., in press), the control group's pre- and post-test scores of English [r] and [l] were analyzed to examine the effectiveness of such speech training on the development of Japanese speakers' English [r] and [l] proficiency. In the current study, their pre- and post-tests scores of English [æ] and [ʌ] (which they did not practice during the training) were used to index test-retest effects in the current study. While the control training was delivered via a smartphone, the control participants took both speech and auditory tests through the online platform, Gorilla.

Auditory Processing Tests

Using the same A×B format as the auditory training (but without feedback), this test was designed to measure participants' auditory acuity towards the target parameter (F2 variation = 1200-1600 Hz). The participants took the auditory processing tests as pre- and post-tests to examine the extent to which their F2 sensitivity changed over time due to the auditory and phonetic training.

Stimuli. A total of 200 stimuli were taken from the auditory processing training materials to create a F2 variation continuum. The tones varied in 200 equal mel-scale steps from 1200 to 1600 Hz. Duration, F1, and F3 were fixed at 100 ms, 478 Hz, and 2371 Hz, respectively.

Procedure. Three tones were presented in an A×B format. In each trial, participants were asked to identify the acoustically odd tone by either clicking a button marked "1" or "3". While the second tone was always the standard stimulus, the target stimulus could be either the first or third sound. The answer "1" was correct if the sequence was Target, Standard, Standard; and "3" if it was Standard, Standard, Target. The standard stimulus was set randomly on a trial-by-trial basis to a value anywhere within the continuum, while the target stimuli were always above the

standard stimulus (with the constraint that the standard level needed to be low enough that the target level fell within the continuum).

The initial difficulty of the task was relatively low, with a large distance between the standard and target stimuli (100 steps along the target continuum). Task difficulty was manipulated via an adaptive staircase method in order to identify the minimum difference between the stimuli that the participants could hear (Levitt, 1971). In this method, the distance between the standard and target stimuli decreased after two consecutive correct answers, but increased after a single incorrect answer. Further, the step size changed after each "reversal", which occurred when a participant made an incorrect response after a sequence of correct responses (with the distance being smaller). A reversal could also occur when a participant made two consecutive correct responses following an easing of the difficulty due to consecutive incorrect responses (with the distance being wider). The task ended after seven reversals or seventy trials, whichever came first. Performance was calculated as the average of the task difficulty level at all reversals from the second onward. Participants' scores were recorded on a 200-point scale (smaller values indicated more sensitivity to F2 variation).

According to the results of prior studies, the test-retest reliability of the adaptive discrimination task was relatively satisfactory in the context of L1 acquisition (r = .75 in (Raz et al., 1987) and L2 acquisition (r = .70 in Saito, Sun, et al., 2020a). To conduct more reliable and robust evaluations of auditory processing abilities, participants in the current study took the same auditory processing tests for the pre- and post-tests. Then, their scores were averaged at pre- and post-tests, respectively.

L2 Speech Perception Tests

Participants were trained on English [æ] and [ʌ] using speech syllables ([hæ] and [hʌ]) during the phonetic training, and nonverbal sounds (varying along the F2 continuum: 1200-1600 Hz) during the auditory training. The goal of the speech perception test was to examine the generalizability of the training gains to more natural L2 word tokens produced by four different talkers. Using the same format as the phonetic training, the speech perception test comprised a forced-choice identification of minimally-paired words. The same test was used in the pre- and post-tests. The performance of the comparison group (who received comparable phonetic training on the different contrast, English [r] and [l]) was used to index test-retest practice effects, if any.

Stimuli. To avoid participants' excessive focus on the target contrast (English [æ] and [ʌ]), the stimuli comprised 40 target tokens together with 80 comparison tokens. The target tokens included 10 English [æ] and [ʌ] minimal pairs produced by four native speakers of British English. For the Phonetic-Only and Auditory-Phonetic groups (who were exposed to English [æ] and [ʌ] during training), the four speakers represented trained voices (M1 and F1) and untrained voices (M4 and F4). The target tokens were singletons including the target phonological contrast in word-initial position. The distracter tokens comprised 80 minimal pairs that included both difficult (e.g., English [r] and [l]) and easy (e.g., English [ɪ] and [ɛ]) phonological contrasts for Japanese listeners of English (Nishi et al., 2008).

The 10 target words were carefully selected (see **Supporting Information**). The influence of frequency was minimized by taking stimuli from the list of the most frequent 2000-word families (Cobb, 2021). As suggested in Ortega et al. (2019), L2 vowel perception could be influenced by phonetic context. Using the methodological procedure in Ortega et al., we matched the consonants preceding the stimuli in terms of place of articulation (8 for labial, 4 for alveolar,

8 for velar), while equally distributing the following consonants (4 for voiceless stops, 4 for voiced stops, 2 for nasals).

Procedure. All test stimuli (40 target items, 80 distracter items) were presented in randomized order via the Gorilla platform. For each trial, participants listened to an audio stimulus, and selected which word they had heard among two choices (in orthographic form). Participants received instructions from the investigator, completed the task in a quiet room with good access to the internet, and contacted the investigator in the case of any technological difficulties.

Results

Screening Process

Among the 108 participants who initially participated in the current project, 98 were included in the final analyses who met the following inclusion criteria. First, they completed all the testing and training sessions without any major delays (n = 5 excluded). Second, their identification accuracy of the two contrasts (not only English [æ] and [ʌ] but also English [r] and [l]) on the pre- and post-tests was beyond chance level (to ensure that they worked on the task with a clear understanding of the procedure and a good amount of attention; n = 3 excluded). Third, no signs of abnormal participation behaviors were observed (e.g., excessively long session time [> 40 minutes], frequent session interruptions due to technological difficulties; n = 2 excluded). The 10% attrition was comparable to what our research team has typically observed in similar longitudinal studies under lab conditions.

Effects of Training on Auditory Processing

Since participants took the F2 discrimination tests twice (pre and post-tests), their averaged pre- and post-test scores were used as the dependent variable. According to the results of normality tests (Kolmogorov-Smirnov), their pre- and post-test scores significantly differed from a normal distribution, D = .090, .096, p = .048, .027. As in the prior studies (e.g., Kachlicka et al., 2019), they were transformed via the square root function. The resulting scores were indistinguishable from a normal distribution (D = .047, .045, p > .200). Descriptive statistics for these tests are summarized in Table 2 and visually plotted in Figure 4.

To investigate the presence of pre-existing differences before the training, participants' transformed F2 scores were submitted to a one-way ANOVA with Group (Auditory-Only, Phonetic-Only, Auditory-Phonetic, Control) as one between-subjects factor. Effect size (partial eta-squared) was calculated and interpreted using Cohen's (1988) benchmarks: η^2 = .01 for small, .06 for medium, and .14 for large. The analysis did not find any significant main effect of Group, $F(3, 94) = 1.136$, p = .339, η^2 = .003, suggesting that the participants' F2 sensitivity was comparable across the four group conditions.

To examine the extent to which participants' auditory processing changed over time, the pre- and post-test scores were submitted to a two-way ANOVA with one between-subjects factor (Group: Auditory-Only, Phonetic-Only, Auditory-Phonetic, Control), and one within-subjects factor (Time: pre-, post-tests).

The analysis of variance confirmed that there were significant main effects for Time, $F(1, 94) = 4.577$, p = .035, η^2 = .046, and significant interaction effects of Group and Time, $F(3, 94) = 4.955$, p = .003, η^2 = .137. Yet, main effects of Group did not reach statistical significance, Group, $F(3, 94) = 1.701$, p = .172, η^2 = .051. To further examine the Group × Time interaction effects, multiple comparison analyses were performed. The results showed that participants in the Auditory-Only and Auditory-Phonetic groups significantly reduced their F2 discrimination thresholds (i.e., achieving more precise auditory processing scores) over time with medium

effects, $F(1, 94) = 7.056, 6.422$, $p = .009, .013$, $\eta^2 = .070, .064$. Neither the Phonetic-Only nor the Control groups changed their auditory processing abilities, $F(1, 94) = 0.172, 2.167$, $p = .679, .109$, $\eta^2 = .002, .027$.

Table 2

Descriptive Statistics of Auditory Processing (F2 Discrimination) Scores at Pre- and Post-Tests

	Pre-Test				Post-Test			
	M	*SD*	95% CI		*M*	*SD*	95% CI	
			Lower	*Upper*			*Lower*	*Upper*
Auditory-Only	7.999	1.458	7.352	8.645	6.956	1.596	6.248	7.664
Phonetic-Only	7.015	1.506	6.346	7.682	6.852	1.410	6.226	7.477
Auditory-Phonetic	7.902	1.705	7.125	8.678	6.883	1.963	5.989	7.777
Control	7.650	2.529	6.753	8.547	8.255	2.169	7.486	9.024

Note. Smaller auditory scores indicate more precise acuity towards F2 variation (1200-1600 Hz)

Effects of Training on L2 Speech Proficiency

To examine the extent to which the different types of training facilitated participants' L2 speech proficiency, binomial logistic mixed effects models were performed. Participants' accuracy score for each item was coded 0 (for incorrect response) and 1 (for correct response) and used as dependent variables. Fixed effects included Group (Auditory-Only, Phonetic-Only, Auditory-Phonetic, Control), Time (pre, post), Talker (trained, untrained), and Auditory Processing (F2 discrimination scores). Random effects comprised participant ID (1-98) and stimulus ID (1-40). Due to the small sample size, random intercepts but not slopes were included. All results (Models 1-3) are summarized in Table 3 and visually plotted in Figure 4.

In Model 0, we first checked whether the four different groups differed prior to training. Using participants' pre-test logit scores (0 for incorrect; 1 for correct) as dependent variables, the model did not find significant main effects of Group (β = .065, z = 0.871, p = .383). The null results suggest that the four groups' L2 speech perception proficiency (English [r] and [l] accuracy) was comparable at the beginning of the project.

In Model 1, we then examined the extent to which the groups differentially improved their L2 speech abilities over time. The model included participants' pre and post-test logit scores as dependent variables relative to two predictor variables (Group, Time). Statistically significant effects were found for Time (β = .0554, z = 4.123, p = < .001), and Time × Group (β = .145, z = 3.174, p = .001). Using participants' averaged vowel perception scores (%), post-hoc multiple comparison analyses showed that a significant, small-to-medium improvement was observed over time for the Auditory-Only (6.3% gain; 65.6 → 71.8%; F(1, 94) = 7.856, p = .006, η^2 = .077), Phonetic-Only (5.5% gain; 68.9 → 74.3%; F(1, 94) = 10.315, p = .002, η^2 = .099), and Auditory-Phonetic groups (4.0% gain; 71.3 → 75.4%; F(1, 94) = 4.129, p = .045, η^2 = .042). However, no significant improvement was found for the Control group (66.6 → 64.9%; F(1, 94) = 1.100, p = .297, η^2 = .012). This showed that the gains among the experimental groups (Auditory-Only, Phonetic-Only, and Auditory-Phonetic) were not due to test-retest effects in the current study.

In Model 2, we examined the extent to which the Group × Time interaction effects could be mediated by the talker conditions. This tested the question of whether the gains of the Phonetic-Only and Auditory-Phonetic groups were affected by their familiarity with the talkers (i.e., trained vs. untrained). To this end, main effects of Talker and interaction effects of Group, Time, and Talker were added to the original model. Interestingly, while the Time × Group interaction effects remained significant (β = .186, z = 3.257, p = .001), the three-way interaction terms did not reach statistical significance (β = .027, z = 1.190, p = .233). The results suggest that training was facilitative of the development of speech perception regardless of talker conditions.

In Model 3, we examined how participants with diverse biographical profiles differentially benefited from the training. The analyses focused only on the three experimental groups (Auditory-Only, Phonetic-Only, and Auditory-Phonetic). Three different mediating factors tapped into participants' individual differences in (a) pre-existing auditory processing ability (i.e., F2 discrimination scores at pre-tests [0-200 points]), (b) immersion experience (whether participants had any study- and living-abroad experience [1 for yes, 0 for no]), and (c) in-session gains (the extent to which participants demonstrated improvement over the course of the training).

The last predictor (in-session gains) was coded 1 for greater gains and 0 for lesser gains as follows. As reported below, participants' performance at every session (Days 1-6) was

recorded for auditory training (0-100 points) and phonetic training (0-100%). Based on the median values for their gain scores (Day 6 minus Day 1 scores), the participants in the Auditory-Only and Phonetic-Only groups were divided into two subgroups, respectively (greater vs. lesser gains). As for those in the Auditory-Phonetic group, greater vs. lesser gains were calculated in accordance with the median values of averaged gain scores in auditory and phonetic training. The data were coded categorically rather than continuously because the raw improvement scores from the different training programs were not directly comparable. Although they were recorded on a 100-point scale, they were based on different units and concepts (enhancing auditory sensitivity vs. phonetic identification accuracy).

The model found significant main effects of Time (β = .364, z = 2.048, p = .040) but not interaction effects of Time × Group (β = .090, z = 0.554, p = .579). When it comes to the participants' individual difference variables (auditory processing, immersion experience, in-session gains), none of the main and interaction effects reached statistical significance ($p > .05$). The results suggest that all the participants in the experimental groups equally improved their English [æ] and [ʌ] accuracy regardless of their group conditions (Auditory-Only, Phonetic-Only, and Auditory-Phonetic) and their different learner profiles (pre-existing aptitude, immersion experience, and training performance).

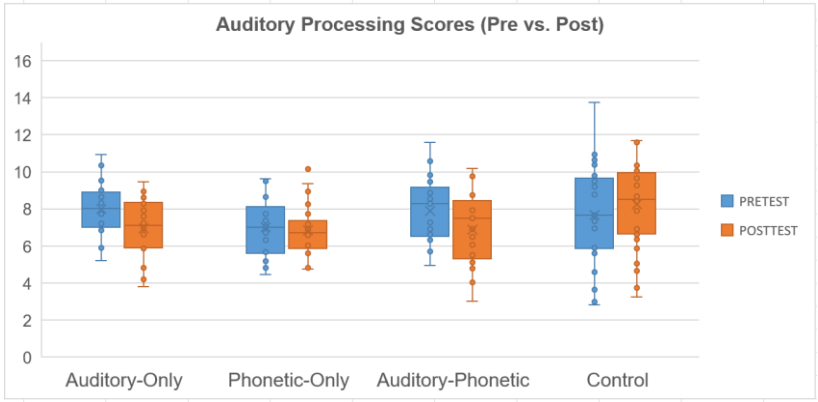
Table 3

Summary of Mixed Effects Modeling Analyses of Group Gains Relative to Talker and Auditory Processing Conditions

	Fixed effects	*β*	*SE*	*z*	*p*	Random effects	Variance	*SD*	R^2
Model 0	Intercepts	.809	.218	3.700	< .001	Participants	.108	.329	.128
	Group	.065	.075	0.871	.383	Stimulus	.857	.926	
Model 1	Intercepts	.391	.272	1.433	.151	Participants	.153	.391	.234
	Group	.130	.079	1.643	.100	Stimulus	.828	.910	
	Time	.554	.134	4.123	< .001*				
	Time × Group	.145	.045	3.174	.001*				
Model 2	Intercepts	.591	.535	1.104	.269	Participants	.153	.391	.234
	Group	.130	.079	1.641	.100	Stimulus	.828	.910	
	Time	.555	.134	4.122	< .001*				
	Talker	.134	.307	0.438	.661				
	Group × Time	.186	.057	3.257	.001*				
	Group × Time × Talker	.027	.023	1.190	.233				
Model 3	Intercepts	1.149	.622	1.847	.064	Participants	.141	.376	.253
	Group	.113	.145	0.783	.433	Stimulus	.923	.961	
	Time	.364	.178	2.048	.040*				
	Auditory processing	-.001	.001	-0.294	.768				
	Immersion experience	.225	.209	1.075	.282				
	In-session gains	.200	.224	0.890	.373				
	Group × Time	.090	.164	0.554	.579				
	Group × Time × Auditory processing	-.001	.001	-0.899	.368				
	Group × Time × Immersion experience	.074	.055	1.332	.183				
	Group × Time × In-session gains	.011	.060	0.191	.848				

Note. * for $p < .05$

4A: Auditory Processing Abilities



4B: L2 Speech Perception Abilities

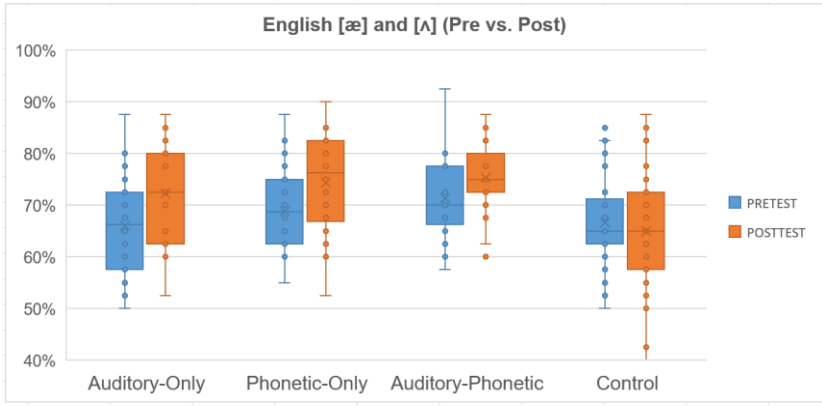


Figure 4

Group Gains Between Pre- and Post-Test Sessions. 4A represents the effects of auditory and phonetic training on F2 discrimination scores (transformed scores; lower values indicate more precise acuity towards F2 variation). 4B represents the effects of auditory and phonetic training on English [æ] and [ʌ] identification accuracy (%).

Distracter Stimuli

Besides the 40 target stimuli (English [æ] and [ʌ] minimal pairs), the L2 speech tests included the 80 distracter stimuli (English [r] and [l] minimal pairs). They were included to measure test-retest effects among the experimental groups who did not engage in any intensive exposure to English [r] and [l] between pre- and post-tests (Audio-Only, Phonetic-Only, Auditory-Phonetic). Using participants' logit scores for English [r] and [l] as dependent variables as per Group and Time as predictors, another set of binomial logistic mixed-effects analysis did not find any significant factors, including Group (β = .016, *SE* = .042, *z* = .377, *p* = .708), Time (β = .005, *SE* = .015, *z* = .375, *p* = .709), or Group and Time (β = .013, *SE* = .022, *z* = .602, *p* = .549). The null results suggest that no learning took place on the participants' English [r] and [l] proficiency when participants took the same speech perception tests twice and received the phonetic and auditory training related to English [æ] and [ʌ].

In-Session Performance

The impact of training on auditory processing and L2 speech perception performance was also assessed for each individual session for each treatment condition. Descriptive statistics are visually summarized in Figure 5. Given that participants completed the training session using their own computers, the precision of the data (especially regarding reaction time) could have been affected by a range of factors (e.g., internet speed). Therefore, the results presented here should be interpreted as suggestive patterns. Here, the analyses of participants' in-session performance were restricted to within-group comparisons (Days 1-6). However, we did not conduct any analyses for the between-group comparisons (Auditory-Only, Phonetic-Only, and Auditory-Phonetic). For each session, not only were the assessment criteria incompatible in auditory training (0-100 points for perceptual sensitivity) vs. phonetic training (0-100% for vowel identification), but the amount of time for each training was also essentially different (30 minutes for Auditory-Only and Phonetic-Only; 15 minutes for Auditory-Phonetic).

- **Auditory-Only (3 Hours of Auditory Training)**: Participants' F2 discrimination accuracy was recorded on a 100-point scale for each session (Days 3-8). Larger values indicate a more precise ability to perceive the target acoustic parameter (F2 = 1200-1600 Hz). These scores were submitted to a one-way repeated-measure ANOVA to examine the extent to which F2 sensitivity improved over the course of six 30-minute sessions. The analysis yielded significant main effects of Time, $F(1, 21) = 5.166$, $p = .034$, $\eta^2 = .197$. According to multiple comparison analyses (alpha set to .008 for five comparisons via Bonferroni correction), participants significantly enhanced their F2 discrimination abilities with large effects after the first 30-minute session, i.e., between Days 1 and 2 ($M = 61.4 \rightarrow 69.5$ out of 100), $F(1, 21) = 16.862$, $p < .001$, $\eta^2 = .455$. Yet, their performance was comparable for the rest of the project: Days 2 vs. 3 ($M = 69.5 \rightarrow 67.8$; $F = .555$, $p = .465$, $\eta^2 = .026$), Day 3 vs. 4 ($M = 67.8 \rightarrow 69.7$; $F = .777$, $p = .465$, $\eta^2 = .036$), Days 4 vs. 5 ($M = 69.7 \rightarrow 68.0$; $F = .561$, $p = .462$, $\eta^2 = .026$), and Days 5 vs. 6 ($M = 68.0 \rightarrow 68.9$; $F = .124$, $p = .728$, $\eta^2 = .006$).
- **Phonetic-Only (3 Hours of Phonetic Training)**: Participants' averaged English [æ] and [ʌ] accuracy scores were recorded for each session (0-100%; Days 3-8). Larger values indicate more targetlike L2 speech proficiency. A one-way repeated-measures ANOVA was performed, yielding a significant main effect for Time with large effects, $F(1, 21) = 49.122$, $p < .004$, $\eta^2 = .701$. The results of multiple comparison analyses were performed to take a closer look at how participants enhanced their vowel identification accuracy

over the course of the six 30-minute sessions. Significant improvement was only found between Days 1 and 2 (M = 76.7% → 81.1%; F = 9.809, p = .001, η^2 = .371) and between Days 3 and 4 (M = 83.3% → 87.0%; F = 19.763, p = < .001, η^2 = .485). Performance appeared to be unchanged in the other sessions: Days 2 and 3 (M = 81.1% → 83.3%; F = 4.161, p = .054, η^2 = .165), Days 4 and 5 (M = 87.0% → 87.3%; F = .123, p = .730, η^2 = .006); and Days 5 and 6 (M = 87.3% → 87.8%; F = .541, p = .509, η^2 = .021) (alpha set to .008, Bonferroni-corrected).

- **Auditory-Phonetic (1.5 Hours of Auditory Training + 1.5 hours of Auditory Training)**: Unlike the Audio-Only and Phonetic-Only groups, participants in the Audio-Phonetic group spent half of the time on auditory training (1.5 hours) and the other half on phonetic training (1.5 hours) over the course of the six days. Regarding training effects, a one-way repeated-measures ANOVA found significant effects for Time, F(1, 20) = 12.258, p = .002, η^2 = .380. According to the results of multiple comparisons (alpha set to .008, Bonferroni-corrected), their Day 1 performance (M = 52.7 out 100) was significantly different from Day 2 (M = 64.1; F = 11.798, p = .003, η^2 = .371), Day 3 (M = 65.4; F = 14.614, p = .001, η^2 = .422), Day 5 (M = 66.4; F = 13.069, p = .002, η^2 = .395), and Day 6 (M = 67.6; F = 13.069, p = .002, η^2 = .395). However, no significant differences were found for the rest of the contrasts (i.e., p > .008 for Days 2 vs. 3 vs. 4 vs. 5 vs. 6). In terms of L2 speech proficiency, a one-way repeated-measures ANOVA demonstrated significant effects of Time, F(1, 20) = 15.043, p < .001, η^2 = .429. Follow-up analyses further revealed that significant improvements were observed between Days 1 and 2 (M = 72.8% → 76.7%; F = 9.360, p = .006, η^2 = .319), Days 2 and 4 (M = 76.7% → 80.5%; F = 13.273, p = .002, η^2 = .399), Days 3 and 6 (M = 79.3% → 83.3%; F = 11.566, p = .003, η^2 = .366).

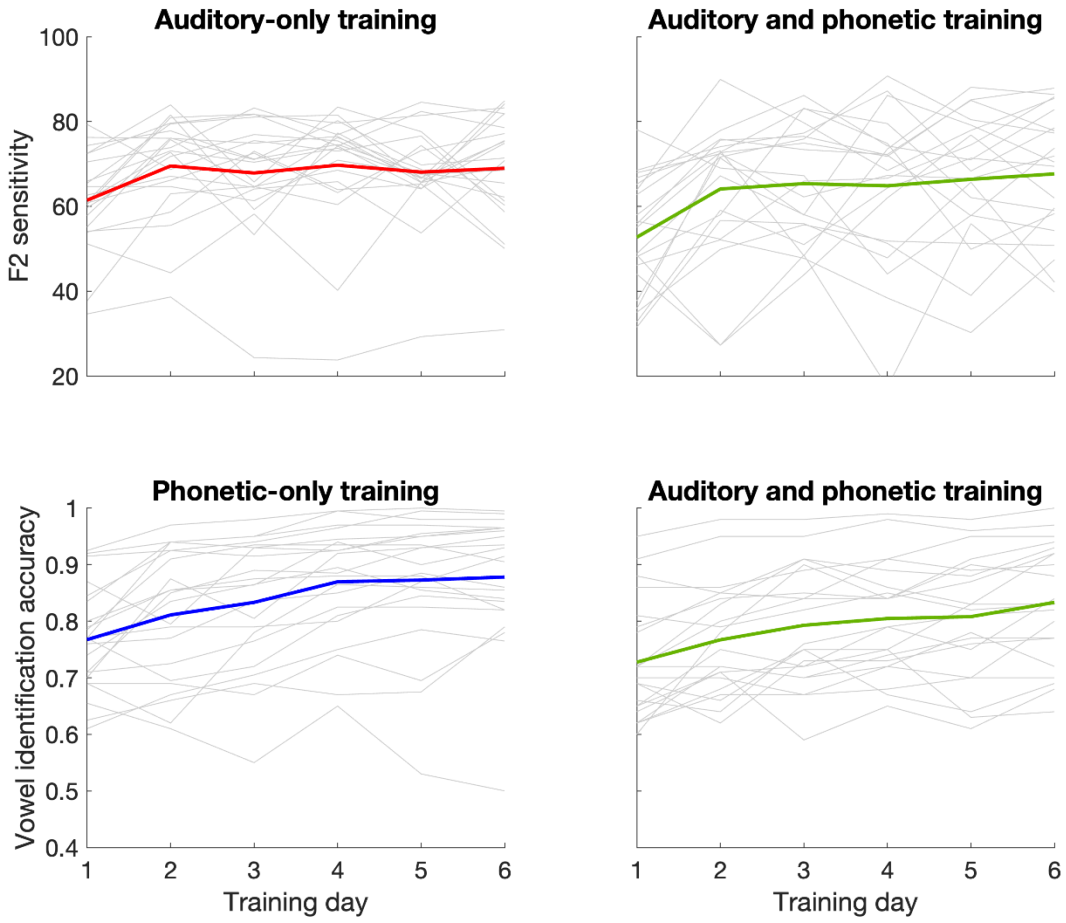


Figure 5
Within-Session Performance of Auditory Processing and English [æ] and [ʌ] Among Three Experimental Groups: Auditory Training Only (3 Hours), Phonetic Training Only (3 Hours), and Auditory & Phonetic Training (1.5 + 1.5 Hours)

Discussion

To disentangle the complex associations between auditory processing and post-pubertal L2 speech learning, the current investigation set out to examine the extent to which provision of auditory versus phonetic training could facilitate the development of auditory and phonetic abilities. Unlike prior studies in hearing research concerning L1 behaviors on a *global* level (e.g., McArthur et al., 2008 for the effects of frequency discrimination on L1 reading difficulty), the current study took a first step towards examining the relative effectiveness of auditory and phonetic training for post-pubertal L2 speech learning. As such, we investigated whether training discrimination of auditory dimensions (Japanese speakers' sensitivity to 1200-1600 Hz) can trigger the development of specific L2 speech skills (the identification of English [æ] and [ʌ]).

Overall, all training methods (Auditory-Only, Phonetic-Only, and Auditory-Phonetic) significantly enhanced different areas of participants' auditory and speech abilities with medium effects. Provision of auditory training helped improve both auditory sensitivity (F2 discrimination of 1200-1600 Hz) and L2 speech proficiency (the identification of English [æ] and [ʌ]) whether it was combined with phonetic training or not. In contrast, the gains of phonetic training (Phonetic-Only) were limited to speech perception (English [æ] and [ʌ]).

Both Phonetic-Only and Auditory-Phonetic engaged in the high variability phonetic training. Throughout the training, the participants were presented with written words, pictures and sounds while intensively exposed to English [æ] and [ʌ] exemplars produced by multiple talkers. As such, they were guided to attend to both the phonetic and semantic information of the stimuli. The amount of the gains in the current study, i.e., 5.5% for Phonetic-Only and 4.0% for Auditory-Phonetic, aligned with the meta-analysis of previous training studies (Sakai & Moorman, 2018). Different from Phonetic-Only and Auditory-Phonetic, the Auditory-Only group engaged in auditory training wherein they were not exposed to the English [æ] and [ʌ] exemplars at all. The training materials comprised only non-verbal sounds without any opportunities to process input for meaning. According to the results, however, the Audio-Only group resulted in similar gains (i.e., 6.3%), and their L2 phonetic learning patterns were comparable to the other groups who enjoyed the benefits of phonetic training, Phonetic-Only and Auditory-Phonetic (5.5%, 4.0%).

These asymmetric findings (i.e., transfer effects were found for auditory training, but not for phonetic training) could be considered as empirical support for the directional nature of the perception-acquisition link (enhanced auditory precision → L2 speech development). That is, one's sensitivity to certain key, domain-general acoustic cues (F2 = 1200-1600Hz) promotes speech learning on a domain-specific level (English [æ] vs. [ʌ]). Considered alongside the existing cross-sectional and longitudinal evidence in L1 and L2 acquisition (e.g., Kachlicka et al., 2019), the results support the theoretical views (a) that language learning draws upon auditory processing throughout the lifespan (Goswami, 2015); and (b) that the systems used for L1 acquisition remain intact and germane to L2 acquisition (Flege & Bohn, 2021).

From a methodological perspective, the current study was designed to clarify how specific dimensions of auditory and speech abilities were linked at a fine-grained level (F2 variation vs. English [æ] and [ʌ] perception). Our unique approach and findings shed some light on why the results of L1 hearing research on the generalizability of auditory training gains to language development have remained mixed (e.g., Schäffler et al., 2004 vs. Strehlow et al., 2006 for significant vs. non-significant effects of pitch and intensity discrimination training). These previous studies have focused on the *global* relationships between auditory processing and

overall language skills (e.g., word recognition, pseudoword repetition, vocabulary size, and spelling accuracy).

Here, it is important to remember that whereas the acquisitional value of auditory processing for phonetic and phonological learning is straightforward (the context of the current study), the strength of the link may be weak when it comes to their associations with higher-order linguistic abilities (reading, listening, writing, and speaking). In the context of L2 acquisition, research has shown that the attainment of global linguistic skills is tied to a range of other sublinguistic skills (e.g., De Jong et al., 2015 for the role of vocabulary, grammar, and pronunciation for global L2 speaking proficiency) and cognitive abilities (for the roles of working memory and attention control in L2 listening proficiency, see Vafaee & Suzuki, 2020). Thus, it is important to acknowledge the current study as initial evidence for a potential causal link between auditory processing and phonological competence. Future studies could further pursue to what degree auditory training can enhance various areas of linguistic competence (phonology *and* lexicogrammar), and then explore whether they lead to global language gains in the long run.

Another intriguing point of discussion concerns the lack of transfer effects when it comes to phonetic training and auditory processing abilities. The findings discussed here align with previous studies showing that mere exposure to L2 phonemes does not necessarily facilitate the development of adult L2 learners' new acoustic representations (Ingvalson et al., 2011 for the nil effects of similar training methods on Japanese speakers' F3 sensitivity which is a main acoustic correlate of English [r] and [l] contrast).

One explanation could be that those who engaged in phonetic training may have learned how to identify the target contrast (English [æ] vs. [ʌ]) using cues other than F2 variation. For example, English [æ] and [ʌ] can also be distinguished based on the duration cue (Umeda, 1975). Japanese listeners tend to prioritize such durational differences because a similar acoustic contrast (i.e., short-long vowel distinction) is present in their L1 (Strange et al., 2011). Further, it has been shown that inexperienced L2 listeners overly rely on the temporal (rather than spectral) aspects of new vocalic contrasts as an interlanguage strategy (Flege et al., 1997). To obtain more advanced L2 speech proficiency, it is important for L2 learners to encode both the spectral *and* temporal properties of new sounds in a well-balanced fashion (see Flege et al., 1997 for the use of spectral vs. temporal cues among L2 listeners with less than 1 year of immersion vs. 7 years).

To help L2 learners attain nativelike perceptual strategies (spectral encoding for English [æ] vs. [ʌ]), intensive exposure to phonetically rich, multi-talker input alone (e.g., high variability phonetic training) may not be sufficient to promote the detection, learning, and automatization of optimal L2 cue weighting strategies because speech contrasts have redundant information (e.g., duration and F2 for English [æ] vs. [ʌ]), and/or because those with relatively low auditory precision may have difficulty processing such complex acoustic signals (Perrachione et al., 2011). Therefore, we call for future research to examine how the provision of phonetic training can help post-pubertal L2 learners to selectively focus on target relevant acoustic parameters in the context of speech sounds despite irrelevant variation in other parameters and explore whether such training can enhance both auditory processing and phonetic abilities.

Future Directions

With an eye towards future replication and extension studies, we address a range of conceptual and methodological issues that that scholars should further elaborate, expand, and refine. First, although the current study demonstrates the generalizability of the auditory training

to L2 speech learning, the mechanisms underlying such results need to be further examined with much caution and more methodological rigor. Although the tasks used in this study are assumed to tap into participants' perceptual acuity to a particular acoustic parameter (A × B discrimination for F2 sensitivity), it remains subject to further investigation whether and to what degree such behavioral tasks can assess participants' auditory perception skills without involving and confounding with other cognitive abilities (e.g., attentional control; Snowling et al., 2018). To this end, a range of neural measures of sound processing have been suggested to tap into pre-attentive perceptual acuity (e.g., the frequency following response; White-Schwoch et al., 2017).

Second, the link between auditory training and phonological competence needs to be re-examined in various areas of L2 speech acquisition with different levels of difficulty. As stated in McAllister et al.'s (2002) Feature Hypothesis, the focus of the current study (Japanese speakers' English [æ] and [ʌ] acquisition) could be considered as a relatively easy instance of L2 speech acquisition. Although Japanese speakers initially assimilate the new phones (English [æ] vs. [ʌ]) into one L1 counterpart (Japanese [a]), they can quickly develop two new separate phonetic categories after a few hours of training (e.g., Lambacher et al., 2005). This is because they can perceive the primary acoustic correlates of the contrast (i.e., F2 = 1200 Hz vs. 1600 Hz) by re-finetuning their already-existing F2 representation which they regularly use for the perception of L1 vowels (e.g., Japanese [i] and [u]). It would be intriguing to examine the generalizability of the auditory training effects in a relatively difficult instance of L2 speech acquisition. One such example is Japanese speakers' English [r] and [l] acquisition. Japanese speakers likely show tremendous difficulty learning English [r] and [l] even after years of immersion experience arguably because the primary acoustic parameter for English [r] and [l] (i.e., F3 = 1800 Hz vs. 2500 Hz) is not actively used in the L1 phonetic system (Ingvalson, et al., 2011). Future studies should examine the extent to which provision of auditory training can facilitate the development of the new acoustic representations (e.g., F3 for Japanese speakers) and then promote L2 speech learning (e.g., Japanese speakers' English [r] and [l] acquisition).

Finally, in the current study, L2 speech proficiency was measured via highly analytic, language-focused tasks (i.e., forced-choice identification of English [æ] and [ʌ]). Attaining high-level performance in such behavioral tasks inevitably draws on participants' L2 perception skills as well as a range of other executive functions. Similarly, some scholars have argued that L2 speech acquisition needs to be examined via not only controlled identification tasks but also spontaneous production tasks so as to mirror how L2 learners actually and spontaneously access the target language in real-life settings (Saito & Plonsky, 2019). To further scrutinize the multifaceted nature of the relationship between auditory processing and L2 speech learning, future studies are called for which adopt a range of outcome measures to cover the perceptual, cognitive, phonetic, and linguistic profiles of adult L2 learners from various angles. With such a design, scholars would be able to comprehensively examine how auditory training can alter the route and outcomes of L2 speech learning while statistically controlling for individual differences in cognitive abilities (e.g., Snowling et al., 2018), tracking precognitive neural encoding of sounds within both training and testing sessions (e.g., White-Schwoch et al., 2017), and/or eliciting L2 speech via controlled and communicatively-oriented speech tasks (e.g., Saito & Plonsky, 2019).

References

Abrahamsson, N., & Hyltenstam, K. (2009). Age of Onset and Nativelikeness in a Second Language: Listener Perception Versus Linguistic Scrutiny. *Language Learning*, *59*(2), 249–306. https://doi.org/10.1111/j.1467-9922.2009.00507.x

Anvari, S. H., Trainor, L. J., Woodside, J., & Levy, B. A. (2002). Relations among musical skills, phonological processing, and early reading ability in preschool children. *Journal of Experimental Child Psychology*, *83*(2), 111–130. https://doi.org/10.1016/S0022-0965(02)00124-8

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

Barriuso, T. A., & Hayes-Harb, R. (2018). High Variability Phonetic Training as a Bridge from Research to Practice. *CATESOL Journal*, *30*(1), 177–194.

Bavin, E. L., Grayden, D. B., Scott, K., & Stefanakis, T. (2010). Testing Auditory Processing Skills and their Associations with Language in 4—5-year-olds. *Language and Speech*, *53*(1), 31–47. https://doi.org/10.1177%2F0023830909349151

Boets, B., Wouters, J., van Wieringen, A., De Smedt, B., & Ghesquière, P. (2008). Modelling relations between sensory processing, speech perception, orthographic and phonological ability, and literacy achievement. *Brain and Language*, *106*(1), 29–40. https://psycnet.apa.org/doi/10.1016/j.bandl.2007.12.004

Cobb, T. (2021). *Compleat Lexical Tutor*. https://www.lextutor.ca/

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

Cutler, A., & Butterfield, S. (1992). Rhythmic cues to speech segmentation: Evidence from juncture misperception. *Journal of Memory and Language*, *31*(2), 218–236. https://doi.org/10.1016/0749-596X(92)90012-M

Darcy, I., Park, H., & Yang, C.-L. (2015). Individual differences in L2 acquisition of English phonology: The relation between cognitive abilities and phonological processing. *Learning and Individual Differences*, *40*, 63–72. https://doi.org/10.1016/j.lindif.2015.04.005

De Jong, N. H., Groenhout, R., Schoonen, R., & Hulstijn, J. H. (2015). Second language fluency: Speaking style or proficiency? Correcting measures of second language fluency for first language behavior. *Applied Psycholinguistics*, *36*(2), 223–243. https://doi.org/10.1017/S0142716413000210

Derwing, T. M., & Munro, M. J. (2013). The Development of L2 Oral Language Skills in Two L1 Groups: A 7-Year Study: Development of L2 Oral Skills. *Language Learning*, *63*(2), 163–185. https://doi.org/10.1111/lang.12000

Douglas, S., & Willatts, P. (1994). The relationship between musical ability and literacy skills. *Journal of Research in Reading*, *17*(2), 99–107. https://doi.org/10.1111/j.1467-9817.1994.tb00057.x

Dubinsky, E., Wood, E. A., Nespoli, G., & Russo, F. A. (2019). Short-Term Choir Singing Supports Speech-in-Noise Perception and Neural Pitch Strength in Older Adults With Age-Related Hearing Loss. *Frontiers in Neuroscience*, *13*, 1153. https://doi.org/10.3389/fnins.2019.01153

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175–191. https://doi.org/10.3758/BF03193146

Flege, J. E., & Bohn, O.-S. (2021). The Revised Speech Learning Model (SLM-r). In R. Wayland (Ed.), *Second Language Speech Learning* (1st ed., pp. 3–83). Cambridge University Press. https://doi.org/10.1017/9781108886901.002

Flege, J. E., Bohn, O.-S., & Jang, S. (1997). Effects of experience on non-native speakers' production and perception of English vowels. *Journal of Phonetics*, *25*(4), 437–470. https://doi.org/10.1006/jpho.1997.0052

Flege, J. E., & Liu, S. (2001). The effects of experience on adults' acquisition of a second language. *Studies in Second Language Acquisition*, *23*(4), 527–552. https://doi.org/10.1017/S0272263101004041

Flege, J. E., Takagi, N., & Mann, V. (1995). Japanese Adults can Learn to Produce English /I/ and /l/ Accurately. *Language and Speech*, *38*(1), 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*(2), 1161–1173. https://doi.org/10.1121/1.414884

Fostick, L., Taitelbaum-Swead, R., Kreitler, S., Zokraut, S., & Billig, M. (2020). Auditory Training to Improve Speech Perception and Self-Efficacy in Aging Adults. *Journal of Speech, Language, and Hearing Research*, *63*(4), 1270–1282. https://doi.org/10.1044/2019_JSLHR-19-00355

Ghaffarvand Mokari, P., & Werner, S. (2019). On the Role of Cognitive Abilities in Second Language Vowel Learning. *Language and Speech*, *62*(2), 260–280. https://doi.org/10.1177/0023830918764517

Goswami, U. (2015). Sensory theories of developmental dyslexia: Three challenges for research. *Nature Reviews. Neuroscience*, *16*(1), 43–54.

Hämäläinen, J. A., Salminen, H. K., & Leppänen, P. H. T. (2013). Basic Auditory Processing Deficits in Dyslexia: Systematic Review of the Behavioral and Event-Related Potential/ Field Evidence. *Journal of Learning Disabilities*, *46*(5), 413–427. https://doi.org/10.1177%2F0022219411436213

Hawkins, S., & Midgley, J. (2005). Formant frequencies of RP monophthongs in four age groups of speakers. *Journal of the International Phonetic Association*, *35*(2), 183–199. https://doi.org/10.1017/S0025100305002124

Hayes, E. A., Warrier, C. M., Nicol, T. G., Zecker, S. G., & Kraus, N. (2003). Neural plasticity following auditory training in children with learning problems. *Clinical Neurophysiology*, *114*(4), 673–684. https://doi.org/10.1016/S1388-2457(02)00414-5

Henshaw, H., & Ferguson, M. A. (2013). Efficacy of Individual Computer-Based Auditory Training for People with Hearing Loss: A Systematic Review of the Evidence. *PLOS ONE*, *8*(5), e62836. https://doi.org/10.1371/journal.pone.0062836

Hornickel, J., & Kraus, N. (2013). Unstable Representation of Sound: A Biological Marker of Dyslexia. *Journal of Neuroscience*, *33*(8), 3500–3504. https://doi.org/10.1523/JNEUROSCI.4205-12.2013

Hu, X., Ackermann, H., Martin, J. A., Erb, M., Winkler, S., & Reiterer, S. M. (2013). Language aptitude for pronunciation in advanced second language (L2) Learners: Behavioural

predictors and neural substrates. *Brain and Language*, *127*(3), 366–376. https://doi.org/10.1016/j.bandl.2012.11.006

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*(2), 255–274. https://doi.org/10.1017/S1366728911000447

Iverson, P., Hazan, V., & Bannister, K. (2005). Phonetic training with acoustic cue manipulations: A comparison of methods for teaching English /r/-/l/ to Japanese adults. *The Journal of the Acoustical Society of America*, *118*(5), 3267–3278. https://doi.org/10.1121/1.2062307

Joanisse, M. F., & Seidenberg, M. S. (1998). Specific language impairment: A deficit in grammar or processing? *Trends in Cognitive Sciences*, *2*(7), 240–247. https://doi.org/10.1016/S1364-6613(98)01186-3

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

Kalashnikova, M., Goswami, U., & Burnham, D. (2019). Sensitivity to amplitude envelope rise time in infancy and vocabulary development at 3 years: A significant relationship. *Developmental Science*, *22*(6), e12836-n/a. https://doi.org/10.1111/desc.12836

Kempe, V., Bublitz, D., & Brooks, P. J. (2015). Musical ability and non-native speech-sound processing are linked through sensitivity to pitch and spectral information. *British Journal of Psychology*, *106*(2), 349–366. https://doi.org/10.1111/bjop.12092

Kidd, G. R., Watson, C. S., & Gygi, B. (2007). Individual differences in auditory abilities. *The Journal of the Acoustical Society of America*, *122*(1), 418–435. https://doi.org/10.1121/1.2743154

Lamb, S. J., & Gregory, A. H. (1993). The Relationship between Music and Reading in Beginning Readers. *Educational Psychology (Dorchester-on-Thames)*, *13*(1), 19–27. http://dx.doi.org/10.1080/0144341930130103

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*(2), 227–247. https://doi.org/10.1017/S0142716405050150

Larson-Hall, J. (2015). *A Guide to Doing Statistics in Second Language Research Using SPSS and R* (2nd ed.). Routledge. https://doi.org/10.4324/9781315775661

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*(6), 3757–3768. https://doi.org/10.1121/1.3506351

Leong, C. X. R., Price, J. M., Pitchford, N. J., & Heuven, W. J. B. van. (2018). High variability phonetic training in adaptive adverse conditions is rapid, effective, and sustained. *PLOS ONE*, *13*(10), e0204888. https://doi.org/10.1371/journal.pone.0204888

Levitt, 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

Linck, J. A., Hughes, M. M., Campbell, S. G., Silbert, N. H., Tare, M., Jackson, S. R., Smith, B. K., Bunting, M. F., & Doughty, C. J. (2013). Hi-LAB: A New Measure of Aptitude for High-Level Language Proficiency. *Language Learning*, *63*(3), 530–566. https://doi.org/10.1111/lang.12011

Logan, J. S., Lively, S. E., & Pisoni, D. B. (1991). Training Japanese listeners to identify English /r/ and /l/: A first report. *The Journal of the Acoustical Society of America*, *89*(2), 874–886. https://doi.org/10.1121/1.1894649

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. *Journal of phonetics*, *30*(2), 229-258. https://doi.org/10.1006/jpho.2002.0174

McArthur, G. M., Ellis, D., Atkinson, C. M., & Coltheart, M. (2008). Auditory processing deficits in children with reading and language impairments: Can they (and should they) be treated? *Cognition*, *107*(3), 946–977. https://doi.org/10.1016/j.cognition.2007.12.005

Merzenich, M. M., Jenkins, W. M., Johnston, P., Schreiner, C., Miller, S. L., & Tallal, P. (1996). Temporal Processing Deficits of Language-Learning Impaired Children Ameliorated by Training. *Science*, *271*(5245), 77–81. https://doi.org/10.1126/science.271.5245.77

Micheyl, C., Delhommeau, K., Perrot, X., & Oxenham, A. J. (2006). Influence of musical and psychoacoustical training on pitch discrimination. *Hearing Research*, *219*(1), 36–47. https://doi.org/10.1016/j.heares.2006.05.004

Mora, J. C., & Valls-Ferrer, M. (2012). Oral Fluency, Accuracy, and Complexity in Formal Instruction and Study Abroad Learning Contexts. *TESOL Quarterly*, *46*(4), 610–641. https://doi.org/10.1002/tesq.34

Mora, J. C. & Mora-Plaza, I. (2019) Contributions of cognitive attention control to L2 speech learning. In Nyvad, A. M., Hejná, M., Højen, A., Jespersen, A. B., & Sørensen, M. H. (eds.) A Sound Approach to Language Matters - In Honor of Ocke-Schwen Bohn. Dept. of English, School of Communication & Culture, Aarhus University, Denmark. 477-499.

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

Nishi, K., & Kewley-Port, D. (2008). Nonnative speech perception training using vowel subsets: Effects of vowels in sets and order of training. *Journal of Speech, Language, and Hearing Research*, *51*(6), 1480–1494.

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*(1), 576–588. https://doi.org/10.1121/1.2931949

Omote, A., Jasmin, K., & Tierney, A. (2017). Successful non-native speech perception is linked to frequency following response phase consistency. *Cortex*, *93*, 146–154. https://doi.org/10.1016/j.cortex.2017.05.005

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

Penner, Z., Wymann, K., & Weissenborn, J. (2001). On the prosody/lexicon interface in learning word order. A study of normally developing and language impaired children. In J. Weissenborn & B. Höhle (Eds.), *Approaches to Bootstrapping: Phonological, lexical, syntactic and neurophysiological aspects of early language acquisition* (Vol. 1). John Benjamins Publishing Company.

Perrachione, T. K., Lee, J., Ha, L. Y. Y., & Wong, P. C. M. (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

Pires, M. M., & Schochat, E. (2019). The effectiveness of an auditory temporal training program in children who present voiceless/voiced-based orthographic errors. *PLOS ONE*, *14*(5), e0216782. https://doi.org/10.1371/journal.pone.0216782

Qin, Z., Zhang, C., & Wang, W. S. (2021). The effect of Mandarin listeners' musical and pitch aptitude on perceptual learning of Cantonese level-tones. *The Journal of the Acoustical Society of America*, *149*(1), 435–446. https://doi.org/10.1121/10.0003330

Ramus, F. (2003). Developmental dyslexia: Specific phonological deficit or general sensorimotor dysfunction? *Current Opinion in Neurobiology*, *13*(2), 212–218. https://doi.org/10.1016/S0959-4388(03)00035-7

Raz, N., Willerman, L., & Yama, M. (1987). On sense and senses: Intelligence and auditory information processing. *Personality and Individual Differences*, *8*(2), 201–210. https://doi.org/10.1016/0191-8869(87)90175-9

Russo, N. M., Skoe, E., Trommer, B., Nicol, T., Zecker, S., Bradlow, A., & Kraus, N. (2008). Deficient brainstem encoding of pitch in children with Autism Spectrum Disorders. *Clinical Neurophysiology*, *119*(8), 1720–1731. https://doi.org/10.1016/j.clinph.2008.01.108

Saito, K., Kachlicka, M., Sun, H., & Tierney, A. (2020). Domain-general auditory processing as an anchor of post-pubertal second language pronunciation learning: Behavioural and neurophysiological investigations of perceptual acuity, age, experience, development, and attainment. *Journal of Memory and Language*, *115*, 104168. https://doi.org/10.1016/j.jml.2020.104168

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. (2020a). *Brief Report: Test-Retest Reliability of Explicit Auditory Processing Measures* (p. 2020.06.12.149484). https://doi.org/10.1101/2020.06.12.149484

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

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

Schäffler, T., Sonntag, J., Hartnegg, K., & Fischer, B. (2004). The effect of practice on low-level auditory discrimination, phonological skills, and spelling in dyslexia. *Dyslexia*, *10*(2), 119–130. https://doi.org/10.1002/dys.267

Smith, J. (2007). *Introduction to Digital Filters: With Audio Applications*. W3K Publishing.

Snowling, M. J., Gooch, D., McArthur, G., & Hulme, C. (2018). Language Skills, but Not Frequency Discrimination, Predict Reading Skills in Children at Risk of Dyslexia. *Psychological Science*, *29*(8), 1270–1282. https://doi.org/10.1177/0956797618763090

Strange, W., Hisagi, M., Akahane-Yamada, R., & Kubo, R. (2011). Cross-language perceptual similarity predicts categorial discrimination of American vowels by naïve Japanese listeners. *The Journal of the Acoustical Society of America*, *130*(4), EL226–EL231. https://doi.org/10.1121/1.3630221

Strehlow, U., Haffner, J., Bischof, J., Gratzka, V., Parzer, P., & Resch, F. (2006). Does successful training of temporal processing of sound and phoneme stimuli improve reading and spelling? *European Child & Adolescent Psychiatry*, *15*(1), 19–29. https://doi.org/10.1007/s00787-006-0500-4

Sun, H., Saito, K., & Tierney, A. (2021). A longitudinal investigation of explicit and implicit auditory processing in L2 segmental and suprasegmental acquisition. *Studies in Second Language Acquisition*, *43*(3), 551–573. https://doi.org/10.1017/S0272263120000649

Talcott, J. B., Witton, C., McLean, M. F., Hansen, P. C., Rees, A., Green, G. G. R., & Stein, J. F. (2000). Dynamic Sensory Sensitivity and Children's Word Decoding Skills. *Proceedings of the National Academy of Sciences - PNAS*, *97*(6), 2952–2957.

Thomson, R. I. (2012). Improving L2 Listeners' Perception of English Vowels: A Computer-Mediated Approach: Improving L2 Listeners' English Vowel Perception. *Language Learning*, *62*(4), 1231–1258. https://doi.org/10.1111/j.1467-9922.2012.00724.x

Tierney, A., Gomez, J. C., Fedele, O., & Kirkham, N. Z. (2021). Reading ability in children relates to rhythm perception across modalities. *Journal of Experimental Child Psychology*, *210*, 105196–105196.

Tierney, A. T., Krizman, J., & Kraus, N. (2015). Music training alters the course of adolescent auditory development. *Proceedings of the National Academy of Sciences*, *112*(32), 10062–10067. https://doi.org/10.1073/pnas.1505114112

Umeda, N. (1975). Vowel duration in American English. *The Journal of the Acoustical Society of America*, *58*(2), 434–445. https://doi.org/10.1121/1.380688

Vafaee, P., & Suzuki, Y. (2020). The relative significance of syntactic knowledge and vocabulary knowledge in second language listening ability. *Studies in Second Language Acquisition*, *42*(2), 383–410. https://doi.org/10.1017/S0272263119000676

Werker, J. F. (2018). Perceptual beginnings to language acquisition. *Applied Psycholinguistics*, *39*(4), 703–728. https://doi.org/10.1017/S0142716418000152

Whiteford, K. L., & Oxenham, A. J. (2018). Learning for pitch and melody discrimination in congenital amusia. *Cortex*, *103*, 164–178. https://doi.org/10.1016/j.cortex.2018.03.012

White-Schwoch, T., Nicol, T., Warrier, C. M., Abrams, D. A., & Kraus, N. (2017). Individual Differences in Human Auditory Processing: Insights From Single-Trial Auditory Midbrain Activity in an Animal Model. *Cerebral Cortex*, *27*(11), 5095–5115. https://doi.org/10.1093/cercor/bhw293

Whitton, J. P., Hancock, K. E., Shannon, J. M., & Polley, D. B. (2017). Audiomotor Perceptual Training Enhances Speech Intelligibility in Background Noise. *Current Biology*, *27*(21), 3237-3247.e6. https://doi.org/10.1016/j.cub.2017.09.014

Witton, C., Swoboda, K., Shapiro, L. R., & Talcott, J. B. (2020). Auditory frequency discrimination in developmental dyslexia: A meta-analysis. *Dyslexia*, *26*(1), 36–51. https://doi.org/10.1002/dys.1645