Cognition 229 (2022) 105236

Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/cognit

Original articles

Auditory precision hypothesis-L2: Dimension-specific relationships between auditory processing and second language segmental learning

Kazuya Saito [a,*], Magdalena Kachlicka [a], Yui Suzukida [a], Katya Petrova [a], Bradford J. Lee [b], Adam Tierney [c]

[a] University College London, United Kingdom
[b] Fukui University of Technology, Japan
[c] Birkbeck, University of London, United Kingdom

ARTICLE INFO

Keywords:
Auditory processing
Second language speech
Aptitude
Segmentals
English-as-a-second-language

ABSTRACT

Growing evidence suggests a broad relationship between individual differences in auditory processing ability and the rate and ultimate attainment of language acquisition throughout the lifespan, including post-pubertal second language (L2) speech learning. However, little is known about how the precision of processing of specific auditory dimensions relates to the acquisition of specific L2 segmental contrasts. In the context of 100 late Japanese-English bilinguals with diverse profiles of classroom and immersion experience, the current study set out to investigate the link between the perception of several auditory dimensions (F3 frequency, F2 frequency, and duration) in non-verbal sounds and English [r]-[l] perception and production proficiency. Whereas participants' biographical factors (the presence/absence of immersion) accounted for a large amount of variance in the success of learning this contrast, the outcomes were also tied to their acuity to the most reliable, new auditory cues (F3 variation) and the less reliable but already-familiar cues (F2 variation). This finding suggests that individuals can vary in terms of how they perceive, utilize, and make the most of information conveyed by specific acoustic dimensions. When perceiving more naturalistic spoken input, where speech contrasts can be distinguished via a combination of numerous cues, some can attain a high-level of L2 speech proficiency by using nativelike and/or non-nativelike strategies in a complementary fashion.

Many second language (L2) learners start learning a target language in adulthood. Post-pubertal learners' L2 speech proficiency could be essentially different from monolinguals' performance, arguably because L2 speech learning takes place in the same space where the first language (L1) system has already been established, inevitably resulting in the interaction between L1 and L2 behaviours (Baker, Trofimovich, Flege, Mack, & Halter, 2008). According to the major theoretical account of adult L2 speech acquisition (i.e., the Revised Speech Learning Model [SLM-r]; Flege & Bohn, 2021), however, the capacity to learn new sounds remains active throughout the lifespan, and germane to post-pubertal L2 speech learning. As shown in the existing literature, many adult L2 learners can attain *advanced* L2 speech proficiency as long as they access a target language through a long-term residence in L2 speaking environments (Derwing & Munro, 2013 for global oral proficiency; Saito & Brajot, 2013 for segmetnals; Trofimovich & Baker, 2006 for suprasegmentals; cf. Muñoz, 2014 for classroom L2 learners with vs. without study-abroad experience). The experience effects can be clearly observed especially when L2 learners regularly interact with native and advanced L2 speakers (Flege & Liu, 2001).

Here, advanced L2 speech proficiency is defined as the ability that a highly functional and successful L2 user possesses. In perception, they can hear L2 sounds quickly and accurately under various phonetic, lexical, and speaker conditions (Bradlow, 2008). In production, their pronunciation forms can be highly intelligible, comprehensible, and fluent (Derwing & Munro, 2013 for the relationship between intelligibility, comprehensibility, and nativelikeness in L2 speech). Notably, there still exists much individual variation in terms of how much highly experienced, functional, and advanced L2 learners can ultimately attain after years of immersion experience. It has been shown that some individuals reach *near-nativelike* L2 speech proficiency which is *almost* indistinguishable from monolingual native speakers of the target L2 (Abrahamsson & Hyltenstam, 2008), and that the incidence of near-nativelike L2 performance has been tied not only to experience-related factors (i.e., quantity, quality, and intensity of L2 immersion) but also

* Corresponding authors at: University College London, Institute of Education, 20 Bedford Way, WC1H0AL, United Kingdom.
E-mail address: k.saito@ucl.ac.uk (K. Saito).

https://doi.org/10.1016/j.cognition.2022.105236
Received 21 May 2021; Received in revised form 11 January 2022; Accepted 25 July 2022
Available online 23 August 2022

to a range of learners' perceptual-cognitive abilities (i.e., aptitude). That is, "all else being equal, aptitude differences would predict differences in language learning success" (Doughty, 2019, p. 107).

However, the question remains as to what aptitude factors help determine the degree of speech learning which an individual will demonstrate at every L2 input opportunity. In the current investigation, we hypothesize that one important factor may be L2 learners' capacity to represent, encode, and integrate domain-general, spectral and temporal dimensions of sound—in other words, auditory processing. Under the framework of the auditory-processing-precision theory in L1 acquisition, domain-general auditory processing is viewed as an anchor of language learning as it can be an affecting factor determining the way learners encounter, parse, and process a new language on domain-specific levels (Goswami, 2015; Mueller, Friederici, & Mañnel, 2012; Tallal, 2004; Tierney & Kraus, 2014). More recently, a growing body of evidence has suggested that individual differences in domain-general auditory processing can help predict variability across various dimensions of L2 speech learning (i.e., auditory precision hypothesis - L2; e.g., Kachlicka, Saito, & Tierney, 2019 for speech perception; Saito, Kachlicka, Sun, & Tierney, 2020 for speech production).

The auditory precision hypothesis – L2 is in line with Flege and Bohn's (2021) SLM-r in that both consider the quantity, quality, and intensity of experience as the primary determinant of L2 speech learning: Provided that post-pubertal learners engage in long-term immersion experience, they can become successful, functional L2 users with advanced L2 speech proficiency. Echoing the aptitude account of high-level L2 acquisition (Abrahamsson & Hyltenstam, 2008; Doughty, 2019), the auditory precision hypothesis – L2 further links the attainment of near-native L2 speech proficiency not only to experience but also to aptitude (i.e., domain-general auditory processing), especially when the target structures entail much learning difficulty (e.g., Japanese speakers' English [r]-[l] acquisition; for details, see below). Here, the auditory precision hypothesis – L2 assumes that the underlying mechanisms in L2 speech learning are associated with learners' processing of sounds on domain-general pre-categorical, rather than domain-specific, speech levels (cf. Miyawaki et al., 1975).

Despite the growing amount of supporting evidence for the auditory precision hypothesis – L2, the extant studies have generally set out to collect broad measures of auditory processing by collapsing across various stimulus dimensions. Whereas individuals are known to differ in terms of the robustness of their abilities to process single acoustic dimensions (e.g., pitch, formants, duration, and amplitude; Kidd, Watson, & Gygi, 2007), it has remained unclear whether dimension-specific auditory processing might be particularly strongly linked to the acquisition of speech sound contrasts that are robustly distinguished by a particular dimension. To address this concern, we investigated the dimension-specific link between auditory processing and speech category acquisition in one specific, well-researched, and extremely difficult instance of L2 segmental learning - the acquisition of English [r]-[l] (e.g., "rock" vs. "lock") - by 100 late Japanese learners of English with diverse biographical, perception, and proficiency backgrounds.

1. Domain-general auditory processing in L1 acquisition

In order to master the perception and production of unique L2 speech sound contrasts and categories, language learners must first perceive small differences between sounds within auditory dimensions. Such pre-categorical auditory encoding abilities allow learners to detect the presence of multi-peaked distributions and assign separate peaks to separate linguistic categories (Toscano & McMurray, 2010). Individuals differ greatly in their abilities to encode various dimensions of sounds (e.g., pitch, formants, duration, and amplitude; Kidd et al., 2007). There is cross-sectional evidence for the correlations between individual differences in auditory processing in normal-hearing children and a range of language outcomes (e.g., speech-in-noise perception, vocabulary use, literacy, and phonological awareness; Anvari, Trainor, Woodside, & Levy, 2002; Bavin, Grayden, Scott, & Stefanakis, 2010; Boets, Wouters, Van Wieringen, De Smedt, & Ghesquiere, 2008; Douglas & Willatts, 1994; Lamb & Gregory, 1993; Talcott et al., 2000; Tierney, Gomez, Fedele, & Kirkham, 2021). Furthermore, Kalashnikova, Goswami, and Burnham's (2019) recent longitudinal investigation showed that auditory processing was a predictor of L1 vocabulary development within the first three years of life.

In contrast, when children lack precise perception of acoustic dimensions, it may lead to greater sensory overlap between linguistic categories, slowing down the acquisition of phonological knowledge and possibly other varieties of linguistic knowledge as well. Children with developmental language delays have been demonstrated to be more likely to show a range of difficulties with the perception of abstract non-verbal sounds (e.g., impaired frequency modulation detection, Talcott et al., 2000; Wright & Conlon, 2009; less precise frequency discrimination, Ahissar, Protopapas, Reid, & Merzenich, 2000; McArthur & Bishop, 2004, 2005; less precise perceptual separation of sound and noise bursts, Montgomery, Morris, Sevcik, & Clarkson, 2005; Rosen & Manganari, 2001; Wright et al., 1997; less precise perception of temporal patterns, Casini, Pech-Georgel, & Ziegler, 2018; Goswami et al., 2002).

Despite this large body of work investigating the relationship between auditory processing and language learning, whether auditory processing plays a causal role in language acquisition remains under question for several reasons. First, research on children with sensorineural hearing loss has found conflicting results, with some reports suggesting that they are at greater risk for developing a language delay (Halliday & Bishop, 2005; Halliday, Tuomainen, & Rosen, 2017), but others suggesting no consequences for language learning (Halliday & Bishop, 2006). Overall, only some children with auditory processing difficulties also show delayed language acquisition, suggesting that auditory impairment is a risk factor that can make language delays more likely, but not a foregone conclusion. Second, auditory processing tests are not pure measures of perception, but also draw on a range of cognitive skills. Some researchers therefore suggest that the link between auditory processing and language may instead reflect the influence of conflating factors such as attention (Snowling, Gooch, McArthur, & Hulme, 2018) or statistical learning (Jaffe-Dax, Frankel, & Ahissar, 2017). Third, while theories regarding the relevance of auditory processing for language learning assume that impaired auditory processing can lead to impaired speech perception, there is surprisingly little evidence supporting this hypothesis. Admittedly, the hypothesis is difficult to test in L1 speech perception, given that even young children are at the ceiling of perception of native speech sound contrasts. However, L1 speech perception can be brought below ceiling even in adults via the addition of background noise, and prior research has found that non-verbal auditory perception and perception of L1 speech in noise are generally uncorrelated (Surprenant & Watson, 2001; Watson, Jensen, Foyle, Leek, & Goldgar, 1982). Moreover, even individuals with severe difficulties in perception of a single acoustic dimension do not experience major disadvantages in speech perception and production. For example, individuals with amusia, who have extreme difficulty with pitch perception but preserved perception of other acoustic dimensions, report no difficulties with speech in everyday life (Liu, Patel, Fourcin, & Stewart, 2010).

In summary, prior research on the relationship between auditory processing and L1 language skills has produced two seemingly contradictory findings: difficulties with auditory processing tend to be linked to L1 developmental language delays, but not to difficulties with L1 speech perception, especially in adulthood. How can these findings be reconciled? First, speech tends to be *redundant*, meaning that a given linguistic feature tends to be conveyed by many different acoustic dimensions simultaneously (Lisker, 1986). Moreover, skill in perception of different auditory dimensions can vary independently, meaning that impaired perception of one dimension can coincide with preserved perception of other dimensions (Karlin, 1942; Kidd et al., 2007; Stankov

& Horn, 1980; Surprenant & Watson, 2001; Watson et al., 1982). One possibility, then, is that difficulties with the perception of one acoustic dimension could be compensated for by utilizing alternate sources of information in the speech signal. This structural redundancy model of speech perception is supported by Jasmin, Dick, Holt, and Tierney (2020), who found that individuals with difficulty processing and remembering pitch compensate by relying upon other acoustic cues more heavily (e.g., duration) to perceive music and speech. However, as these compensatory strategies [1] take time to develop, auditory processing difficulties may have more severe consequences in the early stages of language acquisition. In the current study, we take a first step towards testing this hypothesis by examining the relationship between individual differences in auditory processing and the speed and ultimate attainment of L2 acquisition in adulthood.

2. Domain-general auditory processing in L2 acquisition

Extending the auditory precision framework, originally developed to explain the source of developmental phenomena in L1 acquisition, some researchers have argued that auditory processing could be even more important when applied to post-pubertal L2 speech learning (i.e., auditory precision hypothesis – L2;Kachlicka et al., 2019; Saito, Kachlicka, Sun, & Tierney, 2020). Adult L2 learners tend to lack sufficient communicatively authentic conversation opportunities with native speakers to form accurate phonemic representations of the target language, even if they stay in L2-speaking countries (Derwing & Munro, 2013). The difficulties caused by limited input could be compounded if the learners also struggle with auditory processing, which would further constrain the extent to which they could utilize and process the already limited auditory input. Moreover, with respect to the quality of experience, adult L2 learners differ from infants learning their L1 in that their auditory representations are already fully established and fine-tuned to their L1's phonetic systems. When learners are exposed to new L2 speech sounds, not only do they need to accurately encode the auditory input, but they also must revise and refine their perceptual strategies to recognize and categorize the new segmental and suprasegmental features of the new language (see McAllister, Flege, & Piske, 2002 for the relative weights of spectral and temporal cues for different instances of L2 vowel acquisition). This may place additional demands on auditory processing, requiring listeners to focus on dimensions (and particular values within dimensions) that they have grown used to ignoring.

Growing evidence supports the existence of a significant relationship between auditory processing and adult L2 speech learning. Short-term training studies have shown that learners with more precise auditory processing are likely to show larger gains when receiving training on new sounds that they have never, or rarely, heard before (e.g., Wong & Perrachione, 2007). As for the outcomes of long-term L2 speech learning in immersive settings, individual differences in auditory processing demonstrate medium-to-strong associations with various aspects of adult L2 speech learning, even after biographical variables (e.g., age of acquisition, length of residence, the frequency of L2 use) are controlled for. These linguistic domains include L2 phonology (Kachlicka et al., 2019; Saito, Kachlicka, Sun, & Tierney, 2020), L2 vocabulary (Silbert et al., 2015), and L2 grammar (Kachlicka et al., 2019; Saito et al., 2020). When the analyses concern classroom L2 learners without much experience abroad, however, the influence of auditory processing remains unclear (e.g., Silbert et al., 2015), arguably because these learners may have limited access to auditory input to process, which is characteristic of foreign language settings (Muñoz, 2014).

[1] In the current study, strategies are defined as listeners' idiosyncratic ways of accessing, orchestrating, and weighting a range of domain-general acoustic cues (e.g., pitch, formants, and duration) while perceiving nonverbal and/or speech sounds (Jasmin, Dick, et al., 2020, Jasmin, Sun, & Tierney, 2020).

2.1. Motivation for the current study

Although we have accumulated substantial empirical evidence indicating a significant relationship between auditory processing and post-pubertal L2 speech learning, it is noteworthy that the existing studies have captured both audition and acquisition at a *broad* level. L2 speech proficiency was assessed via global measures, such as composite speech perception and grammaticality judgement tasks (e.g., Kachlicka et al., 2019), and linguistically trained coders' global evaluation of pronunciation accuracy and vocabulary appropriateness (e.g., Saito, Kachlicka, Sun, & Tierney, 2020). Auditory processing was operationalized as averaged scores across multiple discrimination tasks, wherein participants were tested on a wide range of acoustic dimensions (duration, amplitude, pitch, and formant frequency) with performances averaged to form composite measures (see Kachlicka et al., 2019 for methodological details). In essence, we have yet to determine whether, and to what degree, the ability to perceive fine differences along the specific dimensions of sensitivity to duration, amplitude, pitch, and formants relate to acquisition of knowledge about specific phonological contrasts.

To our knowledge, Chandrasekaran, Sampath, and Wong (2010) is the only exception, providing a detailed picture of the relationship between auditory processing of a specific dimension and L2 speech learning behaviours. In their training study, non-tonal language users (L1 English speakers) received several hours of explicit training on how to use lexical tones to learn novel words in an artificial language. According to the results, those with greater learning gains demonstrated more precise domain-general auditory processing of the most relevant cue: identification and discrimination of pitch direction, rather than height. Building on this line of work on L2 *suprasegmental* sensitivity and *short-term* learning, the current study set out to examine the mechanisms underlying one well-researched instance of L2 *segmental* sensitivity and *long-term* learning: the acquisition of English [r] and [l] among 100 Japanese speakers in classroom and immersion settings in comparison to 10 native speakers of English.

Japanese speakers' English [r]-[l] acquisition is considered to be one of the most difficult instances of adult L2 speech learning, and it can serve as "a productive testing ground for general principles of learning and claims about adult neural plasticity" (Bradlow, 2008, p. 294). According to the Perceptual Assimilation Model (Best & Tyler, 2007), Japanese speakers' acquisition of English [r]-[l] is an example of Single Category acquisition; the two phones are perceived as poor exemplars of the Japanese alveolar tap [ɾ] (Guion, Flege, Akahane-Yamada, & Pruitt, 2000). Indeed, it has been shown that attaining high-level proficiency requires a tremendous amount of immersion experience (e.g., for 20+ years, Flege, Takagi, & Mann, 1996) and/or many hours of intensive training (e.g., for 30+ hours, Bradlow, Pisoni, Akahane-Yamada, & Tohkura, 1997).

One reason why acquisition of the English [r]-[l] contrast is notoriously difficult for Japanese learners is that it requires them to learn to focus on acoustic information that is not generally relevant for speech sound categorization in Japanese. English [r] and [l] differ in the frequencies of formants F3 (1900–2000 Hz vs. 2400–2800 Hz) and F2 (1200–1400 Hz vs 1600–2000 Hz), as well as in duration (50–100 ms vs. 5-20 ms). However, the most reliable and salient information across various phonetic and lexical contexts is provided by F3 (Espy-Wilson, 1992). Since Japanese speakers do not use F3 variation as a primary acoustic cue for differentiating their L1 sound contrasts, they are used to ignoring this dimension. Indeed, they show much difficulty in the perception of synthesized English [r]-[l] exemplars differing in F3 dimensions (Iverson et al., 2003). Moreover, the task of acquiring a robust F3 representation is resistant to the effects of short-term training (Ingvalson, Holt, & McClelland, 2011).

To remedy the extreme difficulty in perceiving English [r]-[l] based on F3 information, it has been reported that Japanese speakers rely on F2 and duration cues, a strategy which is suboptimal (Ingvalson,

McClelland, & Holt, 2011). This non-nativelike strategy could be ascribed to the fact that Japanese speakers typically rely on F2 and duration variation to differentiate approximant and vowel contrasts in their L1 (e.g., [j]-[w] for F2, long-short vowels for duration). However, it is important to note that the alternative cues may not be as reliable (relative to F3 variation) as they are highly context-dependent. For example, although F2 in English [r] is lower than English [l] especially when the sounds precede back vowels ("road-load"), F2 in English [r] and [l] preceding front vowels could be relatively high, rendering Japanese speakers' reliance on partial cues (F2 and duration cues) ineffective and suboptimal ("read-lead"). F2 in English [l] could also be low when it is velarized ("clear vs. dark [l]"; see Flege, Takagi, & Mann, 1995 for the role of phonetic contexts in Japanese speakers' English [r]-[l] acquisition).

2.2. Research objectives and predictions

Comparing 100 Japanese speakers' English [r]-[l] proficiency (perception and production) and relevant auditory profiles (F3 and F2 sensitivity), we aimed to investigate how the precision of processing of individual acoustic dimensions relates to L2 speech proficiency. According to the L2 acquisition literature, the nature of L2 speech learning experience is essentially different in classroom settings (input is restricted to several hours of instruction per week) versus naturalistic settings (there is ample, context-rich, and interactive input; Muñoz, 2014). To cover a range of adult L2 speakers, our dataset comprised 50 English-as-a-Foreign-Language (EFL) learners in Japan without any experience overseas and 50 experienced Japanese residents in English-speaking countries (length of residence [LOR] = 1–20 years).

Adult L2 speech proficiency is commonly defined as a multifaceted phenomenon. Learners must decode, discriminate, and identify acoustic and articulatory features, and in addition convert them to abstract representations to revise existing speech categories and/or develop new speech categories on phonetic and phonological levels (Best & Tyler, 2007 for Perceptual Assimilation Model; Flege & Bohn, 2021 for Speech Learning Model). Further, such proficiency needs to be assessed on both perception and production levels (e.g., Nagle & Baese-Berk, 2021) via a range of task modalities (e.g., Saito & Plonsky, 2019 for controlled vs. spontaneous; Leong, Price, Pitchford, & van Heuven, 2018 for with vs. without noise; Toscano, McMurray, Dennhardt, & Luck, 2010 for categorical vs. gradient). While we acknowledge that a range of possible task options are available, Japanese speakers' composite speech proficiency in the current investigation was operationalized and analyzed via two different tasks. Following the methodological practices in the existing literature on Japanese speakers' English [r]-[l] acquisition in instructed (e.g., Bradlow et al., 1997 for perception; Saito, 2013a for production) and naturalistic settings (Ingvalson, McClelland, & Holt, 2011 for perception; Saito, 2013b for production), perception was operationalized as participants' abilities to perceptually identify contrasting natural phonemes in a minimal pair context (i.e., forced-choice identification task) and produce the target phonemes while using language freely, naturally, and spontaneously (i.e., timed picture description task). For the use of other relevant tasks and the potential influence of such tasks on the findings, see the Future Directions section.

It is important to point out some evidence that Japanese speakers acquire English [r] and [l] at different rates (see Aoyama, Flege, Guion, Akahane-Yamada, & Yamada, 2004). According to the cross-linguistic analyses of English [r], English [l], and the Japanese tap [ɾ], English [r] could be clearly distinguishable from Japanese [ɾ] with lower F2, F3 and longer duration. Yet, there is much overlap between English [l] and Japanese [ɾ] (Hattori & Iverson, 2009). In existing studies (e.g., Bradlow, 2008 for a methodological review), many researchers have approached this topic, analyzing how Japanese speakers can perceive and produce the difference between English [r] and [l] in minimal pair contexts. In the current investigation, the main focus of the analyses lay in Japanese speakers' English [r]-[l] proficiency. Following the methodological practices, their proficiency was operationalized as the identification and spontaneous pronunciation of English [r] and [l] minimal pairs. However, we did not conduct any separate scrutiny of the extent to which Japanese speakers differentially acquired English [r] and [l] sounds and distinguish the English [r]-[l] contrast.

With respect to predictions, as stated in the SLM-r (Flege & Bohn, 2021), post-pubertal Japanese speakers can attain advanced English [r]-[l] proficiency especially after long-term residence in an English-speaking environment. That is, experienced Japanese speakers were assumed to show approximately 80–90% identification accuracy (e.g., Flege et al., 1996) and highly intelligible pronunciation forms (e.g., Flege et al., 1995). At the same time, however, their English [r]-[l] proficiency would also be characterized as individual variation with some achieving almost nativelike L2 perception (> 90% accuracy) and production (judged to be highly targetlike without any trace of L1 Japanese). Here, we speculated that the incidence of near-native English [r]-[l] proficiency could be not only linked to participants' experience profiles (extensive immersion) but also to their different auditory abilities to process the three acoustic cues relevant to the English [r]-[l] contrast (F3, F2, and duration; Doughty, 2019). This could be arguably because the degree of Japanese speakers' high-level English [r]-[l] acquisition could be essentially driven by their domain-general (rather than domain-specific, phonetic and phonological) processing of sounds (cf. Miyawaki et al., 1975).

Whereas acquisition of English [r]-[l] requires Japanese learners to make use of information from F3 as a *primary* cue, they are accustomed to ignoring this cue as it is irrelevant when it comes to the acoustic correlates of L1 Japanese approximant categories (Iverson et al., 2003). Thus, we predicted that perception and production of this speech sound contrast would be primarily linked to performance on a test of discrimination of sounds based on F3 frequency. Notably, L1 acquisition literature has begun to show that the acquisitional function of auditory dimension skills is *dimension-specific*. Even if one has difficulty in perception of a primary dimension, they can still perceive sound contrasts owing to their capacity to attend to other relevant (though secondary, less reliable) cues (Jasmin, Dick, et al., 2020). However, the development of such compensatory strategies may require a substantial amount of conversational input. Thus, we predicted that some Japanese speakers with greater F2 and duration sensitivity may attain near-nativelike English [r]-[l] proficiency regardless of their individual differences in F3 sensitivity, especially after they have accumulated more naturalistic and immersion experience in English-speaking environments. The predictive power of F2 sensitivity could be observed more clearly among experienced Japanese learners of English ($n = 50$) but not among EFL learners in Japan ($n = 50$).

3. Method

3.1. Participants

3.1.1. English-as-a-foreign-language & experienced L2 learners

A total of 50 freshman EFL learners were recruited at a private university in Japan (3 females, 47 males). While the project was advertised in a range of English classes at the university, the first 50 volunteer participants were included in the analyses. All of them were engineering majors with no experience living overseas at the time of the project.

3.1.2. Experienced L2 learners

A total of 50 Japanese residents (40 females, 10 males) were recruited in the UK ($n = 38$) and the USA ($n = 12$) using the same data collection platform and procedure as for the EFL learners. Digital flyers were posted on multiple community websites, clarifying the following recruitment criteria. Since our interest lay in naturalistic L2 speech learning, potential recruits had to reside in one of the three English-speaking countries (UK, USA, or Canada), and their age of arrival had to be beyond 18 years old. At this stage, over 80 interested participants

contacted our team. Given that the main focus of the project concerned the analyses of regular and active Japanese users of L2 English, we carefully selected such participants. The data collection was stopped when we reached $n = 50$ experienced Japanese participants. A total of five trained assistants individually met all the candidates for a background screening interview. The quantity and quality of the participants' L2 experience in two different contexts (EFL vs. immersion) were detailed in the Measures of L2 Experience section (Section 3.7).

3.1.3. L1 speakers

To provide the nativelike baseline data, a total of 10 native speakers of American English were recruited (5 males, 5 females; $M_{age} = 25.2$ years, *Range* = 22–28) to engage in speech and auditory processing tests. To this end, a digital flyer was posted on several community websites and social media. For each interested participant, a researcher had an individual online meeting to provide guidance and instruction on how to use the Gorilla platform.

3.2. General procedure

Due to the current global pandemic, testing and interview sessions took place online via the Gorilla platform (Anwyl-Irvine, Massonnié, Flitton, Kirkham, & Evershed, 2020). To monitor participants' performance, a team of research assistants were trained to assist participants in understanding, implementing, and completing each task. After a series of pilot studies, we revised, streamlined, and refined the online data collection procedure to accommodate the pros and cons of the platform. The assistants communicated with participants via email, met each participant online via a video-conferencing tool, surveyed their L2 experience backgrounds, and checked their eligibility for the project. To help participants familiarize themselves with the task procedure, and complete them online without any confusion, they first joined a short practice session (5 min) wherein they listened to a short English speech (not included in the pre- and post-tests) and recorded their voices to check the volumes, microphones, internet speed, and sound systems on their computers. After we confirmed that the participants' performance was recorded correctly, and that they had a clear understanding of the procedure, they proceeded to the main tasks (speech perception/production, auditory processing) and completed them with the guidance from the assistants. Whenever participants had questions and/or encountered problems, they could consult with the assistants who remained online with their microphones muted for the duration of the experimental session. While the individual data collection sessions were done in English (as the assistants were non-Japanese), all the task instructions were displayed in Japanese on the computer screen. Prior to the individual sessions, the participants were also given an electronic brochure which detailed the objectives of the project and the task procedures in Japanese.

In the current study, participants' L2 speech proficiency was determined by assessing the accuracy of their perception and production of the difficult sound contrast between English [r] and [l] via a forced-choice identification task (Bradlow et al., 1997) and a timed picture description task (Saito, 2013a). Auditory processing was evaluated by way of three dimension-specific AXB discrimination tasks. To minimize the extent to which participants explicitly focused on the target phonetic contrast (English [r] and [l]) during the picture description task, participants engaged in the tasks in the following order: (a) picture description task, (b) identification task, and (c) auditory processing task.

3.3. Measures of auditory processing

In previous studies investigating the role of perception of F3 in English [l]-[r] categorization, scholars assessed L2 learners' sensitivity to, and reliance on, F3 when it was manipulated in synthesized speech within the context of other synthesized phonemes (e.g., Iverson et al., 2003; Ingvalson, Holt, & McClelland, 2011; Ingvalson, Holt, & McClelland, 2012). Our goal here was different. We set out to test the hypothesis that the ability to discriminate sounds based on the frequency of the third formant would be related to skill in perception of [l] and [r], even when the sounds to be discriminated were abstract non-verbal sounds presented in isolation (i.e., not in the context of other phonemes). As such, the sounds for the auditory processing tests were created by synthesizing a complex tone with a static fundamental frequency of 100 Hz, 60 harmonics, 10 ms linear amplitude ramps at onset and offset, and three formants imposed using a parallel formant filter bank (Smith, 2007). The resulting sounds were clearly *artificial* and did not sound like natural speech. As described below, the sounds were characterized with completely flat F0 and formant contours, flat harmonic spectrum, and only three formants. Importantly, only one single target dimension (F3, F2, or duration) was manipulated at a time in the discrimination task. This crucial aspect of the tasks made them "auditory" tasks rather than speech perception tasks. While naturalistic speech stimuli could be dynamic and acoustically complex, with multiple cues that could be used to detect a particular speech feature, the stimuli used in the current task were static and comparatively simple, forcing participants to detect a specific acoustic cue.

For each of the three tests, a continuum of 200 stimuli was created, spanning a range of values along the target dimension (F3 frequency, F2 frequency, and duration). At each trial, three tones were presented in an AXB format, with X matching either tone A or B. Participants were tasked with identifying the odd tone by either clicking a button marked "1" or "3". The initial difficulty of the task was relatively low, with a large distance between the baseline and target stimuli (100 steps along the target continuum). Task difficulty was manipulated dynamically via an adaptive staircase method by decreasing the distance between target and baseline stimuli after two consecutive correct answers and increasing the distance between target and baseline stimuli after a single incorrect answer.

Moreover, over the course of the test, the degree to which the task difficulty changed after correct and incorrect answers gradually changed, starting out relatively large and then becoming smaller over time. This enabled the algorithm to gradually "zero in" on the difference between stimuli that participants could just barely hear. Specifically, the step size changed after each "reversal". A trial counted as a reversal when, after a period in which the task gradually became more difficult, the participant logged an incorrect response, resulting in the task becoming easier again. Alternately, a reversal was also marked when, after a period in which the task gradually became easier, the participant got two answers correct in a row. The step sizes were 20, 10, 4, 1, 1, 1, 1 for the first through seventh reversals, respectively. The task ended after seven reversals or seventy trials, whichever came first. The baseline level was set randomly on a trial-by-trial basis to a value anywhere within the continuum, while the target level was above baseline (with the constraint that the baseline level needed to be low enough that the target level fell within the continuum). Performance was calculated as the average of the task difficulty levels at all reversals from the second onward.

Test-specific stimulus parameters were as follows. For the F2 discrimination test, F2 varied in 200 equal mel-scale steps from 1200 to 1500 Hz, while duration, F1, and F3 were fixed at 100 ms, 478 Hz, and 2371 Hz, respectively. For the F3 discrimination test, the duration of the stimuli was fixed at 100 ms, F2 was fixed at 1088 Hz, and F3 varied in 200 equal mel-scale steps from 1601 to 3400 Hz. For the duration discrimination test, the duration of the stimuli varied linearly in 200 steps from 20 to 60 ms, while F1, F2, and F3 were fixed at 478, 1345, and 2371 Hz, respectively. According to the results of prior validation studies, test-retest reliability of the discrimination tasks was found to be relatively high in the context of L1 acquisition ($r = 0.75$ in Raz, Willerman, & Yama, 1987) and L2 acquisition ($r = 0.701$ in Saito, Sun, & Tierney, 2020).

3.4. Follow-up experiment

To assess Japanese speakers' domain-general auditory acuity to three relevant cues for the English [r] and [l] contrast, nonverbal sounds were used as stimuli. Following and extending the methodological paradigm in Guion et al. (2000), two preliminary experiments were conducted to assess two methodological features of the auditory processing tests: (a) to what degree the nonverbal stimuli could be perceived as English phonemes (category goodness judgements); and (b) to what degree they were perceived by listeners as speech versus non-speech (speech-likeness judgements).

3.4.1. Participants

A total of 10 native listeners of English ($n = 5$ for General American; $n = 5$ for Received Pronunciation) were recruited via the Prolific recruitment platform. 6 listeners had prior linguistic training at post-secondary school levels and 3 reported having English language teaching experience. They reported no hearing difficulty. At their convenience, the listeners participated in the category goodness and speech-likeness judgements in this order. While using their own computers in a quiet room, they completed the judgement sessions via the online data collection platform, Gorilla (Anwyl-Irvine et al., 2020).

3.4.2. Materials

A total of 24 nonverbal stimuli (used in the auditory processing tests) and 8 speech stimuli (not included in the current study) were prepared for the judgement sessions. For the nonverbal stimuli, 12 represented the most [r]-like exemplars from the F3 discrimination task (with top 4 lowest F3 values selected from 200 stimuli created with equal mel-scale steps from 1601 to 3400 Hz), the F2 discrimination task (with top 4 lowest F2 values selected from 200 stimuli created with equal mel-scale steps from 1200 to 1500 Hz) and the duration discrimination task (with top 4 longest duration values selected from 200 stimuli varying linearly from 20 to 60 ms); and 12 represented the most [l]-like exemplars from the F3 discrimination task (with top 4 highest F3 values), the F2 discrimination task (with top 4 highest F2 values) and the duration discrimination task (with top 4 shortest duration values). For the speech stimuli, the open syllables (English [ra] and [la]) were extracted from natural words ("rock" and "lock") produced by two male and two female native speakers of General American. The investigator carefully listened to the target words, extracted only [ra] and [la], and saved them as WAV files. Each speech sample was presented 3 times to match the number of non-speech trials (i.e., 24 in total).

3.4.3. Category goodness judgements

First, they familiarized themselves with the stimuli by listening to all of them. Then, they listened to a stimulus played in a randomized order, and categorized it into one of the seven categories:

1. English [r] sound (e.g., rock, read)
2. English [l] sound (e.g., lock, lead)
3. English [w] sound (e.g., wood, weed)
4. English [y/j] sound (e.g., year, yield)
5. Other English vowels (e.g., i, e, a, o)
6. Other English consonants (e.g., s, z, b, v, th)
7. None of the above

Upon each category judgement, they then rated the degree of English-likeness on a 7-point scale (*7 = English-like; 1 = not English-like at all*). According to the results, all the listeners not only consistently categorized the speech stimuli as "English [r]"and "English [l]", but also perceived them as more English-like exemplars ($M = 5.51$–5.58 out of 7; *English-like*). In contrast, they categorized the nonverbal stimuli either as "none of the above" or "Other English vowels" with significantly lower goodness ratings ($M = 2.61$–2.85 out of 7; *not English-like at all*).

3.4.4. Speech-likeness judgements

After the category goodness judgement, they listened to the same stimuli (24 nonverbal stimuli, 8 speech stimuli) in a randomized order, and rated each sample for the degree of speech-likeness on a 10-point scale (*10 = Definitely human speech, 1 = not human speech-like at all*). Whereas the speech stimuli were perceived as "definitely speech" ($M = 8.63$–8.7 out of 10), the nonverbal stimuli were considered as "not human speech-like at all" ($M = 2.02$–2.12 out of 10).

Taken together, the follow-up experiments confirmed that although the nonverbal stimuli used in the auditory processing tests represented the key acoustic dimensions to the English [r] and [l] contrast (F3, F2 and duration discrimination), they were not perceived as any of the target English phonemes ([r] and [l]) nor human speech.

3.5. Measure of L2 speech production proficiency

3.5.1. Speaking task

Given that adult L2 speakers can monitor the accuracy of their production when their performance is tested via controlled tasks (e.g., word and sentence reading), scholars have emphasized the importance of assessing L2 production proficiency by use of more spontaneous, ecologically-valid free-speech tasks so as to index how they actually produce L2 sounds in real-life conversational settings (Piske, MacKay, & Flege, 2001). The former task has been found to allow same L2 learners to demonstrate more nativelike pronunciation than the latter task (Major, 2001). In the current investigation, following the procedure in Saito (2013a, 2013b), a timed picture description was adopted wherein participants described a series of pictures with time limit (5 s for planning and 30 s for picture description). To help low-proficient L2 learners produce a certain length of spontaneous L2 speech without much dysfluency, three word prompts were provided per photo. There were a total of 20 pictures: $n = 10$ for the main analyses (experimental stimuli) and $n = 10$ for distracters (control stimuli). For the experimental stimuli, one of the three word prompts featured the target singletons, wherein English [r] and [l] appeared on word-initial positions. As such, participants were guided to produce English [r] (*read, race, row, rock, wrong*) and [l] (*lead, lace, low, lock, long*). The control stimuli highlighted another difficult phonological contrast for Japanese speakers of English (i.e., [æ]-[ʌ]; *mad, mud, bad, bud, fan, fun, cap, cup, cat, cut*).

The 10 experimental stimuli were carefully chosen taking into account the lexical and phonetic status of the words. Since word frequency was found to affect L2 speech performance (Flege et al., 1995 for production, and 1996 for perception), all target lexical items were carefully chosen from a list of the most common 3000 words, according to the Lexical Tutor (Cobb, 2000). As such, they can be assumed to be frequent and familiar words to the participants, which at the very least should have been learnt as minimum requirements at high schools across Japan (McLean, Hogg, & Kramer, 2014). To control for the influence of phonetic context on L2 speech perception, the place of the following vowels was equally distributed in terms of height and backness (see Supplementary Information).

3.5.2. Task procedure

The task was presented on the Gorilla online platform (Anwyl-Irvine et al., 2020). Participants first received instruction from a trained assistant on the procedure. To ensure that their microphone and sound system functioned properly, a check-up function was set up in Gorilla, wherein participants recorded and listened to their own speech (10 s). Next, they practiced the procedure by engaging in three picture descriptions (excluded from the data analyses). After they familiarized themselves with the procedure, they proceeded to the main task of 20 picture descriptions.

The pictures were delivered in a randomized order. For each picture, one of the word prompts featured a target word. In Fig. 1, for example, "read" is a target singleton (for a full list of the target words, see Supporting Information). To indicate the amount of time left for planning (5

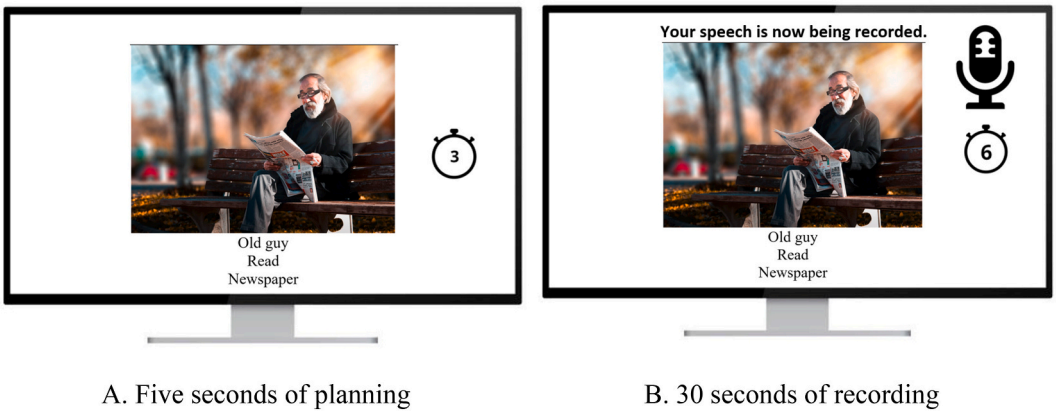

Fig. 1. Screenshots of online timed picture description task.

s; Fig. 1A) and task completion (30 s; Fig. 1B), a timer was displayed on the righthand corner of the screen throughout task completion. To provide ample speech data for analyses, participants were explicitly told to produce a few sentences while explaining each picture (instead of reading aloud only target words on the screen).

For technical reasons, two participants failed to record and upload the data to the online platform; moreover, the quality of two participants' production data was judged to be not optimal for listener judgements (e.g., substantially noticeable noise). To provide a baseline estimate of nativelike English [r] and [l] production, 10 native speakers also completed the same production tasks. As a result, the production data of 106 participants (96 Japanese, 10 English) were submitted to the listener analyses.

3.5.3. Listener judgement materials

In many previous studies, Japanese speakers' English [r] and [l] production was elicited via a controlled speech task (word and sentence reading). To help listeners assess only the quality of English [r] and [l], Flege and his colleagues (e.g., Aoyama et al., 2004; Flege et al., 1995) removed most of the vowels and all of the final consonants in English words, leaving just a few glottal pulses following the initial liquid sounds. In the current investigation, however, English [r] and [l] exemplars were embedded in spontaneous speech (where the preceding and following phonetic contexts were not controlled).

Following the procedure for the analyses of spontaneous English [r]-[l] production in Saito (2013a, 2013b), a trained researcher carefully listened to each sample and cut and saved isolated target words including English [r] and [l] in a WAV file in order to avoid introducing too much distortion on rating samples. Efforts were made to ensure that the samples sounded as natural as possible. The researcher put a cursor on the onset of the target word (where any component of [r] and [l] could be heard) and moved towards it by 5 ms steps. The 1060 tokens (5 [r] tokens and 5 [l] tokens × 96 Japanese participants and 10 native English) were normalized for peak amplitude.

To avoid the influence of lexical status (word frequency and familiarity) on listeners' judgements, two methodological decisions were made. First, English [r] and [l] were featured within the most common 3000 words to minimize the influence of the lexical frequency and familiarity of the target items on listeners' judgements (see above). Second, we recruited only linguistically trained listeners who had similar L2 speech analysis experience so that they could fully focus on the quality of English [r] and [l] without being influenced by the lexical factors (see below).

3.5.4. Listeners

A total of five linguistically trained speaking listeners (3 males, 2 females) were recruited. All of them were native speakers of English in the UK and reported an extensive amount of linguistics training (with BA, MA, or PhD degree in linguistics) and/or teaching experience (M = 4 years, *Range* = 3–6 years). The listeners were carefully selected based on their prior experience in similar L2 speech analyses. Due to the pandemic situation, the listening sessions individually took place online and included a preparatory session with a trained researcher. After the training, the raters listened to all the stimuli presented in randomized order via the Gorilla online platform (Anwyl-Irvine et al., 2020).

3.5.5. Rating procedure

Listeners were told that all tokens comprised a set of minimal pair words including either English [r] or [l] at a word-initial position, and they were produced by Japanese learners of English with various proficiency levels or by native speakers of English. They were asked to make their judgements as much as they could by focusing on only the quality of English [r] and [l] component—instead of the entire words. This instruction was crucial in order to avoid the influence of other pronunciation errors typical of Japanese learners on their ratings (e.g., [rɪd] for [rid]), including suprasegmental errors (e.g., monotonous or non-targetlike lexical stress). A "repeat" button was available to allow the listeners to hear an item up to three times before making a judgement. They were told that the Japanese participants may have followed a different model of English pronunciation (General American vs. Received Pronunciation). The listeners were asked to evaluate Japanese speakers' English [r] and [l] forms but without taking into account regional varieties. The listeners reported a great deal of familiarity with both pronunciation models owing not only to their relevant linguistic training experience (e.g., listening to a wide variety of languages including American and British English), but also to frequent encounters via mass media. [2]

After rating three familiarization stimuli not included in the subsequent listening sessions, and checking the sound quality of the platform, the listeners judged the stimuli in two blocks with 5-min breaks at the halfway point. The entire listening session took approximately 90 min.

[2] The influence of different English dialects on the listeners' English [r] and [l] production judgements could be minimal in the current study. Native speakers of English produce [r] differently, such as bunched [r] (i.e., a raised tongue tip and lowered dorsum) and retroflexed [r] (i.e., a raised dorsum and lowered tip), even within the speech of the same speaker (Delattre & Freeman, 1968). Although their articulator positions are different, the two forms of English [r] (bunched vs. retroflexed) require speakers to make the simultaneous constrictions in the labial, palatal, and pharyngeal regions of the vocal tract. This creates an anterior oral cavity that includes the sublingual space, which in turn leads to F3 lowering and the perceived rhotocization. When it comes to English [r] in syllable initial positions (the target stimuli in the current study), the shared articulatory and perceptual features (i.e., the simultaneous constrictions for lower F3) are commonly observed across different regional dialects in English (Espy-Wilson, Boyce, Jackson, Narayanan, & Alwan, 2000).

3.5.6. Rating scales

In the current study, we adopted the 9-point rating procedure which was developed in Saito (2013a, 2013b). Revising and extending Flege et al. (1995), the nine-point descriptors were assumed to reflect the different developmental stages of Japanese speakers' English [r]-[l] acquisition. The listeners judged the quality of word-initial [r] and [l] by choosing one of the response alternatives:

- 1: Nativelike [r]
- 2: Good [r]
- 3: Probably [r]
- 4: Possibly [r]
- 5: Neutral exemplars, neither [r] nor [l]
- 6: Possibly [l]
- 7: Probably [l]
- 8: Good [l]
- 9: Nativelike [l]

The rubric here indexed the accuracy of each participant's English [r] production (i.e., lower is better) and English [l] production (i.e., higher is better). The neutral category (neither English [r] nor English [l]) was included in the midpoint (5 out of 9) because the category represents the interlanguage phase of adult L2 speech learning. As discussed in Major's (2008) Ontogeny Phylogeny Model, L2 learners initially tend to substitute their own L1 counterparts for L2 sounds (e.g., the Japanese tap for both English [r] and [l]; Guion et al., 2000). With increasing awareness of L2 sounds, however, L2 learners make some efforts to stay away from the L1 counterpart (Japanese [ɾ]) and make efforts to approximate the target features (English [r] and [l]) by using various interlanguage strategies (e.g., retracting and fronting tongue body; and lengthening and shortening phonemic length; Saito & Munro, 2014). The acoustic properties of such interlanguage pronunciation forms (F1, F2, F3, duration) appeared to stretch between English [r], English [l] and the Japanese tap (e.g., Lotto, Sato, & Diehl, 2004). In Polka and Strange's (1985) listening experiment with synthesized tokens on a rock-lock continuum, native English listeners perceived stimuli with intermediate spectral (F3, F2) and temporal (F1 transition duration) values as neither /ɹ/nor /1/, but rather as /w/or /d/. Therefore, the interlanguage exemplars could be perceived as neither /ɹ/ nor /l/ (i.e., Descriptor "5" *neutral exemplars*) and needed to be included in the midpoint as they index the mid state of L2 speech learning.

3.5.7. Rating agreement

The inter-listener agreement was relatively high for English [r] samples (Cronbach alpha = 0.93, Inter-Class Correlation = 0.86) and English [l] samples (Cronbach alpha = 0.91, Inter-Class Correlation = 0.85). Similar to Flege et al. (1995) and Saito (2013a, 2013b), the listeners' judgement scores were averaged across to derive a single score for each talker (96 Japanese speakers), averaged across both phonetic contexts (English [r] and [l]). Since the relationship between the magnitude of the scores and pronunciation accuracy was reversed for English [r] and [l], the scores for [r] were adjusted using the formula [10−n]. For example, Participant A received $M = 3.4$ for English [r], her adjusted score would be 6.6. As such, the averaged scores represented the extent to which each participant's production approximated nativelike English [r] and [l] distinction on a 9-point scale (*1 = not nativelike at all, 9 = nativelike*; for the details of the method, see Flege et al., 1995).

3.6. Measures of L2 speech perception proficiency

3.6.1. Materials

Following similar formats to (Ingvalson, McClelland and Holt, 2011), a forced-choice identification task was developed to assess participants' L2 speech perception proficiency. Participants were asked to listen to a total of 320 minimal pairs. The stimuli consisted of 160 experimental tokens (20 English [r]-[l] minimal pairs) and 160 distracter tokens (20 English [ɪ]-[ɛ] and [æ]-[ʌ] minimal pairs) spoken by eight native speakers of British English (four males, four females). Using the same criteria, the 20 English [r]-[l] minimal pairs were carefully selected. All of them were high frequency words. The following vowel conditions were equally distributed (see Supplementary Information). The tokens were recorded at a 40-kHz sampling rate and normalized for peak intensity using the *Praat* speech analysis software (Boersma & Weenink, 2010).

3.6.2. Procedure

The participants first received instruction on the test procedure from a trained research assistant. Before moving to the main experimental task, they practiced three sample questions (using three minimal pairs not included in the main test). Once they heard a stimulus, they were shown the minimal pair on the screen and asked to click the one they thought they had heard. The entire test lasted 30 min. According to the results of Cronbach alpha analyses, the test materials demonstrated high level reliability ($\alpha = 0.90$). The total accuracy scores (%) were used for the subsequent statistical analyses.

3.7. Measures of L2 experience

During the initial online interview, participants' experience-related variables were surveyed by trained assistants. Following and tailoring the existing methodological paradigms for classroom L2 speech learning (i.e., English-as-a-Foreign-Language Questionnaire; Saito & Hanzawa, 2016) and naturalistic L2 speech learning (i.e., Language Contact Profile; Freed, Dewey, Segalowitz, & Halter, 2004), we surveyed not only quantity but also quality variables which the existing literature has found to be related to the rate of learning success (for a summary of participants' L2 learning experience, see Table 1).

Table 1
Summary of participants' biographical backgrounds.

	EFL learners ($n = 50$)				Experienced learners ($n = 50$)			
	M	*SD*	*Range*		*M*	*SD*	*Range*	
			Min	*Max*			*Min*	*Max*
Age (years)	18.6	2.2	18	19	37.5	10.8	19	61
Length of EFL (years)	7.9	2.5	5	15	9.5	3.7	5	20
Pronunciation training (yes, no)	Yes = 17, No = 33				*N/A*			
Current EFL instruction (hours per week)	2.9	1.9	2	15	*N/A*			
Length of residence (years)	0	0	0	0	11.0	8.6	1	29
Age of arrival (years)	*N/A*				22.8	5.8	16	43
Current L2 use/ home (% per day)[a]	*N/A*				83.5	20.5	0	100
Current L2 use/ work on with L1 and fluent L2 speakers (% per day)	*N/A*				53.8	39.7	20	100
Current L2 use/ social with L1 and fluent L2 speakers (% per day)	*N/A*				71.3	24.9	15	100

Note.
[a] L2 use was restricted to aural modes (speaking, listening) with L1 and fluent L2 speakers.

3.7.1. EFL experience (total years of EFL education, hours of current EFL classes, pronunciation training)

Prior research has clearly shown that the outcomes of classroom L2 speech learning can be strongly related to the total years of EFL education (e.g., Muñoz, 2014) and the number of EFL classes that participants are currently registered for (Saito & Hanzawa, 2016). The predictive power of age of EFL learning remains in EFL contexts where L2 input is limited to a few hours of language-focused instruction per week (typically delivered by L2 teachers) without many conversation opportunities outside of classrooms (for a comprehensive overview, see Pfenninger & Singleton, 2019). According to the results of a background questionnaire, certain participants also reported having received some specific pronunciation training on the target L2 speech instance, English [r]-[l] ($n = 17$), allowing us to explore the effects of their explicit phonetic knowledge (for the role of explicit phonetic training and knowledge in L2 speech acquisition, see Saito & Plonsky, 2019 for a meta-analysis). Although we initially attempted to quantify the type of instruction (form vs. meaning) and EFL activities (speaking, listening, writing vs. reading), the participants reported highly homogeneous EFL experience (i.e., exclusively limited to grammar and vocabulary exercise through writing and reading activities). Thus, the variables were not further pursued.

3.7.2. Naturalistic L2 experience (length of residence, age of arrival, current L2 use)

While much research attention has been directed towards the role of age of arrival and length of residence in naturalistic L2 speech learning, there is a consensus that such variables can be used as a rough index of L2 input (Piske, MacKay, & Flege, 2001). The timing and length of immersion does not necessarily concur with how much L2 learners actually use a target language. There is some evidence showing that some prefer to use their L1 without many opportunities to use a L2 throughout their residence in a L2 environment (Derwing & Munro, 2013). As done in the existing studies (e.g., Flege et al., 1995, 1996; Saito, 2013b), the decision was made to include only regular and active L2 users as participants in the current investigation. All participants reported their main language of communication at home and/or work as being English. Thus, the amount of their received L2 input was assumed to be proportional to the length of residence (for a similar quantization of L2 input, see Flege & Wayland, 2019).

While length of residence/age of arrival could be used as an estimated *quantity* of L2 input that the experienced Japanese speakers had during their extensive periods of immersion, the extant research has also shown that the *quality* of L2 input that learners have engaged in at the time of data collection can greatly vary and relate to acquisition. For example, Freed et al. (2004) proposed that such quality variables should tap into how often L2 speakers are using a target language in accordance with (a) different types of contexts (home, work/school, and social settings), (b) activities (speaking, listening, writing, and reading), and (c) interlocutors (L1 speakers, fluent L2 speakers, and non-fluent L2 speakers). Using a tailored version of the Language Contact Profile (Freed et al., 2004), we surveyed the frequency of L2 use at the time of testing on aural modes (speaking and listening) especially with L1 and fluent L2 speakers (but not non-fluent L2 speakers) under three different conditions: home, work/school, and social settings as they were assumed to be most relevant to post-pubertal L2 speech acquisition.

4. Results

In this section, we first report how the Japanese speakers of English and native speakers differed in speech perception and auditory processing performance, and then how variance in the Japanese participants' L2 speech proficiency could be explained by factors related to EFL experience and auditory processing.

4.1. L2 speech proficiency (perception, production)

As summarized in Table 2, the results of 95% confidence interval (CI) analyses showed that Japan-based EFL speakers' identification of English [r] and [l] was around the chance level (50.6–55.4%). In terms of production, these speakers appeared to substitute the Japanese tap (somewhere between English [r] or [l]) for the English [r] and [l] contrast (listener ratings 4.9–5.2; i.e., unintelligible to neutral English [r] and [l]). The experienced Japanese speakers' performance was substantially more accurate in identification (81.0–88.03%) and production (6.6–7.1; i.e., possible to probable English [r] and [l]). Not surprisingly, native speakers' performance was at ceiling. According to the results of Kolmogorov-Smirnov tests, none of the perception and production scores were significantly different from the normal distribution at a $p < .012$ level (Bonferroni corrected) for the EFL Group ($D = 0.122$, 157, $p = .408$, 0.149) or the Experienced Group ($D = 0.196$, 0.116, $p = .035$, 0.473).

To check Japanese participants' English [r]-[l] proficiency according to Group (EFL, Experienced) and Context (Perception, Production), their scores were submitted to a two-way analysis of variance (ANOVA). For the purpose of comparability, the perception and production scores were standardized into *z* scores. The results yielded significant main effects of Group, $F(1, 94) = 241.989$, $p < .001$, $\eta^2 = 0.721$. Yet, neither main effects of Context, nor interaction effects of Group and Context reached statistical significance, $F(1, 94) = 0.169$, 0.848, $p = .690$, 0.359, $\eta^2 = 0.002$, 0.009. The results indicated (a) that the presence of immersion experience determined Japanese speakers' English [r]-[l] proficiency to a great degree ($d = 2.99$); and (b) that their perception and production scores closely aligned with each other within the current dataset ($r = 0.856$, $p < .001$).

4.2. Auditory processing

The descriptive statistics for Japanese and English speakers' auditory processing scores for F3 and F2 (0–200 points) were summarized in Table 3. Note that these scores are thresholds—the size of the difference between stimuli necessary for correct discrimination—and so lower scores indicate better performance. According to Kolmogorov-Smirnov tests, the scores of F3, F2, and duration were normally distributed for the EFL Group ($D = 0.082$, 0.090, 0.101, $p > .05$) and the Experienced Group ($D = 0.113$, 0.094, 0.084, $p > .05$). In terms of the interrelations among F3, F2, and duration processing scores, a Pearson correlation analyses did not find any relationships reaching statistical significance at a $p > .016$ level (Bonferroni corrected), while the EFL participants' F3 and F2 scores were marginally positively correlated ($r = 0.279$, $p = .049$).

Next, we examined whether Japanese and English speakers differed in discrimination of sounds based on F3, F2, and duration via a set of Kruskal-Wallis tests. The non-parametric tests were chosen as the three groups were substantially different in size (50 EFL, 50 Experienced, 10 Native). The results showed that the three groups' performance was

Table 2
Descriptive statistics of Japanese and English Speakers' English [r] and [l] perception and production proficiency.

	M	*SD*	*95% CI*	
			Lower	*Upper*
A. English [r]-[l] perception proficiency (%)				
Japanese EFL learners ($n = 50$)	53.0%	8.4	50.6	55.4
Experienced Japanese learners ($n = 50$)	84.5%	12.2	81.0	88.0
English L1 speakers ($n = 10$)	100%	0.0	100	100
B. English [r]-[l] production proficiency (9-point)				
Japanese EFL learners ($n = 46$)	5.0	0.5	4.9	5.2
Experienced Japanese learners ($n = 50$)	6.9	0.8	6.7	7.2
English L1 speakers ($n = 10$)	7.6	0.4	7.3	7.9

Table 3
Descriptive statistics of Japanese and English Speakers' auditory processing abilities.

	Acoustic dimension (0–200 points)	*M*	*SD*	95% *CI*	
				Lower	Upper
Japanese EFL learners ($n = 50$)	F3	102.3	47.9	88.6	115.9
	F2	87.9	45.0	75.1	100.7
	Duration	68.5	36.0	58.3	78.8
Experienced Japanese learners ($n = 50$)	F3	102.7	53.9	87.7	117.6
	F2	68.2	27.4	60.6	75.7
	Duration	59.5	31.2	50.8	68.1
English L1 speakers ($n = 10$)	F3	99.6	41.9	69.6	129.6
	F2	55.6	34.8	30.7	80.5
	Duration	82.1	34.1	57.6	106.5

Note. Smaller values in auditory processing measures (0–200 points) indicate more precise auditory processing.

similar for F3 variation ($Z = 0.133$, $p = .935$) and duration variation ($Z = 3.979$, $p = .137$), while the group difference reached statistical significance ($Z = 7.746$, $p = .021$). According to the multiple comparison analyses (with an alpha set to $p < .012$; Bonferroni corrected), native speakers of English demonstrated marginally less precise sensitivity to F2 compared to Japanese EFL speakers ($Z = 0.4.718$, $p = .030$), and experienced Japanese learners ($Z = 4.509$, $p = .034$). The results of 95% CI analyses hinted at a possibility that Japanese EFL learners' F2 processing scores were higher (less precise) compared to those of experienced Japanese learners, although there was a small overlap between the two groups (60.6–75.7 vs. 75.1–100.7; see Table 3).

To check the relationship between the three auditory processing factors (discrimination of F3, F2, and duration) and their associations with the relevant experience factors, a set of Pearson correlation analyses were performed. Experienced speakers' current L2 use was averaged across three different contexts (home, work, and social). The alpha level was set to 0.008 (Bonferroni corrected). According to the results summarized in Table 4, participants' auditory processing scores (F3, F2, and duration) did not show statistically significant associations with any of the biographical factors.

In sum, the results suggest (a) that the auditory processing measures tap into relatively independent constructs of participants' sensitivity to F3, F2, and duration variation specific to the English [r] and [l] contrast; (b) that F3 and duration discrimination abilities may be unrelated to the influence of biographical factors; and (c) that L2 learners with some degree of immersion experience may have greater F2 discrimination abilities.

4.3. Relationships between auditory processing, biographical backgrounds, and L2 speech proficiency

To examine the relative *weights* of auditory and biographical factors in L2 speech acquisition, a set of linear mixed-effects regression analyses were performed using the lmer functions from the lme package (Version 1.1–21; Bates, Maechler, Bolker, & Walker, 2015) in the R environment. To provide an overall picture of the link between auditory processing, experience, and L2 speech proficiency across the entire dataset ($N = 100$ Japanese speakers), Model 1 was constructed with participants' English [r]-[l] proficiency scores as dependent variables. In terms of predictors, the model featured one context factor (Perception vs. Production), one group factor (EFL vs. Experienced), and the three auditory processing factors (F3, F2, and duration discrimination). All numerical values were standardized, and dummy variables (Context, Group) were treatment-coded. Participants were entered as random effects. Effect sizes were interpreted in accordance with Cohen's (1988) benchmark, i.e., $R^2 = 0.02$ for small, 0.13 for medium, 0.26 for large.

It was important to first analyze the omnibus model including a total of 100 Japanese speakers (Model 1) because the findings would directly reveal the extent to which the Group variable (the presence of immersion experience) explained various levels of Japanese speakers' English [r] and [l] proficiency. As explained earlier, we predicted that L2 speech acquisition would be mainly predicted by the experience variables (Group: EFL vs. Experienced), and secondarily by the auditory processing variables (F3, F2, and duration discrimination). After Model 1 confirmed the presence of significant main effects of Group, we then conducted the follow-up analyses (Models 2, 3, and 4) wherein we further delved into the interaction effects of Group, i.e., whether the relationship between experience, auditory processing, and L2 speech acquisition was differed across two different contexts, EFL vs. Experienced.

As summarized in Table 5, Model 1 explained 70.6% of the variance in participants' English [r]-[l] proficiency. The effect size of the model could be considered substantially large. In light of the standardized β values, the results suggest that the rate of success was primarily predicted by the presence of immersion experience ($\beta = 1.546$) and secondarily by participants' F3 and F2 (but not duration) discrimination abilities ($\beta = -0.140, -0.126$). To check whether the mediating effects of auditory processing differed across Contexts (perception vs. production) and Group (EFL vs. Experienced), Model 2 was constructed by including both main and interaction effects of Group, Contexts, F3, and F2 factors. The model explained substantially large amounts of variances (i.e., 71.1%). Whereas the main effects remained significant as to Group ($\beta = 1.516$, $p < .001$) and F3 ($\beta = -0.238$, $p = .005$), the interaction effects of Group and F2 were found to be significant ($\beta = -0.335$, $p = .009$). With respect to Contexts, none of the main and interaction effects reached statistical significance ($p > .05$). The results suggest (a) that F3 discrimination serves as a significant predictor across all the context and group conditions; (b) that the strength of the link between F2 discrimination and acquisition could be significantly affected by participants' immersion status (EFL vs. Experienced); and (c) the complex relationships between F3, F2, and EFL and Experienced speakers' L2 proficiency remains comparable in both perception and production dimensions.

To further disentangle the significant Group × F2 interaction effects, we set out to investigate the dimension-specific relationship between auditory processing, experience, and L2 speech acquisition for each group (EFL vs. Experienced). Taking into account more detailed biographical information, two separate models were constructed (Models 3 and 4; see Table 6).

As for Model 3 ($n = 50$ EFL speakers), the model included participants' perception and production scores (English [r] and [l]) as

Table 4
Biographical correlates of F3, F2, and duration discrimination abilities.

	Age (EFL, Experienced)		Length of EFL (EFL, Experienced)		Pronunciation training (EFL)		Current EFL training (EFL)		Length of residence (Experienced)		Age of arrival (Experienced)		Current L2 use (Experienced)	
	r	*p*	*r*	*p*	*r*	*p*	*r*	*p*	*r*	*p*	*r*	*p*	*r*	*p*
F3 sensitivity (0–200 points)	0.074	0.446	0.109	0.280	−0.207	0.149	0.069	0.632	−0.044	0.764	0.258	0.071	0.033	0.821
F2 sensitivity (0–200 points)	−0.183	0.063	−0.154	0.125	−0.164	0.255	0.120	0.408	−0.075	0.045	0.124	0.389	−0.118	0.416
Duration sensitivity (0–200 points)	−0.020	0.884	−0.074	0.465	0.024	0.870	−0.110	0.448	0.211	0.141	0.100	0.490	−0.065	0.656

Note. Smaller values in auditory processing measures (0–200 points) indicate more precise auditory processing; An alpha set to $p < .008$.

Table 5
Summary of the mixed-effects model explaining the perceptual and biographical correlates of L2 speech proficiency (English [r] and [l]).

	Fixed effects	Estimate (β)	*SE*	*t*	*p*
Model 1	Intercept	−0.748	0.071	−10.438	< 0.001*
	Context (Perception vs. Production)	0.024	0.078	0.315	0.753
	Group (EFL vs. Experienced)	1.546	0.081	19.049	< 0.001*
	F3 sensitivity	−0.140	0.039	−3.534	0.001*
	F2 sensitivity	−0.126	0.041	−3.075	0.002*
	Duration sensitivity	−0.036	0.039	−0.936	0.548
	Random effects	Variance	*SD*		
	Participant	0.006	0.081		
	Conditional R^2	Marginal R^2			
	0.870	0.706			
Model 2	Intercept	−0.801	0.073	−10.864	< 0.001*
	Context (Perception vs. Production)	−0.013	0.057	−0.241	0.810
	Group (EFL vs. Experienced)	1.516	0.098	15.379	< 0.001*
	F3 sensitivity	−0.238	0.084	−2.815	0.005*
	F2 sensitivity	−0.021	0.068	−0.316	0.752
	Context × F3 sensitivity	0.148	0.088	1.684	0.095
	Context × F2 sensitivity	−0.060	0.067	−0.892	0.374
	Group × F3 sensitivity	0.075	0.111	0.676	0.500
	Group × F2 sensitivity	−0.335	0.127	−2.625	0.009*
	Context × Context × F3 sensitivity	−0.059	0.114	−0.521	0.603
	Context × Context × F2 sensitivity	0.104	0.127	0.823	0.412
	Random effects	*Variance*	*SD*		
	Participant	0.147	0.384		
	Conditional R^2	Marginal R^2			
	0.857	0.711			

Note.
* Indicates statistical significance ($p < .05$).

Table 6
Summary of the mixed-effects models explaining the perceptual and biographical correlates of L2 speech proficiency (English [r] and [l]) among EFL and experienced Japanese speakers.

	Fixed effects	*Estimate* (β)	*SE*	*t*	*p*
Model 3: EFL speakers	Intercept	−0.578	0.191	3.023	0.004*
	Context (Perception vs. Production)	−0.024	0.074	−0.325	0.747
	F3	−0.241	0.049	−2.851	0.007*
	F2	−0.021	0.040	−0.528	0.601
	Duration	−0.076	0.042	−1.805	0.079
	Total EFL education	0.001	0.017	0.073	0.942
	Pronunciation training	0.222	0.096	2.313	0.026*
	Current EFL training	0.026	0.022	1.180	0.245
	Random effects	Variance	*SD*		
	Participant	0.113	0.198		
	Conditional R^2	Marginal R^2			
	0.525	0.314			
Model 4: Experienced speakers	Intercept	−0.347	0.362	−0.959	0.340
	Context (Perception vs. Production)	0.072	0.128	0.568	0.572
	F3	−0.135	0.062	−2.160	0.033*
	F2	−0.334	0.090	−3.684	<0.001*
	Duration	−0.052	0.071	−0.727	0.469
	Length of residence	0.011	0.007	1.481	0.142
	Age of arrival	0.008	0.011	0.711	0.479
	Current L2 use	0.001	0.006	0.015	0.988
	Random effects	Variance	SD		
	Participant	0.058	0.242		
	Conditional R^2	Marginal R^2			
	0.747	0.266			

Note.
* Indicates statistical significance ($p < .05$).

dependent variables relative to six predictors: discrimination of F3, F2, and duration, total EFL education (total hours), the presence of pronunciation training to English [r] and [l] (yes = 1, no = 0), and the current EFL training (hours per week). As summarized in Table 6, the model accounted for the medium-to-large amount of the variances (31.4%) in EFL participants' English [r]-[l] acquisition. The outcomes primarily related to F3 discrimination ($\beta = -0.241$, $p = .007$) and secondarily to pronunciation training experience ($\beta = 0.222$, $p = .026$).

As for Model 4 ($n = 50$ experienced speakers), the model consisted of English [r] and [l] accuracy scores as dependent variables relative to six predictors: discrimination of F3, F2, and duration, length of residence, age of arrival, and current L2 use. Similar to Model 3, the model explained the medium-to-large amount of the variances (26.6%) in participants' English [r]-[l] acquisition. Yet, none of the experience-related factors reached statistical significance ($p > .05$). Interestingly, participants' L2 speech proficiency was significantly tied to not only F3 discrimination ($\beta = -0.135$, $p = .033$), but also F2 discrimination ($\beta = -0.334$, $p < .001$). Fig. 2 visually plots how Japanese participants' F2 and F3 discrimination abilities were associated with English [r]-[l] proficiency (averaged perception and production z scores), even after all the individual differences in the group condition (EFL vs. Experienced) were controlled for.

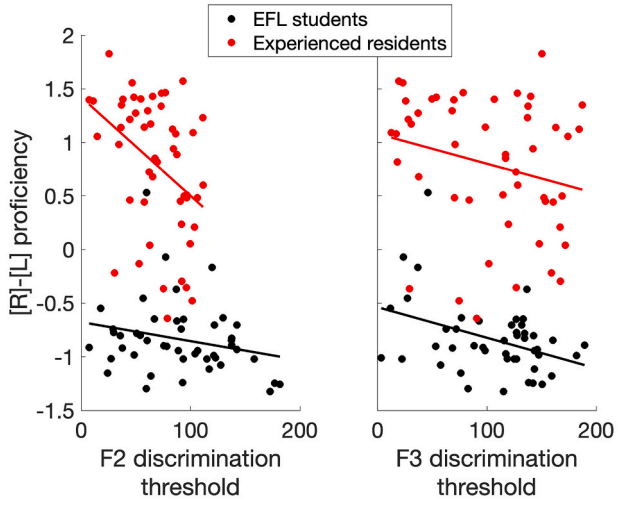

Fig. 2. (Left) Scatterplot displaying the relationship between F2 discrimination threshold (units along a stimulus continuum, lower numbers indicate better performance) and [r]-[l] perception and production ability (composite scores). Participants were inexperienced EFL students living in Japan (black) or experienced Japanese residents of an English-speaking country (red). (Right) Relationship between F3 discrimination threshold and [r]-[l] perception and production ability (composite scores). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

5. Discussion

Growing evidence suggests that individual variation in domain-general auditory perception skills may help predict acquisition of L2 speech proficiency (e.g., Kachlicka et al., 2019 for composite auditory

processing scores and general speech perception proficiency). To date, however, most work has focused on composite measures of auditory processing and speech proficiency, and so it remains an open question how individual differences in auditory processing along specific acoustic dimensions are linked to the acquisition of specific speech sound contrasts. We hypothesized that acquisition of speech sound contrasts relies upon the ability to precisely encode and focus attention on those dimensions that cue each contrast in various ways. To test this hypothesis, the current study focused on a well-researched, relatively difficult instance of L2 adult speech acquisition, that of English [r] and [l] for Japanese learners in classroom and immersion settings. In particular, we examined how specific aspects of Japanese speakers' English [r]-[l] performance (perception, production) were associated with individual differences in domain-general auditory sensitivity to key acoustic features (F3, F2, and duration variation) and experience-related factors (the quantity and quality of immersion and instruction experience).

First and foremost, we found that a substantial amount of variance in Japanese speakers' English [r] and [l] perception was accounted for by the presence of immersion experience (50–60%). That is, Japanese speakers with immersion experience perceived and produced the phonemes more accurately than those without such experience overseas with large effects ($d = 2.99$). The findings were in line with L2 speech literature showing that Japanese speakers can attain advanced English [r] and [l] proficiency after extensive immersion (Flege et al., 1995, 1996) and training experience (Bradlow et al., 1997). They also echoed the neurophysiological findings that L2 speech acquisition can be greatly promoted by the amount of immersion experience (e.g., Winkler et al., 1999) but that little improvement in L2 phonetic discrimination can be observed in classroom settings, especially when instruction is highly language-focused and delivered in the L1 rather than the L2 (e.g., Grimaldi et al., 2014). This in turn lends empirical support to the theoretical view that adult L2 learners can master relatively difficult new sounds (e.g., English [r] and [l]), provided that they access a great deal of naturalistic L2 input extensively, actively, and regularly (e.g., Best & Tyler, 2007 for Perceptual Assimilation Model; Flege & Bohn, 2021 for Speech Learning Model).

Taking into account participants' individual differences in auditory processing abilities, however, the findings here also add that some individual variation in adult L2 speech outcomes can be further explained by factors beyond the presence/absence of immersion experience. Overall, perceptual capacities to discriminate key acoustic dimensions of English [r] and [l] appeared to serve as an additional determinant of the attainment of advanced L2 speech proficiency (explaining 20–30% of the variance). Echoing the aptitude-acquisition view (Doughty, 2019), we argue that aptitude (operationalized in this study as more robust auditory processing abilities) matters for the acquisition of relatively difficult linguistic features. In the case of Japanese speakers' English [r]-[l] acquisition, some individuals may be better able to flexibly and precisely focus attention on the key acoustic information directly relevant to the English [r] and [l] contrast such as variations in F3 (1600–3400 Hz) and F2 (1200–1500 Hz). Such talented learners can better engage in input opportunities under various kinds of input conditions (classroom, immersion), leading to more advanced, robust, and nativelike L2 speech proficiency due to their ability to accurately detect, discriminate, and categorize certain acoustic information relevant to difficult speech sound contrasts. The findings here concur with previous studies showing that successful L2 speech learners likely have greater perceptual-cognitive abilities, such as phonological awareness (Venkatagiri & Levis, 2007), analysis (Hu et al., 2013 for phonemic coding), and memory (Silbert et al., 2015 for associative and working memory).

It is important to note that different types of auditory processing strategies—F3 and F2 discrimination abilities—were differentially associated with English [r] and [l] acquisition under different learning contexts (classroom vs. immersion). Whereas F3 sensitivity played a significant role in every stage of L2 speech learning (classroom *and* immersion), F2 discrimination was predictive of the extent to which participants could refine their English [r] - [l] accuracy in immersion settings. This suggests that compensatory mechanisms may be available only at later stages of adult L2 speech learning.

For Japanese learners, acquisition of the [r]-[l] contrast is extremely difficult as it requires listeners to not only utilize cues which they are accustomed to using (i.e., F2) but also to focus on new cues which they previously ignored (i.e., F3; Iverson et al., 2003). Even though some learners may be endowed with particularly precise F3 processing, and thus attend to F3 variation in English [r] and [l] contrasts, many Japanese learners likely encounter difficulty with the perception of the contrast due to the lack of such auditory skills. According to our findings, however, there is a possibility that other learners can still manage to distinguish English [r] and [l] with great accuracy by looking for other available cues, such as F2. Although F2 variation is a secondary (not necessarily the most reliable nor deciding) cue for the English [r] and [l] contrast, F2 distinguishes between English [r] (1200–1400 Hz) and English [l] (1600–2000 Hz) to some degree (Flege et al., 1995). As a remedy for their difficulty in attending to F3, those with more precise F2 discrimination could resort to this dimension so as to attain more robust, reliable, and advanced L2 speech perception skills in the long run.

That F2 sensitivity only predicted performance among immersion learners suggests that extensive exposure to speech input may by required for individuals to learn that this cue can be useful for perceiving the [r]/[l] contrast. In other words, learners may gradually become more cognizant that speech categorization can be achieved via a combination of various cue strategies (Lisker, 1986), although some strategies (F3 reliance) are more robust/nativelike than others (F2 reliance) (Iverson et al., 2003). The relevance of this cue may take time to learn because [r] and [l] overlap more in F2 than in F3, and so extensive experience may be necessary to separate out the underlying distributional peaks.

Our arguments concur with a structural redundancy model of speech perception (Jasmin, Dick, et al., 2020; Massaro & Cohen, 1993). Under this view, speech perception can be a multifaceted event in that the same information can be conveyed via a range of different cues (e.g., pitch, duration, and amplitude for word stress; voice onset time and pitch in following vowels for stops; Toscano & McMurray, 2010) and individuals are different in their preferences and biases towards which cues are weighted so as to grasp the underlying categories (Kidd et al., 2007). Even if individuals have difficulty with one cue due to biological constraints (e.g., amusics) or prior language backgrounds (e.g., tonal vs. non-tonal language speakers), they may show little difficulty in understanding spoken language as they can develop alternative strategies through much speech perception experience, wherein they can test, identify, and practice a range of cues that they can better detect (Liu et al., 2010). For example, whereas amusics rely on duration rather than pitch for detecting prosodic structures in English (Jasmin, Dick, et al., 2020), Mandarin speakers overly use pitch rather than duration cues (Jasmin, Sun, & Tierney, 2020). By resorting to other available cues (while downplaying perceptually difficult cues), neither of the populations demonstrate significant language impairments.

Extending the existing theories in adult L2 speech learning (Flege & Bohn, 2021 for Speech Learning Model), and applying the redundancy model of speech perception to this context, a range of implications can be made in regard to human speech learning throughout the lifespan. First, even adult learners can continue to master new sounds as a function of ample input using the system used for L1 acquisition (Flege & Bohn, 2021). Here, we add that one crucial component of the system is domain-general auditory perception, i.e. the ability to encode, represent, and integrate basic acoustic cues available in a new language. Thus, individual differences in precise auditory sensitivity play a critical role in determining the rate of success, especially in L2 speech acquisition, not only because their cue weighting processes can inevitably be affected by established L1 speech perception patterns, but also because they generally lack a sufficient amount of authentic input for robust auditory analyses (compared to L1 acquisition).

Whereas the precursor work has shown a broad relationship between individual differences in auditory processing and various areas of L2 proficiency (phonology vs. grammar; perception vs. production; Saito, Kachlicka, Sun, & Tierney, 2020), the current study has revealed the audition-proficiency link at dimension-specific levels. In general, those with sensitivity to the primary acoustic correlate of target phonetic structures (F3 for English [r]-[l] acquisition) take advantage of every input opportunity across any context (classroom vs. immersion). Instead of allocating efforts to the development of a new acoustic representation, which is well-known to be difficult even after years of immersion and focused training (Ingvalson, Holt, & McClelland, 2011, Ingvalson, McClelland, & Holt, 2011 for Japanese speakers' F3 discrimination during the English [r] and [l] perception), our findings have highlighted the complementary use of F2 information as a remedial strategy among more experienced Japanese users of English. Once learners engage in ample naturalistic input and become aware of the redundant nature of speech categorizations, they may gradually begin to attend to any other acoustic information, whether primary or secondary. In this regard, all dimension-specific auditory skills would provide additional benefits to naturalistic L2 speech learning as long as they can be used to help perceive sound contrasts (e.g., resorting to F3 *and/or* F2 for English [r] and [l]), even if they differ from nativelike strategy use (e.g., weighing F3 alone for English [r] and [l]; Iverson et al., 2003).

Finally, the specificity of the relationship between perception of F3 and F2 frequency and success in learning the [r]-[l] contrast in classroom or immersion contexts has crucial implications. First, this suggests that perception along different auditory dimensions forms dissociable skills—in particular, that perception of F2 and F3 can vary somewhat independently, despite the relative acoustic similarity of these two dimensions. This is in line with previous findings that performance on large auditory processing batteries is best modeled as a series of dissociable skills rather than an overarching "auditory quotient" (Karlin, 1942; Kidd et al., 2007; Stankov & Horn, 1980; Surprenant & Watson, 2001; Watson et al., 1982). Second, the specificity of the relationship between F3 and F2 discrimination and [l]-[r] perception argues against the view that relationships between auditory processing and language learning are primarily driven by a shared reliance on broad, higher-order cognitive skills such as attention, memory, or statistical learning (Jaffe-Dax et al., 2017; Snowling et al., 2018). The three auditory processing tests (F3, F2, and duration discrimination) can be assumed to be roughly matched in their overall cognitive demands, given that they used the exact same format. Moreover, the use of a roving comparison stimulus level for all three tests should minimize the role of statistical learning of a perceptual "anchor". As such, we can conclude that the complex relationship between auditory processing and [l]-[r] perception primarily reflects factors specific to auditory perception.

6. Limitations and future directions

With an eye towards future replication and extension studies, there are a range of methodological features that should be further elaborated, expanded, and refined. First, the study used psychoacoustic discrimination tasks, which are widely used measures of auditory processing in cognitive psychology and hearing research (Moore, 2012). Given that this is a behavioural approach, it inevitably entails the possibility that the task taps into not only participants' perception, but also a range of their cognitive abilities, such as phonological short-term memory, attentional control, and processing speed (Snowling et al., 2018). A complementary approach is to measure auditory encoding at preconscious levels using electrophysiological measures (e.g., Coffey, Herholz, Chepesiuk, Baillet, & Zatorre, 2016 for frequency following responses) and relate auditory neural encoding to L2 speech learning success (Kachlicka et al., 2019). It would be interesting in future studies to examine whether/how electrophysiological instruments can capture language learners' sensitivity to various single acoustic dimensions, and whether/how such implicit auditory processing may be associated with their English [r]-[l] proficiency.

Despite the relevance of auditory processing for post-pubertal L2 speech learning (similar to L1 speech learning), individuals vary widely in their pre-existing auditory skills. Another crucial question concerns why certain individuals are equipped with more robust auditory skills than others. It is noteworthy that although no significant correlation was found between participants' auditory processing and biographical backgrounds, at least within the current dataset ($N = 100$ Japanese EFL learners), there was an indication that experienced Japanese speakers appeared to exhibit more precise F2 discrimination than their EFL counterparts. The findings here accord with the view that greater amounts of bilingual experience could have larger effects on auditory skills. For example, background differences in auditory processing have been found when researchers compared two groups of speakers with substantially different backgrounds, such as tonal vs. non-tonal language users (Bidelman, Gandour, & Krishnan, 2011) and simultaneous vs. sequential bilinguals (Krizman, Slater, Skoe, Marian, & Kraus, 2015).

Third, although the current study pointed out that auditory processing could relate to the incidence of near-native L2 speech proficiency among experienced L2 learners, it remains a possibility that those participants with lower auditory processing abilities and less native L2 speech proficiency could continue to improve their proficiency, reaching near-native performance, after even more extensive exposure to L2 speech input. This possibility can be neither ruled out nor confirmed based on the current dataset. Assuming that one dimension of auditory processing abilities (i.e., F2 discrimination/representation) could be amenable to bilingual experience effects, this in turn brings to light the potential of hearing training to enhance L2 learners' auditory sensitivities. If auditory processing determines the degree of benefit which L2 learners can derive from received input (e.g., phonetic instruction in L2 speech learning), it is reasonable to predict that provision of auditory and phonetic training at the same time may boost the rate and ultimate attainment of L2 speech learning. There is much heated discussion on the possibility of focused training as a way to help remedy auditory impairments (for a meta-analysis, see Lawrence et al., 2018). For example, several hours of auditory training have been found to enhance pitch processing (e.g., Carcagno & Plack, 2011) and temporal processing (e.g., Strehlow et al., 2006) in children and adults. However, little is known about the extent to which enhanced auditory processing abilities will in turn lead to enhanced language learning.

The possibility of training F3, F2, and duration discrimination relevant to Japanese speakers' English [r] and [l] acquisition and the ideal method (i.e., use of natural versus synthesized English [r]-[l] tokens) is still open to debate (see Ingvalson, Holt, & McClelland, 2011 vs. Iverson, Hazan, & Bannister, 2005). In both studies (Ingvalson, Holt, & McClelland, 2011; Iverson et al., 2005), such training was delivered via synthesized speech tokens. Although the relevant cues were acoustically enhanced, it remains unclear the extent to which Japanese speakers can actually perceive them, given that they are used to ignoring them during L1 speech processing (the acoustic properties of Japanese [ɾ] substantially overlap with both English [r] and [l]; Hattori & Iverson, 2009).

In conjunction with some empirical evidence that the difficulty in Japanese speakers' English [r]-[l] acquisition lies in domain-general auditory rather than domain-specific speech (phonetic and phonological) levels (Miyawaki et al., 1975 for Japanese speakers' encoding of F3 differences in speech vs. nonspeech sounds), it would be intriguing to provide focused training using non-verbal rather than synthesized speech sounds as training stimuli. Such training is assumed to help enhance Japanese speakers' ability to focus on different types of single dimensions, such as F2 variation (which is partially used in the L1 Japanese system and may be amenable to experience effects), and F3 variation (which is not exploited in L1 Japanese system and resistant to experience effects). The effectiveness of auditory training can be further enhanced if the training stimuli comprised more complex non-verbal sounds simultaneously varying along other, task-irrelevant dimensions.

Fourth, it is worth mentioning that participants' duration

discrimination abilities did not significantly relate to any aspects of biographical backgrounds, nor to L2 English [r] and [l] proficiency. Given that English [r] is longer than English [l], there is empirical evidence that native speakers can rely on duration information to perceive the contrast when other acoustic information is masked (e.g., F3; Underbakke, Polka, Gottfried, & Strange, 1988). Surprisingly, participants with greater duration sensitivity did not necessarily demonstrate advanced L2 speech proficiency. One reason could be methodological. According to the results of the previous research report (Saito, Sun, & Tierney, 2020), although participants' test-reliability was high for composite scores ($r = 0.701$), it became relatively low when we conducted separate analyses for temporal processing tests ($r = 0.284$ for duration discrimination) vs. spectral processing tests ($r = 0.619$ for F2 discrimination). We may need to wait for future studies to conceptualize, elaborate, and refine more reliable instruments to capture the degree of one's temporal precision via using both behavioural and neurophysiological measures.

Fifth, the current study highlighted a total of six variables related to the quantity and quality of L2 input in EFL and naturalistic L2 speech learning (see Table 1). We intentionally minimized the number of predictors because including too many (similar) variables would result in multicollinearity problems and low power for statistical analyses. While considerably more experience information can be surveyed via the existing questionnaires (e.g., Freed et al., 2004 for Language Contact Profile), it is important to acknowledge that the method overly relies on participants' self-report and recollection. Flege and Wayland (2019) proposed the Experience Sampling Method as a new measure of the quantity and quality of L2 input. Under this methodological paradigm, L2 learners report their here-and-now L2 use by answering a simple question (e.g., what language are you using now?) via their smartphone everyday (2–3 times per day); this intensive data collection can take place for a certain period of time (e.g., 1–2 months). To track the influence of the experience on acquisition, their L2 speech proficiency can also be recorded on the same time interval. While useful, to our knowledge, such research has never been conducted due to obvious methodological difficulties. To avoid much attrition, participants need to be highly motivated to do the same task multiple times without any delay and fails. If the data is collected online, the precision of data remains open to discussion because we do not know the extent to which participants carefully monitored their L2 use and then accurately reported it as use as instructed (for a range of attention check measures during online data collection for L2 speech research, Nagle, 2021). In terms of interlocutor type, participants' responses could be still influenced by their own subjective judgements of who can be considered as "L1 speakers," "fluent L2 speakers," vs. "non-fluent L2 speakers." If participants undertake the same L2 speaking tasks multiple times, any change in their performance could be due to test-rest effects. As one interesting direction for future research of this kind, this topic can be scrutinized from quantitative *and* qualitative paradigms by combining a range of instruments, such as questionnaires, online learning log measures, and ethnographic interviews/descriptions of language use (cf. Ranta & Meckelborg, 2013).

Finally, the current study examined Japanese speakers' English [r]-[l] proficiency via the identification and spontaneous pronunciation of the difference between the two phonemes in minimal pair contexts. We must acknowledge that the topic needs to be examined via different tasks and analyses in order to take into account a range of affecting variables. It has been shown that Japanese speakers likely have more learning difficulty in English [l] than English [r] (e.g., Aoyama et al., 2004) because the former is acoustically too similar to be distinguished from the L1 counterpart (Japanese [ɾ]; Hattori & Iverson, 2009). Thus, Japanese speakers tend to initially assimilate English [l] to Japanese [ɾ], and then create a composite category (featuring both English [l] and Japanese [ɾ]), while they can create a new L2 category for English [r] (which is sufficiently different from Japanese [ɾ]; see Best & Tyler, 2007 for their perceptual assimilation account of L2 speech acquisition). To examine the extent to which Japanese speakers have established two new different categories for English [r] and [l], for example, Aoyama et al. (2004) adopted a test in which not only English [r] and [l] tokens but also realizations of other potentially confusable English consonants (e.g., English [w]) were presented together for categorization.

7. Conclusion

The current study investigated dimension-specific relationships between processing of individual auditory dimensions (F3, F2, duration) and success in learning specific L2 speech sound contrasts (English [r] and [l]). We found that while adult L2 learners generally benefit from immersion experience or phonetic instruction to a great degree, learners with robust processing of the crucial acoustic cues for a given speech sound contrast may be better able to demonstrate more advanced and nativelike performance. The findings support the view that dimension-specific auditory processing abilities drive language learning in a *complementary* fashion throughout the lifespan (Jasmin, Dick, et al., 2020; Jasmin, Sun, & Tierney, 2020). In the context of L2 speech acquisition, individuals vary in their abilities to encode single acoustic dimensions not only due to the influence of L1 phonetic systems, but also as a part of pre-existing traits. As a function of increased exposure to naturalistic spoken input, wherein speech contrasts can be distinguished via a combination of numerous cues, some can attain high-level L2 speech proficiency while compensating for any disadvantage in a primary cue by relying on another available one (i.e., a trade-off between F3 and F2 variation for English [r] and [l] acquisition).

CRediT authorship contribution statement

Kazuya Saito: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing - original draft, Writing - review & editing. **Magdalena Kachlicka:** Data curation, Formal analysis, Writing – review & editing. **Yui Suzukida:** Data curation, Formal analysis, Writing – review & editing. **Katya Petrova:** Data curation, Formal analysis, Writing – review & editing. **Bradford Lee:** Data curation, Investigation, Project administration, Writing - review & editing. **Adam Tierney:** Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Resources, Supervision, Visualization, Writing – review & editing.

Acknowledgement

The current project was funded by Leverhulme Trust (RPG-2019-039) and Economic and Social Research Council (ES/S013024/1).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.cognition.2022.105236.

References

Abrahamsson, N., & Hyltenstam, K. (2008). The robustness of aptitude effects in near-native second language acquisition. *Studies in Second Language Acquisition, 30*(4), 481–509. https://doi.org/10.1017/S027226310808073X

Ahissar, M., Protopapas, A., Reid, M., & Merzenich, M. M. (2000). Auditory processing parallels reading abilities in adults. *PNAS, 97*(12), 6832–6837. https://doi.org/10.1073/pnas.97.12.6832

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 behavioural experiment builder. *Behaviour Research Methods, 52*, 388–407. https://doi.org/10.3758/s13428-019-01237-x

Aoyama, K., Flege, J. E., Guion, S. G., Akahane-Yamada, R., & Yamada, T. (2004). Perceived phonetic dissimilarity and L2 speech learning: The case of Japanese/r/and English/l/and/r. *Journal of Phonetics, 32*(2), 233–250. https://doi.org/10.1016/S0095-4470(03)00036-6

Baker, W., Trofimovich, P., Flege, J. E., Mack, M., & Halter, R. (2008). Child—Adult differences in second-language phonological learning: The role of cross-language similarity. *Language and Speech, 51*(4), 317–342. https://doi.org/10.1177/0023830908099068

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

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/10.1177/0023830909349151

Best, C. T., & Tyler, M. D. (2007). Nonnative and second-language speech perception: Commonalities and complementarities. In O. S. Bohn, & M. J. Munro (Eds.), *Language experience in second language speech learning: In honor of James Emil Flege*. John Benjamins Publishing Company.

Bidelman, G. M., Gandour, J. T., & Krishnan, A. (2011). Musicians and tone-language speakers share enhanced brainstem encoding but not perceptual benefits for musical pitch. *Brain and Cognition, 77*(1), 1–10. https://doi.org/10.1016/j.bandc.2011.07.006

Boersma, P., & Weenink, D. (2010). *Praat (version 5.1. 35). Phonetic sciences*. University of Amsterdam.

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

Bradlow, A. R. (2008). Training non-native language sound patterns. Lesson from training Japanese adults on the english /r/−/l/ contrast. (pp. 287–309). In J. G. Hansen-Edwards, & M. L. Zampini (Eds.), *Phonology and second language acquisition*. John Benjamins Publishing Company.

Bradlow, A. R., Pisoni, D. B., Akahane-Yamada, R., & Tohkura, Y. (1997). Training Japanese listeners to identify English /r/ and /l/: IV. Some effects of perceptual learning on speech production. *The Journal of the Acoustical Society of America, 101*(4), 2299–2310. https://doi.org/10.1121/1.418276

Carcagno, S., & Plack, C. J. (2011). Pitch discrimination learning: Specificity for pitch and harmonic resolvability, and electrophysiological correlates. *Journal of the Association for Research in Otolaryngology: JARO, 12*(4), 503–517. https://doi.org/10.1007/s10162-011-0266-3

Casini, L., Pech-Georgel, C., & Ziegler, J. C. (2018). It's about time: Revisiting temporal processing deficits in dyslexia. *Developmental Science, 21*(2), 1–14. https://doi.org/10.1111/desc.12530

Chandrasekaran, B., Sampath, P. D., & Wong, P. C. M. (2010). Individual variability in cue-weighting and lexical tone learning. *The Journal of the Acoustical Society of America, 128*(1), 456–465. https://doi.org/10.1121/1.3445785

Cobb, T. (2000). The compleat lexical tutor. https://www.lextutor.ca/.

Coffey, E. B. J., Herholz, S. C., Chepesiuk, A. M. P., Baillet, S., & Zatorre, R. J. (2016). Cortical contributions to the auditory frequency-following response revealed by MEG. *Nature Communications, 7*, 11070. https://doi.org/10.1038/ncomms11070

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*. New York, NY: Routledge Academic.

Delattre, P., & Freeman, D. (1968). A dialect study of American r's by x–ray motion picture. *Linguistics, 44*, 29–68.

Derwing, T. M., & Munro, M. J. (2013). The development of L2 oral language skills in two L1 groups: A 7-year study. *Language Learning, 63*(2), 163–185. https://doi.org/10.1111/lang.12000

Doughty, C. J. (2019). Cognitive language aptitude. *Language Learning: A Journal of Research in Language Studies, 69*(1), 101–126. https://doi.org/10.1111/lang.12322

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

Espy-Wilson, C. Y. (1992). Acoustic measures for linguistic features distinguishing the semivowels /w j r l/ in American English. *The Journal of the Acoustical Society of America, 92*(2), 736–757. https://doi.org/10.1121/1.403998

Espy-Wilson, C. Y., Boyce, S. E., Jackson, M., Narayanan, S., & Alwan, A. (2000). Acoustic modeling of American English /ɹ/. *Journal of the Acoustical Society of America, 108*, 343–356.

Flege, J., & Bohn, O.-S. (2021). The revised speech learning model. In R. Wayland (Ed.), *Second language speech learning: Theoretical and empirical progress*. Cambridge: Cambridge University Press.

Flege, J. E., & Liu, S. (2001). The effect 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 /r/ and /l/ accurately. *Language and Speech, 38*(1), 25–55. https://doi.org/10.1177/2F002383099503800102

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

Flege, J. E., & Wayland, R. (2019). The role of input in native Spanish Late learners' production and perception of English phonetic segments. *Journal of Second Language Studies, 2*(1), 1–44. https://doi.org/10.1075/jsls.00004.fle

Freed, B. F., Dewey, D. P., Segalowitz, N., & Halter, R. (2004). The language contact profile. *Studies in Second Language Acquisition, 26*(2), 349–356. https://doi.org/10.1017/S027226310426209X

Goswami, U. (2015). Sensory theories of developmental dyslexia: Three challenges for research. *Nature Reviews Neuroscience, 16*(1), 43–54. https://doi.org/10.1038/nrn3836

Goswami, U., Thomson, J., Richardson, U., Stainthorp, R., Hughes, D., Rosen, S., & Scott, S. K. (2002). Amplitude envelope onsets and developmental dyslexia: A new hypothesis. *PNAS, 99*(16), 10911–10916. https://doi.org/10.1073/pnas.122368599

Grimaldi, M., Sisinni, B., Gili Fivela, B., Invitto, S., Resta, D., Alku, P., & Brattico, E. (2014). Assimilation of L2 vowels to L1 phonemes governs L2 learning in adulthood: A behavioral and ERP study. *Frontiers in Human Neuroscience, 8*, 279. https://doi.org/10.3389/fnhum.2014.00279

Guion, S. G., Flege, J. E., Akahane-Yamada, R., & Pruitt, J. C. (2000). An investigation of current models of second language speech perception: The case of Japanese adults' perception of English consonants. *The Journal of the Acoustical Society of America, 107*(5), 2711–2724. https://doi.org/10.1121/1.428657

Halliday, L. F., & Bishop, D. V. M. (2005). Frequency discrimination and literacy skills in children with mild to moderate sensorineural hearing loss. *Journal of Speech, Language, and Hearing Research, 48*(5), 1187–1203. https://doi.org/10.1044/1092-4388(2005/083)

Halliday, L. F., & Bishop, D. V. M. (2006). Is poor frequency modulation detection linked to literacy problems? A comparison of specific reading disability and mild to moderate sensorineural hearing loss. *Brain and Language, 97*(2), 200–2213. https://doi.org/10.1016/j.bandl.2005.10.007

Halliday, L. F., Tuomainen, O., & Rosen, S. (2017). Auditory processing deficits are sometimes necessary and sometimes sufficient for language difficulties in children: Evidence from mild to moderate sensorineural hearing loss. *Cognition, 166*, 139–151. https://doi.org/10.1016/j.cognition.2017.04.014

Hattori, K., & Iverson, P. (2009). English/r/−/l/category assimilation by Japanese adults: Individual differences and the link to identification accuracy. *The Journal of the Acoustical Society of America, 125*(1), 469–479. https://doi.org/10.1121/1.3021295

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

Ingvalson, E. M., Holt, L. L., & McClelland, J. L. (2012). Can native Japanese listeners learn to differentiate/r–l/on the basis of F3 onset frequency? *Bilingualism: Language and Cognition, 15*(2), 255–274.

Ingvalson, E. M., McClelland, J. L., & Holt, L. L. (2011). Predicting native English-like performance by native Japanese speakers. *Journal of Phonetics, 39*(4), 571–584. https://doi.org/10.1016/j.wocn.2011.03.003

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

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

Jaffe-Dax, S., Frankel, O., & Ahissar, M. (2017). *Dyslexics' faster decay of implicit memory for sounds and words is manifested in their shorter neural adaptation., eLife, 6*, Article e20557. https://doi.org/10.7554/eLife.20557

Jasmin, K., Dick, F., Holt, L., & Tierney, A. T. (2020). Tailored perception: Individuals' speech and music perception strategies fit their perceptual abilities. *Journal of Experimental Psychology: General, 149*(5), 914–934. https://doi.org/10.1037/xge0000688

Jasmin, K., Sun, H., & Tierney, A. T. (2020). Effects of language experience on domain-general perceptual strategies. *Cognition, 206*, Article 104481. https://doi.org/10.1016/j.cognition.2020.104481

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), 1–9. https://doi.org/10.1111/desc.12836

Karlin, J. E. (1942). A factorial study of auditory function. *Psychometrika, 7*(4), 251–279.

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

Krizman, J., Slater, J., Skoe, E., Marian, V., & Kraus, N. (2015). Neural processing of speech in children is influenced by extent of bilingual experience. *Neuroscience Letters, 585*, 48–53. https://doi.org/10.1016/j.neulet.2014.11.011

Lamb, S. J., & Gregory, A. H. (1993). The relationship between music and reading in beginning readers. *Educational Psychology, 13*(1), 19–27. https://doi.org/10.1080/0144341930130103

Lawrence, B. J., Jayakody, D. M. P., Henshaw, H., Ferguson, M. A., Eikelboom, R. H., Loftus, A. M., & Friedland, P. L. (2018). Auditory and cognitive training for cognition in adults with hearing loss: A systematic review and Meta-analysis. *Trends in Hearing, 22*, 1–20. https://doi.org/10.1177/2F2331216518792096

Leong, C. X. R., Price, J. M., Pitchford, N. J., & van Heuven, W. J. (2018). High variability phonetic training in adaptive adverse conditions is rapid, effective, and sustained. *PLoS One, 13*(10), 1–24.e0204888. https://doi.org/10.1371/journal.pone.0204888

Lisker, L. (1986). "Voicing" in English: A catalogue of acoustic features signalling /b/ versus /p/ in trochees. *Language and Speech, 29*(1), 3–11. https://doi.org/10.1177/2F002383098602900102

Liu, F., Patel, A. D., Fourcin, A., & Stewart, L. (2010). Intonation processing in congenital amusia: Discrimination, identification and imitation. *Brain: A. Journal of Neurology, 133*(6), 1682–1693. https://doi.org/10.1093/brain/awq089

Lotto, A. J., Sato, M., & Diehl, R. (2004). Mapping the task for the second language learner: The case of Japanese acquisition of /r/ and /l/. In J. Slitka, S. Manuel, & M. Matthies (Eds.), *From sound to sense: 50+ years of discoveries in speech communication* (pp. C181–C186). Cambridge, MA: MIT Press.

Major, R. (2001). *Foreign accent: The ontogeny and phylogeny of second language phonology.* Mahwah, NJ: Lawrence Erlbaum Associates.

Major, R. C. (2008). Transfer in second language phonology. In J. G. Hansen Edwards, & M. L. Zampini (Eds.), *Phonology and second language acquisition*. Philadelphia: John Benjamins.

Massaro, D. W., & Cohen, M. M. (1993). Perceiving asynchronous bimodal speech in consonant-vowel and vowel syllables. *Speech Communication, 13*(1–2), 127–134. https://doi.org/10.1016/0167-6393(93)90064-R

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., & Bishop, D. V. M. (2004). Frequency discrimination deficits in people with specific language impairment. *Journal of Speech, Language, and Hearing Research, 47*(3), 527–541. https://doi.org/10.1044/1092-4388(2004/041)

McArthur, G. M., & Bishop, D. V. M. (2005). Speech and non-speech processing in people with specific language impairment: A behavioural and electrophysiological study. *Brain and Language, 94*(3), 260–273. https://doi.org/10.1016/j.bandl.2005.01.002

McLean, S., Hogg, N., & Kramer, B. (2014). Estimations of Japanese university learners' English vocabulary sizes using the Vocabulary Size Test. *Vocabulary Learning and Instruction, 3*(2), 47–55.

Miyawaki, K., Jenkins, J. J., Strange, W., Liberman, A. M., Verbrugge, R., & Fujimura, O. (1975). An effect of linguistic experience: The discrimination of [r] and [l] by native speakers of Japanese and English. *Perception & Psychophysics, 18*(5), 331–340. https://doi.org/10.3758/BF03211209

Montgomery, C. R., Morris, R. D., Sevcik, R. A., & Clarkson, M. G. (2005). Auditory backward masking deficits in children with reading disabilities. *Brain and Language, 95*(3), 450–456. https://doi.org/10.1016/j.bandl.2005.03.006

Moore, B. C. J. (2012). *An introduction to the psychology of hearing.* Emerald Group Publishing.

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

Muñoz, C. (2014). Contrasting effects of starting age and input on the oral performance of foreign language learners. *Applied Linguistics, 35*(4), 463–482. https://doi.org/10.1093/applin/amu024

Nagle, C. L. (2021). Revisiting perception–production relationships: Exploring a new approach to investigate perception as a time-varying predictor. *Language Learning, 71*(1), 243–279. https://doi.org/10.1111/lang.12431

Nagle, C. L., & Baese-Berk, M. M. (2021). Advancing the state of the art in l2 speech perception-production research: Revisiting theoretical assumptions and methodological practices. *Studies in Second Language Acquisition, 1–26*. https://doi.org/10.1017/S0272263121000371

Pfenninger, S. E., & Singleton, D. (2019). Starting age overshadowed: The primacy of differential environmental and family support effects on second language attainment in an instructional context. *Language Learning, 69*(S1), 207–234. https://doi.org/10.1111/lang.12318

Piske, T., MacKay, I. R., & Flege, J. E. (2001). Factors affecting degree of foreign accent in an L2: A review. *Journal of Phonetics, 29*(2), 191–215. https://doi.org/10.1006/jpho.2001.0134

Polka, L., & Strange, W. (1985). Perceptual equivalence of acoustic cues that differentiate/r/and/l. *The Journal of the Acoustical Society of America, 78*(4), 1187–1197. https://doi.org/10.1121/1.392887

Ranta, L., & Meckelborg, A. (2013). How much exposure to English do international graduate students really get? Measuring language use in a naturalistic setting. *The Canadian Modern Language Review, 69*(1), 1–33. https://doi.org/10.3138/cmlr.987

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

Rosen, S., & Manganari, E. (2001). Is there a relationship between speech and nonspeech auditory processing in children with dyslexia? *Journal of Speech, Language, and Hearing Research, 44*(4), 720–736. https://doi.org/10.1044/1092-4388(2001/057)

Saito, K. (2013a). The acquisitional value of recasts in instructed second language speech learning: Teaching the perception and production of English/ɹ/to adult Japanese learners. *Language Learning, 63*(3), 499–529. https://doi.org/10.1111/lang.12015

Saito, K. (2013b). Age effects on late bilingualism: The production development of/ɹ/by high-proficiency Japanese learners of English. *Journal of Memory and Language, 69*(4), 546–562. https://doi.org/10.1016/j.jml.2013.07.003

Saito, K., & Brajot, F. X. (2013). Scrutinizing the role of length of residence and age of acquisition in the interlanguage pronunciation development of English/ɹ/by late Japanese bilinguals. *Bilingualism: Language and Cognition, 16*(4), 847–863.

Saito, K., & Hanzawa, K. (2016). Developing second language oral ability in foreign language classrooms: The role of the length and focus of instruction and individual differences. *Applied PsychoLinguistics, 37*(4), 813–840. https://doi.org/10.1017/S0142716415000259

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*, Article 104168. https://doi.org/10.1016/j.jml.2020.104168

Saito, K., & Munro, M. J. (2014). The early phase of/ɹ/production development in adult Japanese learners of English. *Language and Speech, 57*(4), 451–469. https://doi.org/10.1177/0023830913513206

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., Kachlicka, M., Alayo, J. R. C., Nakata, T., & Tierney, A. T. (2020). Domain-general auditory processing explains multiple dimensions of L2 acquisition in adulthood. *Studies in Second Language Acquisition, 1–30*. https://doi.org/10.1017/S0272263120000467

Saito, K., Sun, H., & Tierney, A. T. (2020). Brief report: Test-retest reliability of explicit auditory processing measures. *BioRxiv.*. https://doi.org/10.1101/2020.06.12.149484

Silbert, N. H., Smith, B. K., Jackson, S. R., Campbell, S. G., Hughes, M. M., & Tare, M. (2015). Non-native phonemic discrimination, phonological short term memory, and word learning. *Journal of Phonetics, 50*, 99–119. https://doi.org/10.1016/j.wocn.2015.03.001

Smith, J. O. (2007). *Introduction to digital filters: 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

Stankov, L., & Horn, J. L. (1980). Human abilities revealed through auditory tests. *Journal of Educational Psychology, 72*(1), 21–44. https://psycnet.apa.org/doi/10.1037/0022-0663.72.1.21.

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*, 19–29. https://doi.org/10.1007/s00787-006-0500-4

Surprenant, A. M., & Watson, C. S. (2001). Individual differences in the processing of speech and nonspeech sounds by normal-hearing listeners. *The Journal of the Acoustical Society of America, 110*(4), 2085–2095. https://doi.org/10.1121/1.1404973

Talcott, J. B., Witton, C., McLearn, M. F., Hansen, P. C., Rees, A., Green, G. G. R., & Stein, J. F. (2000). Dynamic sensory sensitivity and children's word decoding skills. *PNAS, 97*(6), 2952–2957. https://doi.org/10.1073/pnas.040546597

Tallal, P. (2004). Improving language and literacy is a matter of time. *Nature Reviews Neuroscience, 5*, 721–728. https://doi.org/10.1038/nrn1499

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*, Article 105196. https://doi.org/10.1016/j.jecp.2021.105196

Tierney, A. T., & Kraus, N. (2014). Auditory-motor entrainment and phonological skills: Precise auditory timing hypothesis (PATH). *Frontiers in Human Neuroscience, 8*(849). https://doi.org/10.3389/fnhum.2014.00949

Toscano, J. C., & McMurray, B. (2010). Cue integration with categories: Weighting acoustic cues in speech using unsupervised learning and distributional statistics. *Cognitive Science, 34*, 434–464.

Trofimovich, P., & Baker, W. (2006). Learning second language suprasegmentals: Effect of L2 experience on prosody and fluency characteristics of L2 speech. *Studies in Second Language Acquisition, 28*(1), 1–30. https://doi.org/10.1017/S0272263106060013

Underbakke, M., Polka, L., Gottfried, T. L., & Strange, W. (1988). Trading relations in the perception of /ɹ/ and /l/ by Japanese learners of English. *Journal of the Acoustical Society of America, 84*, 90–100.

Venkatagiri, H. S., & Levis, J. M. (2007). Phonological awareness and speech comprehensibility: An exploratory study. *Language Awareness, 16*(4), 263–277. https://doi.org/10.2167/la417.0

Watson, C. S., Jensen, J. K., Foyle, D. C., Leek, M. R., & Goldgar, D. E. (1982). Performance of 146 normal adult listeners on a battery of auditory discrimination tests. *The Journal of the Acoustical Society of America, 71*(S73). https://doi.org/10.1121/1.2019533

Winkler, I., Kujala, T., Tiitinen, H., Sivonen, P., Alku, P., Lehtokoski, A., Czigler, I., Csépe, V., Ilmoniemi, R. J., & Näätänen, R. (1999). Brain responses reveal the learning of foreign language phonemes. *Psychophysiology, 36*(5), 638–642. https://doi.org/10.1111/1469-8986.3650638

Wong, P. C. M., & Perrachione, T. K. (2007). Learning pitch patterns in lexical identification by native English-speaking adults. *Applied PsychoLinguistics, 28*(4), 565–585. https://doi.org/10.1017/S0142716407070312

Wright, B. A., Lombardino, L. J., King, W. M., Puranik, C. S., Leonard, C. M., & Merzenich, M. M. (1997). Deficits in auditory temporal and spectral resolution in language-impaired children. *Nature, 387*, 176–178. https://doi.org/10.1038/387176a0

Wright, C. M., & Conlon, E. G. (2009). Auditory and visual processing in children with dyslexia. *Developmental Neuropsychology, 34*(3), 330–355. https://doi.org/10.1080/87565640902801882