

Language-dependent cue weighting: An investigation of perception modes in L2 learning

Second Language Research

1–25

© The Author(s) 2019

Article reuse guidelines:

sagepub.com/journals-permissions

DOI: 10.1177/0267658319832645

journals.sagepub.com/home/slr

**Kakeru Yazawa** 

Waseda University, Japan

James Whang

Western Sydney University; Australian Research Council Centre of Excellence for the Dynamics of Language, Australia

Mariko Kondo

Waseda University, Japan

Paola Escudero

Western Sydney University, Australia; Australian Research Council, Centre of Excellence for the Dynamics of Language, Australia

Abstract

This study examines relative weighting of two acoustic cues, vowel duration and spectra, in the perception of high front vowels by Japanese learners of English. Studies found that Japanese speakers rely heavily on duration to distinguish /i:/ and /ɪ/ in American English (AmE) as influenced by phonemic length in Japanese /ii/ and /i/, while spectral cues are more important for native AmE speakers. However, little is known as to whether and how this non-native perceptual weighting can change as a result of L2 learning. By employing computational and experimental methods, the present study shows that Japanese learners of English exhibit different cue weighting depending on which language they think they hear. The experiment shows that listeners use more spectral cues and less durational cues when they think they are listening to 'English' stimuli as opposed to 'Japanese' stimuli, despite the stimuli being identical. This result is generally in line with our computer simulation, which predicts distinct developmental paths in first language (L1) and second language (L2) perception. The Second Language Linguistic Perception (L2LP) model, which incorporates the language mode hypothesis, provides a comprehensive explanation for the current findings.

Corresponding author:

Kakeru Yazawa, Waseda University, 1-6-1 Nishiwaseda, Shinjuku-ku, Tokyo, 169-8050, Japan.

Email: k-yazawa@asagi.waseda.jp

Keywords

American English, cue weighting, high front vowels, Japanese, L2LP, language mode

1 Introduction

In adulthood, perception of speech sounds in a second language (L2) is mediated by the acquired phonology of the first language (L1). Research on second language acquisition (SLA) suggests that difficulties in L2 perception arise from a lack of straightforward, one-to-one correspondences between L1 and L2 sound categories (Best, 1995; Escudero, 2005; Flege, 1995). A well-known example of this is Japanese listeners' inability to discriminate English /l/ and /r/ ([ɹ]), which is caused by the two L2 sound categories mapping to a single L1 sound category, namely Japanese /r/ ([ɾ]) (Hattori and Iverson, 2009; Iverson et al., 2003; MacKain et al., 1981). Since difficulty in perceiving certain sound contrasts leads to difficulties in differentiating words contrasting in those sounds, it is important to consider how and to what extent existing L1 sound categories interact with L2 speech perception. The present study investigates Japanese speakers' use of acoustic cues for perceiving high front vowels in L1 Japanese and L2 American English (AmE)¹ as a function of language-specific perception modes, in relation to predictions made by major L2 perception models.

In AmE, the tense vowel /i:/ is more peripheral in the vowel space and also longer in duration than the lax vowel /ɪ/ (Hillenbrand et al., 1995). Despite the systematic difference in duration, native AmE speakers are known to distinguish this contrast by vowel spectra, with their perception 'hardly affected at all by duration' (Hillenbrand et al., 2000). In other words, vowel length is a phonologically redundant feature for this contrast. However, studies have found that native Japanese speakers rely heavily on duration to distinguish /i:/ and /ɪ/ in AmE, presumably because Japanese has long /ii/ and short /i/ contrasting in the temporal rather than the spectral domain. That is, acoustically long /i:/ in AmE seems to map to phonologically long /ii/ in Japanese, and acoustically short /ɪ/ in AmE to phonologically short /i/ in Japanese (Strange et al., 1998). This assimilation pattern is also reflected in Japanese loanwords from English, e.g. /riibu/ *leave* and /ribu/ *live*. Yet, little is known as to whether Japanese listeners learn to use spectral cues as they become proficient in L2 English. Morrison (2002) conducted a longitudinal study in which native Japanese and Spanish speakers were tested on Canadian English /i:/ and /ɪ/ (which share very similar acoustic properties with AmE /i:/ and /ɪ/), one month and six months after their arrival in Canada. The Japanese listeners showed primarily duration-based perception at both initial and final tests, suggesting that no developmental change occurred within five months. Contrarily, Fox and Maeda (1999) found that short-term perception training with immediate feedback can improve Japanese listeners' perception of /i:/ and /ɪ/ tokens that were manipulated to have roughly the same duration and therefore contrasting only in vowel spectra. The listeners' performance on natural tokens without robust durational cues also improved after training. The result suggests that Japanese listeners can notice and make use of spectral cues if they are given explicit feedback as to whether their categorization is correct. However, it remains unclear whether they will ultimately acquire native-like, primarily spectral perception as a result of naturalistic L2 learning.

Various perception models have been proposed to account for the relationships between L1 and L2 perception and to predict L2 learning outcomes. Among these, the Speech Learning Model (SLM) (Flege, 1995), the Perceptual Assimilation Model (PAM) and its extension to L2 learning (PAM-L2) (Best, 1995; Best and Tyler, 2007) have been extensively used in the SLA literature. SLM would explain Japanese listeners' persistent use of duration in perceiving AmE /i:/ and /ɪ/ as a result of equivalence classification, in which an L2 sound is perceived as equivalent to an existing L1 phonetic category (Japanese /ii/ and /i/, respectively). The L1 and L2 phonetic categories would be perceptually linked as diaphones in a common phonological space, of which properties will eventually resemble one another. Alternatively, a new phonetic category can be formed for an L2 sound if listeners discern at least some of the phonetic differences between L1 and L2 categories (e.g. Japanese /i/ and AmE /ɪ/), in which case Japanese listeners may establish a spectral distinction between AmE /i:/ and /ɪ/. PAM(-L2) makes somewhat similar predictions to SLM. According to the model, Japanese speakers' perception of AmE /i:/-/ɪ/ falls into a Two-Category assimilation pattern, in which each L2 sound is assimilated to a different L1 category (Japanese /ii/ and /i/) in a common L1-L2 space. Since learners would have little difficulty in discriminating minimally contrasting words for those sounds, no further perceptual learning is likely to occur for this assimilation pattern. This indicates that Japanese listeners are likely to maintain duration-based perception for AmE /i:/-/ɪ/. However, the model also proposes an alternative possibility that one of the L2 sounds is phonologically assimilated and yet perceived as phonetically deviant. For example, Japanese listeners may phonologically (functionally or lexically) equate AmE /ɪ/ with Japanese /i/, but the two sounds may be nonetheless easily dissimilated phonetically. In such case, a new category for the deviant L2 sound is reasonably likely to be formed, possibly leading to a spectral distinction between AmE /i:/ and /ɪ/. To summarize, SLM and PAM(-L2) predict that Japanese listeners may maintain duration-based perception for AmE /i:/-/ɪ/ as a result of cross-linguistic assimilation in a common space, but neither model rejects the possibility of new category formation leading to spectral perception.

A more recently proposed model, the Second Language Linguistic Perception (L2LP) model (Escudero, 2005; van Leussen and Escudero, 2015), differs from the above two in that it assumes separate L1 and L2 perception grammars rather than a common L1-L2 phonological space (see Colantoni et al. (2015) for a thorough comparison of these three models). According to the model, Japanese listeners' perception of AmE high front vowels follows a SIMILAR scenario, in which the AmE /i:/-/ɪ/ contrast is perceived as similar to the Japanese /ii/-/i/ contrast. At the initial state of L2 acquisition, learners start with a copy of their existing L1 perception grammar through which L2 sounds are perceived (Full Copying hypothesis). The duplicated L2 perception grammar then gradually modifies itself as it receives perceptual input in the L2, independently from the L1 perception grammar. Eventually, the L2 grammar becomes optimal for perceiving L2 sound contrasts, while the L1 grammar remains optimal for L1 perception (optimal perception hypothesis). As for Japanese listeners' perception of AmE /i:/-/ɪ/, the model predicts that they initially show predominantly duration-based perception in both L1 and L2. However, with increased exposure to the L2 input, their cue weighting in L2 perception gradually shifts from duration to spectra, ultimately resulting in native-like (i.e. predominantly spectral) perception.

On the other hand, their L1 perception will remain unaffected because a shift in cue weighting in the L1 grammar would result in inaccurate perception of L1 sound contrasts, which the listeners would not favor. Importantly, L2LP posits that the L1 and L2 grammars can be activated selectively or in parallel during online speech perception. The model thus predicts that Japanese learners of English would show duration-based perception for high front vowels if the L1 Japanese grammar is active, while they would rely more on spectra and less on duration if the L2 AmE grammar is being activated. Depending on the activation levels of the two grammars, learners may also show an intermediate perceptual behavior in which both durational and spectral cues are used.

L2LP's notion of separate L1 and L2 perception grammars that can be simultaneously activated is based on Grosjean's language mode hypothesis, which is defined as 'the state of activation of the bilingual's languages and language processing mechanisms at a given point of time' (Grosjean, 2001: 2). According to Grosjean, language mode can be seen as a continuum between a monolingual mode and a bilingual mode with varying activation levels of the two languages involved. Activation levels depend on a number of psychosocial and linguistic factors, such as the language of the experimenter, the task, the stimuli, the instructions, and so on. Studies suggest that language mode affects how speech sounds are perceived (for a review, see Simonet, 2016). One of the earliest studies demonstrating such effects is Elman et al. (1977), in which Spanish–English bilinguals were tested on a series of natural stimuli differing in voice onset time or VOT (from /ba/ to /pa/). Each stimulus was preceded by an auditory language-appropriate precursor sentence such as *write the word* or *escriba la palabra* to manipulate the language context. The study found that bilinguals switch their identification of ambiguous stimuli that would be classified as /p/ in Spanish and /b/ in English, depending on which precursor sentence is presented. The effect was larger for more proficient bilinguals, some of whom showed a virtually complete shift between the two conditions. Although the use of the precursor sentences in the study could be somewhat problematic as it can result in phonetic context effects (e.g. the mere presence of [p] in *palabra* may shift the perceptual boundary), García-Sierra et al. (2009) also found significant language-context effects in phonetic judgments between /g/ and /k/ by Spanish–English bilinguals that correlated with their L2 proficiency, without any effect of precursor sentences. Other studies with more sophisticated methodology to elicit different language modes also found similar effects (García-Sierra et al., 2012; Gonzales and Lotto, 2013). While most studies on perceptual mode effects focused on voiced and voiceless obstruents on a VOT continuum, similar effects were found for vowels as well. Escudero (2009) investigated categorization of /ɛ/ and /æ/ by Canadian English (CE) learners of Canadian French (CF), and found that CE learners of CF shift their cue weighting (duration vs. spectra) according to the language context. The degree of such shift correlated with the learners' proficiency in L2 CF. Escudero and Boersma (2002) also found evidence for an L1–L2 intermediate language mode, in which Dutch speakers of Spanish perceived the same vowel tokens differently when they were told to classify 'Dutch' vowels into Dutch vowel categories (i.e. L1 mode), 'Spanish' vowels into Dutch categories (i.e. L1–L2 mode) and 'Spanish' vowels into Spanish categories (i.e. L2 mode). In sum, previous research suggests that language modes can affect L2 learners' perception of both vowels and consonants, of which magnitude appears to be related to L2 proficiency.

Despite their theoretical relevance, perceptual effects of language mode have not received much attention with respect to L2 perception models. The present study therefore investigates whether Japanese speakers' perceptual cue weighting for high front vowels is affected by language contexts (L1 Japanese or L2 AmE), through the use of the L2LP model that incorporates the language mode hypothesis (Escudero, 2005: 118–21). In what follows, a computational implementation of L2LP will be presented to help make the theoretical predictions more explicit (Section II). A perception experiment is then presented in Section III, which manipulates durational and spectral cues in order to investigate whether the reliance on either cue by Japanese learners of English actually changes depending on their language modes ('Japanese' and 'English'). The computational and experimental results are then discussed in Section IV together with implications for SLM and PAM(-L2). Section V provides the conclusions.

II Simulations


In this section, we present a computational implementation of L2LP to simulate Japanese listeners' perceptual acquisition of high front vowels in L1 Japanese and L2 AmE. The model consists of two parts: a phonological grammar that shapes perception and an acquisition mechanism that builds the grammar. The model's perception grammar is based on Stochastic Optimality Theory (Boersma, 1997, 1998), which is a probabilistic extension of Optimality Theory (OT) (Prince and Smolensky, 1993, 2004), and the model's acquisition mechanism is based on the Gradual Learning Algorithm (GLA), an error-driven algorithm for learning optimal constraint rankings from the input data in Stochastic OT (Boersma and Hayes, 2001). We first explain how Stochastic OT represents the process of accessing abstract, phonological categories from varying acoustic inputs through ranked constraints, then how the GLA evaluates the input data to build and update the perception grammar.

I Stochastic Optimality Theory

Stochastic OT was originally developed in phonology, but has also been applied in other areas of linguistics including SLA. Although the theoretical component of L2LP can be implemented with various computational frameworks (e.g. neural network-like modeling; van Leussen and Escudero, 2015), computational implementations of L2LP have generally utilized Stochastic OT (Boersma and Escudero, 2008; Escudero and Boersma, 2004). Stochastic OT combined with the GLA in particular can outperform other learning algorithms such as Naive Bayesian (Escudero et al., 2007), and thus the current study also utilizes the two computational frameworks. The use of Stochastic OT together with the GLA allows L2LP to make very specific predictions as to how L1 and L2 experience shapes one's perception.

Escudero and Boersma (2004) proposed that perception of sound categories can be modeled by Optimality Theoretic, negatively formulated constraints such as 'a value of x on the auditory continuum f should not be mapped to the phonological category y .' Here, we restrict the relevant auditory continua f to the F2/F1 ratio² and duration of vowels, and the relevant phonological categories y to high front vowels in Japanese and AmE. Table 1 illustrates how such constraints can express a native Japanese speaker's

Table 1. Japanese speaker’s perception of a token of Japanese /i/.

[F2/F1 = 7.5 Duration = 120]		[duration = 120] not /ii/	[F2/F1 = 7.5] not /i/	[F2/F1 = 7.5] not /ii/	[duration = 120] not /i/
	/ii/	*!		*	
	/i/		*		*

perception of a token of a Japanese high front vowel with e.g. an F2/F1 ratio of 7.5 (e.g. F1 = 300 Hz, F2 = 2250 Hz) and a duration of 120 ms. At the top of the left-most column is the input, namely the vowel, followed by candidates for the perceptual output, i.e. what the listener perceives given the input. There are four relevant constraints (two auditory dimensions \times two vowel categories), and their ranking determines the vowel category that is perceived. In this example, the highest ranked constraint is ‘[duration = 120 ms] is not /ii/’. Therefore, whereas the two spectral constraints prefer the perception of /ii/, the short /i/ is chosen as the winner.

Stochastic OT differs from traditional OT in that the former allows for probabilistic variations while the latter does not. A vowel with a duration of 120 ms as above, actually falls between the typical lengths of short and long vowels in Japanese, and listeners show variable behavior in perceiving such ambiguous tokens. However, because constraints are strictly ordered in traditional OT, the constraint ranking in Table 1 predicts that a 120 ms-long vowel will be perceived as short without exception. Stochastic OT resolves this problem by arranging constraint rankings on a continuous scale rather than an ordinal one, and allowing the constraint rankings to shift. Each constraint is assigned a ranking value, which represents constraint strength (e.g. 100.0). The ranking value is not static, but rather it is temporarily perturbed by a random value within a set range (e.g. ± 2.0) called evaluation noise at each time of evaluation. The resulting value, called selection point, is used for evaluating the candidates.

In essence, this process is drawing a random sample (i.e. selection point) from a normal distribution with the ranking value as the mean μ and the evaluation noise as the standard deviation σ . For example, if a constraint has a ranking value of 100.0 and evaluation noise of 2.0, a selection point of 100.8, 101.3, 99.6, etc. can be drawn at each evaluation. Since selection points (constraint strengths at time of evaluation) change every time, constraint rankings are not absolute as in regular OT (e.g. $C_1 \gg C_2$) but are probabilistic (e.g. C_1 with a ranking value of 100.0 usually outranks C_2 with a ranking value of 98.0, but C_2 may outrank C_1 in some occasions), enabling Stochastic OT to deal with probability and variation in speech perception and other linguistic phenomena. Note that the ranking of selection points is strict at each evaluation as in regular OT.

2 Gradual Learning Algorithm

Given the probabilistic constraint-based framework of Stochastic OT, the question of perceptual acquisition becomes one of learning what the constraints are and how they should be ranked. For the sake of simplicity, the model is provided with the necessary

Table 2. Japanese child’s misperception of a long token /ii/.

[F2/F1 = 7.5 duration = 120]		[duration = 120] not /ii/	[F2/F1 = 7.5] not /i/	[F2/F1 = 7.5] not /ii/	[duration = 120] not /i/
✓	/ii/	*!→		*→	
✗	/i/		←*		←*

spectral and durational constraints, thus limiting the model’s task to learning just the ranking values of the provided constraints (see Boersma et al. (2003) for discussion on how constraints emerge). Constraint rankings are learned through the Gradual Learning Algorithm (GLA), an error-driven algorithm for learning optimal constraint rankings from the input data in Stochastic OT (Boersma and Hayes, 2001). The GLA is error-driven in that it adjusts the ranking values when there is a mismatch between what the listener perceived based on the most current grammar and what the speaker intended. Ranking values are adjusted by a small number called plasticity (e.g. 1.0), which simulates the listener’s neural plasticity. Plasticity is set to gradually decrease over time, making learning fast but imprecise at an early stage (e.g. infancy) and slow but accurate at a later stage (e.g. adulthood). The gradual learning scheme thus takes into account the well-attested effect of age on L2 acquisition (Flege et al., 1995).

Table 2 illustrates how the GLA updates a perception grammar. Suppose that a Japanese mother produces the word *oniisan* /oniisaN/ ‘older brother’ with a 120ms medial vowel. A Japanese child with the grammar in Table 1 will incorrectly perceive the word as *onisan* /onisaN/ ‘ogre’. Perceptual learning due to semantic or contextual feedback has been shown to occur in native speech for both infants (ter Schure et al., 2016) and adults (Norris et al., 2003) as well as in nonnative speech (Kriengwatana et al., 2016). Therefore, having incorrectly perceived a short vowel (shown by ‘✓’), the semantic context may tell the child that a long vowel should have been perceived instead (shown by ‘✗’) based on lexical knowledge and the semantic context. When the learner notices a mismatch, the GLA strengthens the constraints that led to the perception of the incorrect winner by raising the ranking values (shown by ‘←’) and weakens the constraints that would lead to the correct form by lowering the ranking values (shown by ‘→’), in order to increase the probability of perceiving the same token correctly as /ii/ at the next evaluation. The ranking values are increased and decreased by the learner’s current plasticity.

L2LP proposes that the mechanism behind L2 perceptual learning is the same as in L1 but with the caveat that the learner’s plasticity and current L1 perception grammar is copied over to serve as the base for L2 learning. Since the L2 grammar is a copy, the L1 grammar remains unchanged, and because plasticity decreases with age (i.e. number of iterations), L2 learning is predicted to be slower.

3 Simulating L1 Japanese and L2 AmE perception

In order to simulate Japanese listeners’ perception of Japanese /ii/ and /i/ and AmE /i:/ and /ɪ/, we first collected detailed acoustic information of the target sounds in the two languages.

Table 3. Mean F1, F2, F2/F1 and duration of high front vowels in AmE and Japanese.

Language	Vowel	F1 (Hz)	F2 (Hz)	F2/F1	Duration (ms)
AmE	/i:/	342	2,322	6.79	243
AmE	/ɪ/	427	2,034	4.76	192
Japanese	/ii/	294	2,206	7.50	188
Japanese	/i/	302	2,091	6.92	63

Table 3 shows the mean F1, F2, F2/F1 ratio and duration of high front vowels in the two languages as produced by male native speakers. The acoustic values for AmE were taken from Hillenbrand et al. (1995), in which 45 male monolingual AmE speakers read aloud a randomized list of isolated /hVd/ syllables (e.g. ‘heed’ and ‘hid’), three times each. Using the same procedure, we recorded 20 male native Japanese speakers’ production of /hVda/ (i.e. /hiida/ and /hida/)³ in Japanese. All speakers were from the greater Tokyo area and had not lived outside of Japan for more than one year (mean age = 25.1). Their utterances were recorded with a Sony F-780 microphone (sampling rate 44.1 kHz, 16-bit quantization) in an anechoic chamber at Waseda University. As can be seen, the AmE vowels differ in both spectral and duration values, whereas the Japanese vowels differ predominantly in duration.

The vowels also differ in their frequency distributions. According to the CMU pronouncing dictionary, /ɪ/ appears approximately 1.5 times more frequently than /i:/ in AmE. According to the Corpus of Spontaneous Japanese (Maekawa, 2003), approximately 90% of Japanese vowels are short, while the remaining 10% are long (Bion et al., 2013). We thus assume that short /i/ is nine times more frequent than long /ii/ in Japanese. The simulations make use of the mean acoustic values in Table 3 as well as the above frequency distributions to train the model.

In order to precisely model the perceptual space, we first chose a range of F2/F1 ratio from 4.65 (F1 = 430 Hz and F2 = 2000 Hz; most /ɪ/-like) to 6.76 (F1 = 340 Hz and F2 = 2300 Hz; most /i:/-like) and a range of duration from 70 ms to 240 ms. The spectral range was chosen so that a monolingual Japanese speaker would perceive only /i/-like vowel qualities,⁴ while a monolingual AmE speaker would hear the spectral difference between /i:/ and /ɪ/. The duration range was chosen to cover the entire durational variability of high front vowels in both languages. We then divided each range into 21 logarithmically equal ‘bins’ (in log₂, following Escudero and Boersma (2004)), since human speech perception tends to be logarithmic rather than linear. Each bin had a pair of constraints, one prohibiting the perception of the long or tense vowel category (/ii/ or /i:/) and the other prohibiting the perception of the short or lax category (/i/ or /ɪ/). Here, we assumed that Japanese /ii/ and AmE /i:/, as well as Japanese /i/ and AmE /ɪ/, are representationally equal in Japanese speakers’ perception grammar.⁵ We thus used 84 (2 auditory continua × 21 bins × 2 vowel categories) constraints to model the perception of L1 Japanese and L2 AmE.

The procedure of the simulation is as follows. Initially, the virtual learner has a ‘blank’ perception grammar in which all 84 constraints has the same ranking values of 100.0. The evaluation noise is fixed at 2.0 in both L1 and L2 simulations. The learner then starts acquiring Japanese, receiving tokens of Japanese /ii/ and /i/ occurring randomly at different frequencies (10% and 90%, respectively). The acoustic values (i.e. F2/F1 and duration) of

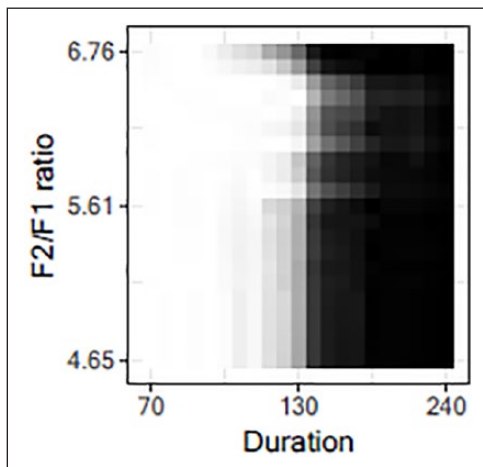


Figure 1. Model's perception of first language (L1) Japanese /ii/ (black) and /i/ (white).

a token are randomly drawn from normal distributions with the mean F2/F1 and duration being those in Table 3 and the standard deviations being 0.1 for F2/F1 and 0.4 for duration (in \log_2). The choice of the standard deviations is, although somewhat arbitrary, based on actual observations in our production data (F2/F1 = 0.17 and duration = 0.27 for /ii/; F2/F1 = 0.20 and duration = 0.39 for /i/). The acoustic values are then rounded to the nearest bins to be evaluated by the relevant constraints. Whenever there is a mismatch between the perceived form and the intended form, the model updates the ranking values of relevant constraints by adding or subtracting the plasticity value. The plasticity is initially set at 1.0, which gradually decreases by a factor of 0.7 every 10,000 tokens (i.e. current plasticity \times 0.7). The parameter settings for evaluation noise and plasticity are based on previous studies (Boersma and Escudero, 2008; Escudero and Boersma, 2004).

Figure 1 shows that the model learns a strict duration-based perception when trained on 40,000 tokens of Japanese /ii/ (black) and /i/ (white). The figure was obtained by feeding 441 F2/F1-duration pairs (21 spectral bins \times 21 duration bins) as input to the model 1,000 times. Darker color indicates that long /ii/ is more likely to be perceived. Despite the low frequency of /ii/ in the input data, the learner successfully acquired a clear length distinction between /ii/ and /i/, without any apparent influence of vowel spectra.

To simulate L2 AmE learning, the L1 Japanese model above was trained on AmE vowels. Following the L2LP model's Full Copying hypothesis, the initial state for L2 acquisition was a copy of the model's L1 perception grammar, in which there was a perfect correspondence between Japanese /ii/ and AmE /i:/ and between Japanese /i/ and AmE /i/. That is, for example, the constraint '[duration = 70ms] is not /ii/' in the L1 perception grammar was copied as '[duration = 70ms] is not /i:/' to the L2 perception grammar with the same ranking value. In the same way as L1 acquisition, the learner received tokens of AmE /i:/ and /i/ occurring randomly at different frequencies (40% and 60%, respectively). The acoustic values were randomly drawn from normal distributions with the means from Table 3. The standard deviations of the normal distributions were again 0.1 for F2/F1, but 0.8 for duration.⁶ The standard

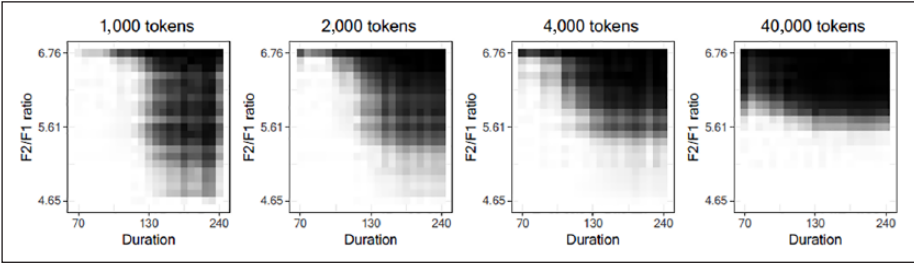


Figure 2. Virtual Japanese listener’s perception of AmE /i:/ (black) and /ɪ/ (white).

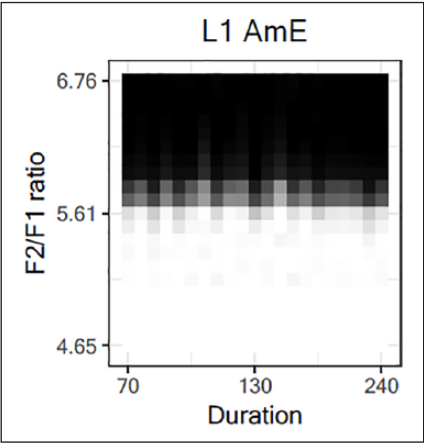


Figure 3. Model’s perception of L1 AmE /i:/ (black) and /ɪ/ (white).

deviation for duration was doubled for L2 AmE simulation for two reasons. Firstly, AmE high front vowels are expected to show more variability in duration as it is not a deterministic cue, whereas Japanese long and short vowels should exhibit more systematic variation in duration. Secondly, we consider that the durations reported in Hillenbrand et al. (1995) may be unnaturally long, perhaps because they were extracted from very careful speech. In fact, other studies such as Nishi et al. (2008) report much shorter values (e.g. 100ms for /i:/ and 86ms for /ɪ/ in citation form). Thus, we attempted to make sure that shorter durations do occur in the listening environment by simply increasing the standard deviation for duration. The plasticity was inherited from L1 acquisition and continued to decrease at the same rate. For comparison, we also modeled the perception of AmE /i:/ and /ɪ/ by a native AmE speaker with the same parameters (except that the plasticity was initialized to 1.0).

Figure 2 shows the outcome of the learner’s acquisition of /i:/ and /ɪ/ after receiving 1,000, 2,000, 4,000 and 40,000 tokens of AmE vowels. The figure was obtained in the same way as Figure 1. Darker shades indicate the likelihood of tense AmE /i:/ perception. As can be seen, a gradual shift in cue weighting from duration to spectra occurred as the learner received more input. The final stage of L2 learning is very close to the simulated native AmE listener’s perception (Figure 3), although the learner is slightly more likely to perceive tokens with short durations as /ɪ/.

Note that in principle, this shift in perceptual cue weighting occurs in the copied L2 grammar only; the L1 perception grammar is considered to remain intact. Therefore, when the learner is in L1 Japanese mode, predominantly duration-based responses as in Figure 1 should be observed, whereas when in L2 AmE mode, the learner will rely more on spectral cues and less on durational cues as in Figure 2. We tested these predictions by comparing the results of the simulations to real listeners' perception, which is presented in Section III.

III Experiment

I Participants

Thirty-two native Japanese speakers participated in the experiment (20 female, 12 male, mean age = 21.5). Twenty-seven of the participants were graduate or undergraduate students at Waseda University, while others were graduates of the University or of other universities in Japan. Demographic information was obtained by a questionnaire the participants answered prior to the experiment. All of them had received compulsory English language education in secondary schools in Japan (age 13–18), which focused primarily on reading and writing skills. Participants had also received some English instruction during college, of which quality and quantity varied depending on the courses they enrolled in. In addition, 15 participants had lived in the United States; 11 of them had spent less than a year (seven to twelve months) on an undergraduate study abroad program, while the remaining four had spent more (e.g. four years) at varying ages. The other 17 participants had not lived outside of Japan for more than one month. None of the participants reported any history of hearing impairment.

2 Stimuli

The stimuli were 49 synthetic vowels differing in spectral and duration values (Figure 4), created using the Klatt synthesizer (Klatt and Klatt, 1990) in Praat (Boersma and Weenink, 2018). The F1 and F2 values co-varied in seven logarithmically equal steps (\log_2), with F1 ranging from 340 Hz to 430 Hz and F2 ranging from 2000 Hz to 2300 Hz. The duration values ranged from 70 ms to 240 ms in another seven logarithmic steps. The spectral and durational ranges are therefore identical to the ones used in the simulations. The stimuli on the top row have the most /i:/-like spectral properties, while those on the bottom row are spectrally most /ɪ/-like. For statistical analysis, the spectral steps were assigned a number from '1' to '7' from low to high so that a high spectral step indicates a tense vowel quality (white numbers in black circles in Figure 4). Likewise, the duration steps were assigned seven numbers so that a high duration step indicates long vowel duration (black numbers in white circles in Figure 4). Fundamental frequency (F0) was fixed at 140 Hz, following Hillenbrand et al.'s measured F0 values for /i:/ (130 Hz) and /ɪ/ (130 Hz). Intensity was fixed at 70 dB.

3 Procedure

To test participants' perception in L1 Japanese and L2 English, the experiment included Japanese (JP) and English (EN) sessions. Participants were informed that they would

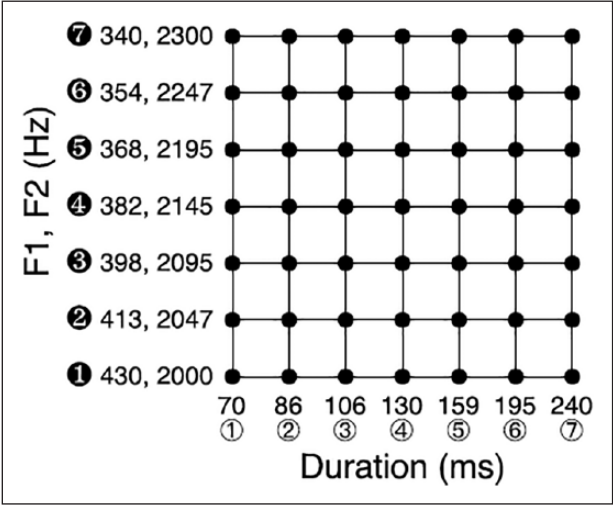


Figure 4. Acoustic values of 49 stimuli used in the experiment. White numbers in black circles represent spectral steps and black numbers in white circles represent duration steps.

hear sounds from different languages (i.e. Japanese or English) between sessions, although they in fact heard exactly the same set of stimuli as in Figure 4 in both sessions. In order to manipulate the participants’ language modes, a pre-recorded audio instruction of the task was first played for the participants immediately before each session, where the instructions were recorded in Japanese by a male native Japanese speaker for the JP session and in English by a male native AmE speaker for the EN session. The experimenter, who was a Japanese speaker of English, also interacted with the participants in the language of the session. After the pre-recorded instruction, participants heard the 49 synthetic stimuli repeated five times in random order, for a total of 245 trials and responses per session.

In the JP session, participants had to decide whether the sound they heard was /ii/ or /i/ in Japanese for each stimulus by clicking either an illustration of *oziiisan* /oziisan/ ‘elderly man’ or that of *ozisan* /ozisan/ ‘middle-aged man’ displayed on a computer screen. Illustrations instead of letters were used to avoid orthographic interference. The EN session followed a similar procedure, where the task was to identify each of the stimuli as either /i:/ or /ɪ/ in English by clicking either an illustration of *sheep* /ʃi:p/ or that of *ship* /ʃɪp/.

The two sessions were consecutive, and session order was counter-balanced across participants to control for order effects. Sixteen participants attended the JP session first, and the other 16 attended the EN session first. Participants were tested individually in an anechoic chamber at Waseda University. The experiment was run on a Macintosh computer using Praat’s ExperimentMFC (multiple forced choice). The audio instructions and stimuli were played at a comfortable volume via Sennheiser HD 380 Pro headphones. The whole experiment took approximately 30 to 40 minutes to complete.

4 Analysis

In order to quantitatively investigate the participants' relative reliance on spectral and durational cues, logistic regression analysis was applied to the obtained response data. Logistic regression is a type of regression analysis where the dependent variable is categorical (and usually binary, e.g. /ii/ or /i/), which is suitable for analyzing identification response data from speech perception experiments (Morrison, 2007). The dependent variable is expressed as log odds, i.e. natural logarithm of the probability that an event occurs (e.g. participant chooses /ii/) divided by the probability that its complementary event occurs (e.g. participant does not choose /ii/, i.e. /i/ is chosen). The logistic regression model used in our study is given in (1):

$$\text{Ln}\left(\frac{P}{1-P}\right) = \alpha + \beta_{\text{spec}} \times \text{spectral step} + \beta_{\text{dur}} \times \text{duration step} \quad (1)$$

In the equation, P is the probability that the participant chooses /ii/ in the JP session or /i:/ in the EN session. The constant α is the intercept of the regression model. The coefficients (β s) show to what extent the seven spectral and seven duration steps cause a change in the log odds of a participant's response. These coefficients therefore can be taken as a participant's reliance on each of the cues in identifying the vowels. For example, if β_{spec} is small and β_{dur} is large, it means that the participant's reliance on vowel spectra is low and his or her reliance on vowel duration is high. As explained in Section III.2, numbers were assigned to the steps so that a large spectral step equates with tense quality and a large duration step equates with long duration.

The two coefficients can also be used to calculate cue weighting as in (2), which represents relative weighting of spectral cues over durational cues (Casillas, 2015; Escudero et al., 2009):

$$\text{Cue weighting} = \frac{\beta_{\text{spec}}}{\beta_{\text{spec}} + \beta_{\text{dur}}} \quad (2)$$

where a value above 0.5 means that vowel spectra is weighted heavier than duration.

Visual inspection of the data as combined with the logistic regression analysis revealed that a few participants showed unexpected perceptual behavior, and their data were excluded from further statistical analysis. Firstly, one participant chose tense /i:/ when the spectral step was low in the EN session (i.e. she seems to have mixed up the labels), which was indicated by a negatively large β_{spec} . Another two participants showed unexpected perception in the JP session, where long /ii/ was perceived when the spectral steps were low, again leading to negative β_{spec} . Although we are unsure of why these participants exhibited such perception patterns, unintended associations between the stimuli and the illustrations might have been established during the experiment.

Furthermore, to directly compare the participants' responses with the simulation results, logistic regression analysis was also applied to the virtual learner's perception. The virtual learner, who was trained first on 40,000 Japanese tokens and subsequently on 1,000, 2,000, 4,000 and 40,000 AmE tokens (in the same way as Section II), 'participated

in the experiment' where the 49 stimuli in Figure 4 were presented five times in each session mimicking stimuli presentation for real participants. Separate L1 Japanese and L2 AmE grammars were used for the JP and EN sessions, respectively. In addition, a virtual native AmE learner who was trained on 40,000 AmE tokens was also tested on the stimuli for the EN session.

5 Results

Table 4 provides by-participant and -session results of logistic regression analyses. Participants have been sorted in the order of cue weighting (from low to high) in the EN session. When aggregated (Figure 5), the results suggest that although duration is a stronger cue in general, participants tend to use more spectral cues and less durational cues in the EN session than in the JP session. We fitted linear mixed effects (LME) models to the response data (except for the three excluded participants), which tested whether β_{spec} , β_{dur} and cue weighting were significantly affected by a fixed effect of session (EN or JP) with participant and session order (EN first or JP first) as random intercepts. We used the lme4 package (Bates et al., 2015) to build the models and the lmerTest package (Kuznetsova et al., 2017) to calculate estimates and statistical significance in R (R Core Team, 2017). The analysis found that session indeed affected β_{spec} , β_{dur} and cue weighting. In the EN session, participants' responses were significantly more dependent on β_{spec} (estimate = 0.45, s.e. = 0.17, $t = 2.63$, $p = .014$) and significantly less dependent on β_{dur} (estimate = -0.48, s.e. = 0.22, $t = -2.19$, $p = .037$) than in the JP session. Accordingly, their cue weighting was significantly larger in the EN session compared to the JP session (estimate = 0.34, s.e. = 0.10, $t = 3.50$, $p = .002$). These results indicate that participants relied more on vowel spectra and less on vowel duration when they thought they were listening to English as opposed to when they thought they were listening to Japanese.

While no effect of participants' L2 proficiency or experience was found,⁷ there was substantial individual variability in the participants' response patterns. This is illustrated in Figure 6, in which darker color indicates more frequent perception of /i:/ for (EN session) and /ii/ (JP session). As can be seen, Participant 3 relied exclusively on duration in both sessions, Participant 19 relied on both duration and spectra in the EN session but only on duration in the JP session, and Participant 23 relied exclusively on spectra in the EN session but exclusively on duration in the JP session. These differences are also reflected in their cue weighting in the EN session, i.e. -0.03, 0.46 and 0.95, respectively. As for the EN session, more than half of the participants whose cue weighting was below 0.5 used duration as the primary cue, whereas several others whose weighting was above 0.5 can be thought to rely primarily on vowel spectra. On the other hand, cue weighting tended to be very small for the JP session, suggesting a strong reliance on duration across participants. Yet, one participant, Participant 21 (Figure 6, far right), showed a surprising but intriguing perception pattern in the JP session. She was tested in the EN session first, and her perception was largely dependent on vowel spectra not only in the EN session but also in the JP session. That is, she showed native AmE-like, spectrally oriented perception in the EN session, which she continued to use in the subsequent JP session.⁸ Her strong reliance on vowel spectra is reflected in her relatively large β_{spec} and cue weighting in both sessions. This interesting perceptual pattern is discussed further in Section IV.

Table 4. Result of logistic regression analysis for each participant per session. Excluded participants shaded in gray.

ID	β_{spec}		β_{dur}		Weighting	
	JP	EN	JP	EN	JP	EN
1	-0.05	-0.21	2.55	3.16	-0.02	-0.07
2	0.10	-0.17	4.29	2.96	0.02	-0.06
3	-0.03	-0.11	2.34	3.48	-0.01	-0.03
4	0.08	-0.03	1.54	1.43	0.05	-0.02
5	-0.05	-0.03	3.46	4.28	-0.02	-0.01
6	-0.24	0.00	2.37	3.43	-0.11	0.00
7	-0.10	0.01	2.43	1.67	-0.04	0.01
8	-0.16	0.02	2.45	2.74	-0.07	0.01
9	0.21	0.07	1.90	2.51	0.10	0.03
10	0.09	0.07	1.58	2.32	0.05	0.03
11	-0.02	0.08	1.52	2.19	-0.01	0.03
12	-0.08	0.13	1.83	2.54	-0.04	0.05
13	0.29	0.16	1.74	2.23	0.14	0.07
14	0.04	0.18	2.75	2.36	0.01	0.07
15	0.12	0.14	1.85	1.62	0.06	0.08
16	-0.08	0.24	2.33	2.40	-0.04	0.09
17	-0.05	0.22	2.20	1.94	-0.02	0.10
18	0.18	0.29	1.95	1.32	0.09	0.18
19	0.10	0.54	2.03	0.63	0.05	0.46
20	0.22	0.88	1.47	0.38	0.13	0.70
21	2.46	1.23	-0.23	0.40	1.10	0.75
*22	-0.31	0.43	0.41	0.14	-3.15	0.76
23	0.20	3.93	2.79	0.22	0.07	0.95
24	0.36	1.69	1.87	0.07	0.16	0.96
*25	-1.87	3.40	0.13	0.07	1.08	0.98
26	0.10	1.24	2.78	-0.04	0.03	1.03
27	0.62	2.20	2.18	-0.11	0.22	1.05
28	-0.10	1.02	0.74	-0.07	-0.15	1.07
29	0.00	1.01	1.35	-0.07	0.00	1.07
30	-0.26	2.04	1.27	-0.19	-0.26	1.10
*31	0.54	-1.09	0.04	0.40	0.92	1.58
32	0.07	0.23	2.30	-0.09	0.03	1.67

The participants’ responses are directly comparable with those of the virtual learner, which is presented in Table 5. L1 Japanese grammar is characterized by a small β_{spec} and large β_{dur} , resulting in a very small cue weighting. This is comparable to most participants’ responses in the JP session and to some participants’ responses in the EN session whose cue weighting is below 0.5. As the virtual learner received more input in L2 AmE,

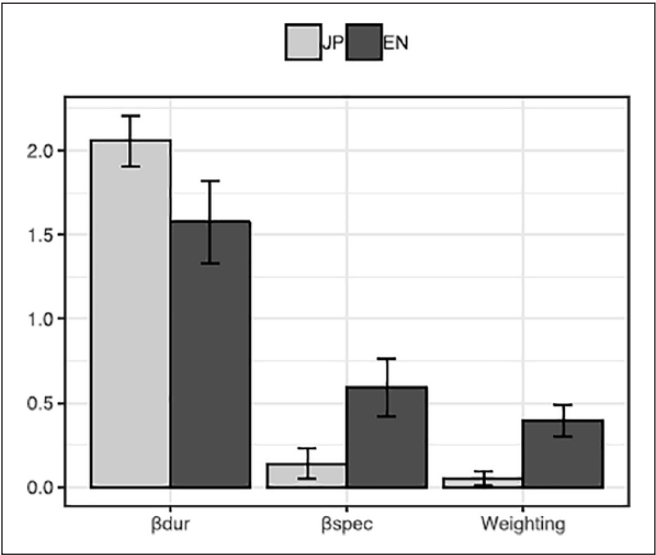


Figure 5. Mean β_{dur} , β_{spec} and cue weighting with ± 1 standard errors based on the logistic regression analysis per session, excluding participants *22, *25 and *31.

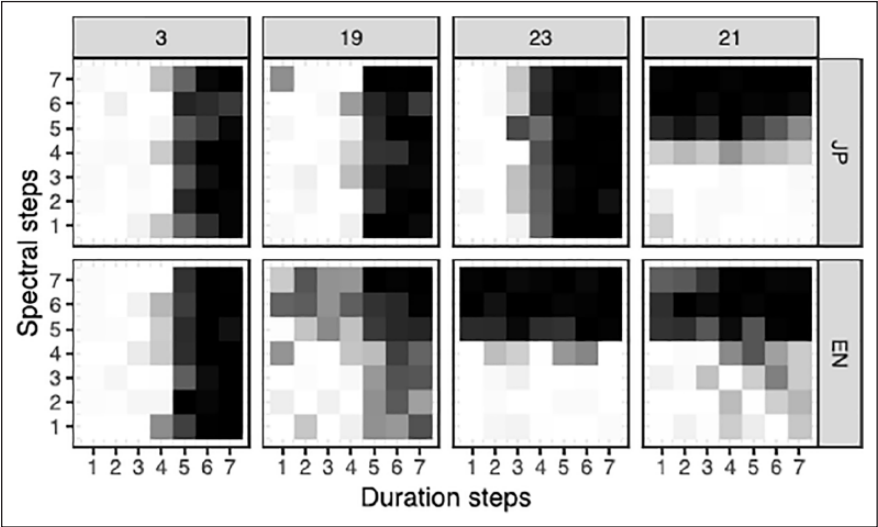


Figure 6. Response patterns of Participants 3, 19, 23 and 21 per session.

the model showed larger β_{spec} and smaller β_{dur} , gradually increasing its cue weighting and thus becoming more ‘native-like’. The real participant’s responses in the EN session whose cue weighting is above 0.5 are, to some extent, comparable to the L2 AmE (e.g. 40,000 tokens) and L1 AmE grammars.

Table 5. Result of logistic regression analysis for the virtual learner.

Grammar	Input	β_{spec}	β_{dur}	Weighting
L1 Japanese	40,000	0.07	1.98	0.03
L2 AmE	1,000	1.28	1.21	0.52
L2 AmE	2,000	1.37	1.17	0.54
L2 AmE	4,000	1.74	0.84	0.67
L2 AmE	40,000	2.65	0.53	0.83
L1 Japanese	40,000	4.50	-0.04	1.01

IV Discussion

I Overall summary

This paper examined through simulations and experiments whether Japanese listeners’ perceptual cue weighting for high front vowels change according to the language context (i.e. L1 Japanese or L2 AmE). Our simulations predicted that, given an adequate amount of L2 input, learners would develop separate perception grammars or perception modes that are appropriate for each language. More specifically, while learners’ perception for L1 Japanese would remain duration-based, their perception for L2 AmE would become more dependent on spectra and less dependent on duration. Learners would be able to switch between the L1 and L2 perception modes and consequently show different cue weighting according to the given language context.

The experimental results supported the simulation predictions. In general, participants’ perception in the JP session was mostly dependent on duration, whereas they relied significantly more on spectral cues and significantly less on durational cues in the EN session, despite the stimuli being exactly the same. However, the degree of such shift in cue weighting varied from individual to individual, ranging from drastic to virtually undetectable. In addition, an unexpected perceptual pattern was also observed where spectral cues were predominantly used in both sessions.

2 Effects of language mode

The current study’s experimental findings are compatible with predictions made by the L2LP model and Grosjean’s language mode hypothesis. As mentioned in the Introduction, L2LP interprets language modes as a selective or parallel activation of separate L1 and L2 grammars during speech perception. More specifically, the model extends Grosjean’s ideas and sees language modes as a continuum from L1 monolingual mode through L1–L2 bilingual mode to L2 monolingual mode, of which control is learned with L2 experience. For the purposes of this study, the language mode continuum would be monolingual L1 Japanese on one end, where listeners rely exclusively on durational cues, and monolingual L2 English on the other, where listeners rely exclusively on spectral cues. The L1–L2 bilingual mode would be intermediary, where both cues may be used. The observed shift in cue weighting in the experimental results can be interpreted as a result of different activation levels of the two grammars. The results are also comparable to our computational

implementation of L2LP, which provided very specific predictions as to how listeners' cue weighting may change according to the language context.

As Grosjean (2001) notes, bilinguals differ as to the extent they travel along the language mode continuum, which can be part of the reason why great individual variability was observed. Take for example Participants 3, 19, 21 and 23 from Figure 6. Participant 23 exhibited a predominantly duration-based perception pattern in the JP session and a spectra-based perception pattern in the EN session, providing an example of successful switching based on the target language. Participant 19 exhibited predominantly duration-based perception in the JP session but a mixture of spectra- and duration-based perception in the EN session, indicating that both languages might have been activated in the latter session. Participant 21, on the other hand, exhibited spectra-based perception in both EN and JP sessions. Given that she was tested in the EN session first, it could be the case that her L2 English perception grammar was strongly activated first, then was not switched off when she was tested in the subsequent JP session. Participant 3 is a similar case but in reverse, where having been tested in the JP session first, a duration-based perception pattern is seen throughout both JP and EN sessions due to a strong activation of their monolingual L1 Japanese mode.

Although we attempted to manipulate the listeners' language mode between sessions and successfully elicited different responses, it should be noted that establishing a language context is not a simple task. In fact, studies that tested mode effects in perception prior to Elman et al. (1977) failed to demonstrate such effects (Caramazza et al., 1974; Williams, 1977). Elman et al. (1977) pointed out that language mode might not have been maintained throughout the identification task in these studies, which could also apply to the current study as language modes were manipulated immediately before each experimental session. In addition, the experimenter in the current study was not a native AmE speaker, potentially hindering mode switching from L1 Japanese to L2 AmE. In future research, subjects could be 'reminded' of the current language context by presenting language-appropriate precursor sentences and/or having a native speaker of each language as an experimenter. Alternatively, since the very use of acoustic precursors might influence perception, language context should perhaps be 'embedded' within the stimuli themselves. For example, Gonzales and Lotto (2013) used a pair of pseudowords, *bafri* and *pafri*, to test Spanish–English bilinguals' perception of /b/ and /p/ in both languages. Crucially, the *ri* portion was pronounced with a tap [ɾ] in Spanish and an approximant [ɹ] in English, which was the only signal of language context (all instructions and conversations were in English). And yet, the bilinguals did show responses that were appropriate for each context. To test the perception of Japanese and English high front vowels, phonological palatalization of certain consonants preceding high front vowels in Japanese (e.g. /si/ becoming [çi]) and the lack thereof in English (e.g. [si:]) could be useful in preparing stimuli where language context is signaled by the consonant.

3 Implications for SLM and PAM(-L2)

While L2LP in conjunction with the language mode hypothesis can straightforwardly explain the current study's experimental results, SLM and PAM(-L2) provide alternative interpretations. Whereas both models predicted that Japanese listeners may maintain duration-based perception for the L2 contrast, which was indeed the case for some

participants, other participants' stronger reliance on spectral cues in the EN session would indicate that a new category had been formed for either /i:/ or /ɪ/, or perhaps both. SLM would explain that, for some learners, equivalence classification resulted in persistent use of duration, whereas for others, noticing phonetic differences between the L1 and L2 categories led to new category formation. From the perspective of PAM(-L2), some learners may have equated the L1 and L2 contrasts both phonetically and phonologically, whereas others may have equated the contrasts only phonologically while being aware of the phonetic dissimilarities. Yet, it is not very clear for which L2 sound new category formation might have occurred. One possibility is that the lax vowel /ɪ/ is prone to new category formation. Strange et al. (2011) reports that, contrary to Strange et al. (1998), Japanese listeners can perceive AmE /ɪ/ as Japanese /e/ instead of /i/, which could be due to dialectal or individual differences in the stimuli between the two studies. Thus, it is possible that Japanese listeners discern the spectral differences between AmE /ɪ/ and Japanese /i/ depending on the specific phonetic realizations and establish a new phonetic category for the L2 sound. On the other hand, new category formation may not be very likely for AmE /i:/ as it was most likely perceived as Japanese /ii/ in both studies.

What the two models must address more explicitly is whether and how L1 and L2 categories can be simultaneously activated within a common space, in real time. Although new category formation can account for the shift in cue weighting found in the current study, it does not adequately explain other studies' finding that a listener can show L1-like, L2-like and L1–L2 intermediate perception according to the language context (Escudero and Boersma, 2002). A couple of studies on bilingual VOT production by Antoniou et al. (2010, 2011) illustrate this point. In these studies, early L2-dominant Greek–English bilinguals produced word-initial and word-medial /p, t, b, d/ in different language contexts: Greek (monolingual L1 mode), English (monolingual L2 mode) and code-switching (bilingual or L1–L2 intermediate mode). It was found that, even though the bilingual' VOTs did not differ from control monolingual speakers' in either L1 Greek or L2 English, they exhibited more Greek-like VOTs in English production when code-switching. These results suggest that the bilinguals had established distinct phonetic categories specific to each language, which interacted dynamically in real time. Antoniou (2010) claims that bilinguals integrate both languages in a common phonetic space, and can selectively attend to language-specific phonetic information depending on the situational language context. Incorporation of such an idea would reinforce the theoretical principles of SLM and PAM(-L2).

4 Avenues for future research

One possible direction for future research is to conduct a perception study where L1–L2 intermediate mode is elicited, as was the case for the code-switched production in Antoniou et al. (2011). If listeners show L1–L2 intermediate perception in the bilingual context, it would be further evidence of language-specific perception modes (or categories) that interact dynamically during online speech perception. This would be a challenging task given the difficulty of mode manipulation, but perhaps the method of Escudero and Boersma (2002) can be used. In the study, Dutch–Spanish bilinguals

classified the same set of CVC tokens in three different tasks: (1) categorizing ‘Dutch’ tokens into Dutch vowel categories (monolingual L1 mode), (2) categorizing ‘Spanish’ tokens into Dutch vowel categories (bilingual mode) and (3) categorizing ‘Spanish’ tokens into Spanish vowel categories (monolingual L2 mode). The second task, which essentially simulates a real-time loan adaptation scenario, is of particular interest because it requires listeners to activate both L1 and L2 representations. The other two tasks, which resemble the JP and EN sessions in the current study, can be referred to as base-lines. As another option, stimuli for perception experiments may also be code-switched (e.g. L2 tokens embedded in L1 carrier sentences and vice versa) to encourage simultaneous activation of L1 and L2 modes.

Another important avenue for future research is to test L2 proficiency effects. As mentioned in Section I, previous research suggests that the magnitude of mode effects tends to be larger for more proficient learners, which was also demonstrated by a recent study on perceptual mode effects in early and late bilinguals (Casillas and Simonet, 2018). A series of studies on phonetic drift by Chang (2011, 2012, 2013) also show that L2 experience can affect L1 production more in novice learners than in experienced learners, suggesting that novice learners have trouble separating L1 and L2 systems. However, the current study did not find any effect of L2 proficiency or experience on perception patterns (see Note 7). Given that only four of the participants had spent more than one year in the USA, a firm L2 monolingual mode may not have been established for the majority of our participants based on their limited exposure to naturalistic L2 input. A follow-up study with L1 Japanese speakers who have resided in the USA for an extended period of time could help elicit clearer effects of proficiency.

Finally, although our computational implementation of L2LP provided explicit predictions for the current study’s findings, some limitations need to be addressed. First, the simulations admittedly simplified the learning scenario by integrating F1 and F2 as ratios because the focus was specifically on the relative cue weighting between vowel spectra and duration. However, both F1 and F2 can have independent effects (Escudero and Chládková, 2010), which the current simulations cannot capture. Thus, future work could implement three-dimensional simulations where F1, F2 and duration are all used. Second, it was assumed in the simulations that spectral and durational cues are equally used, which may not hold true. Bohn (1995) found that native speakers of Spanish and Mandarin, neither of which uses duration to differentiate vowel contrasts, relied heavily or exclusively on vowel duration to distinguish English /i:/–/ɪ/ and /ɛ:/–/æ/ that are mainly distinguished by vowel spectra by native English speakers. Likewise, Escudero et al. (2009) found that Spanish learners of Dutch favored vowel duration over vowel spectra to categorize Dutch /a:/–/a/, which is also distinguished chiefly by vowel spectra by native Dutch speakers. These results indicate that vowel duration can be psychoacoustically salient regardless of its phonemic status in a particular language, which may have affected Japanese listeners’ responses in the current study as well. Perceptual salience could be implemented in the simulations by e.g. increasing plasticity for the auditory dimension of duration. Another property of the current simulations to consider is that they assumed perfect correspondence between AmE /i:/–/ɪ/ and Japanese /ii/–/i/, which perhaps is overly simplistic. As PAM(-L2) explains, it is possible that an L2 category is phonologically, but not phonetically, assimilated into an L1 category. This could be

modeled by a perception grammar that contains multiple levels of representations (e.g. acoustic, phonetic, phonemic and lexical levels) as in Boersma (2011) and van Leussen and Escudero (2015) rather than the simple acoustic- to-phonological grammar assumed in the current study.

V Conclusions

The present study provides computational and empirical evidence that Japanese learners of English employ different cue weighting (duration vs. spectra) to differentiate perceptually similar L1 and L2 contrasts (i.e. Japanese /ii/–/i/ and AmE /i:/–/ɪ/) depending on the given language mode. The experiment found that Japanese listeners used more spectral cues and less durational cues when in ‘English’ mode than in ‘Japanese’ mode, of which magnitude varied but could be extensive for some individuals. Our computational implementation of the L2LP model, which incorporates the language mode hypothesis, predicted and explains the experimental results well. Mode effects in bilingual speech perception have theoretical implications for other L2 perception models as well, and are worth further investigation in future research.

Acknowledgements

We thank Jeff Moore for providing his voice for the English audio instruction and Risa Matsubara for her help with recording the Japanese production data. We would also like to show our gratitude to three anonymous reviewers and the editor for their valuable comments and suggestions that greatly improved the manuscript.

Declaration of Conflicting Interest

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by the Japan Society for the Promotion of Science (grant numbers 15H02729, 18J11517); the Australian Research Council Centre of Excellence for the Dynamics of Language (grant number CE140100041); and the Australian Research Council (grant number FT160100514).

Notes

1. AmE was chosen as the target variety of English because it is used in formal English language education in Japan and therefore is most familiar to the learners.
2. F2/F1 ratio was used to represent vowel tenseness in a single value (larger = more tense).
3. The vowel /a/ was added to /hVd/ because Japanese does not allow a stop coda except for geminates. Lexical pitch accent was placed on the first mora.
4. According to Strange et al. (1998), AmE /ɪ/ is most likely categorized as Japanese /i/ by native Japanese speakers (approximately 60–80%).
5. This assumption comes from not only acoustic similarities between the L1 and L2 sounds but also other factors such as orthography and loanwords. For example, the fact that /ɪ/ is

often spelled as 'i' in English (e.g. *ship*, *pick*, *this*) can lead Japanese speakers to establish a representational connection between English /ɪ/ with Japanese /i/ rather than /e/. Such orthographic factors have been known to affect L2 speech perception (Detey and Nespoulous, 2008; Escudero and Wanrooij, 2010). In addition, Japanese loanwords from English words containing /ɪ/ are usually transcribed with /i/ in Japanese orthography, which may further reinforce the connection.

6. Actual standard deviations are not reported in Hillenbrand et al. (1995).
7. We also ran LME models to test whether the participants' experience of living in the USA affected β_{spec} , β_{dur} and cue weighting a fixed effect, which yielded non-significant results. Although Participant 32, who had lived in the USA for four years, showed a drastic shift in cue weighting between sessions, three other participants (Participants 16, 17, 18) who had spent more than a year in the USA showed only a subtle shift.
8. Note that session order was set as a random effect in the LME models and therefore was controlled for in the statistical analysis. We also ran additional models including session order as a fixed effect, which yielded non-significant results. It is thus speculated that session order did not have a general effect on listeners' responses, although it could have affected perception at an individual level.

ORCID iD

Kakeru Yazawa  <https://orcid.org/0000-0002-0528-6103>

References

- Antoniou M (2010) One head, two languages: Speech production and perception in Greek–English bilinguals. Unpublished PhD thesis, Western Sydney University, Sydney, Australia.
- Antoniou M, Best CT, Tyler MD, and Kroos C (2010) Language context elicits native-like stop voicing in early bilinguals' productions in both L1 and L2. *Journal of Phonetics* 38: 640–53.
- Antoniou M, Best CT, Tyler MD, and Kroos C (2011) Inter-language interference in VOT production by L2-dominant bilinguals: Asymmetries in phonetic code-switching. *Journal of Phonetics* 39: 558–70.
- Bates D, Mächler M, Bolker BM, and Walker SC (2015) Fitting linear mixed-effects models using lme4. *Journal of Statistical Software* 67: 1–48.
- Best CT (1995) A direct realist view of cross-language speech perception. In: Strange W (ed.) *Speech perception and linguistic experience: Issues in cross-language research*. Timonium, MD: York Press, pp. 171–204.
- Best CT and Tyler MD (2007) Nonnative and second-language speech perception: Commonalities and complementarities. In: Munro MJ and Bohn OS (eds) *Language experience in second language speech learning: In honor of James Emil Flege*. Amsterdam: John Benjamins, pp. 13–34.
- Bion RA, Miyazawa K, Kikuchi H, and Mazuka R (2013) Learning phonemic vowel length from naturalistic recordings of Japanese infant-directed speech. *PLOS ONE* 8: e51594.
- Boersma P (1997) How we learn variation, optionality, and probability. In: *Proceedings of the Institute of Phonetic Sciences of the University of Amsterdam: Volume 21*. University of Amsterdam, pp. 43–58.
- Boersma P (1998) Functional Phonology: Formalizing the interactions between articulatory and perceptual drives. Unpublished PhD thesis, University of Amsterdam, Amsterdam, Netherlands.
- Boersma P (2011) A programme for bidirectional phonology and phonetics and their acquisition and evolution. In: Benz A and Mattausch J (eds) *Bidirectional Optimality Theory*. Amsterdam: John Benjamins, pp. 33–72.

- Boersma P and Escudero P (2008) Learning to perceive a smaller L2 vowel inventory: An Optimality Theory account. In: Avery P, Dresher E and Rice K (eds) *Contrast in phonology: Theory, perception, acquisition*. Berlin: Mouton de Gruyter, pp. 271–302.
- Boersma P and Hayes B (2001) Empirical tests of the Gradual Learning Algorithm. *Linguistic Inquiry* 32: 45–86.
- Boersma P and Weenink D (2018) *Praat: Doing phonetics by computer: Version 6.0.37* [computer program]. Available at: <http://www.praat.org/> (accessed February 2019).
- Boersma P, Escudero P, and Hayes R (2003) Learning abstract phonological from auditory phonetic categories: An integrated model for the acquisition of language-specific sound categories. In: Solé MJ, Recasens D, and Romero J (eds) *Proceedings of the 15th International Congress of Phonetic Sciences*. Barcelona, pp. 1013–16.
- Bohn OS (1995) Cross-language speech perception in adults: First language transfer doesn't tell it all. In: Strange W (Ed.) *Speech perception and linguistic experience: Issues in cross-language speech research*. Timonium, MD: York Press, pp. 275–300.
- Caramazza A, Yeni-Komshian G, and Zurif EB (1974) Bilingual switching: The phonological level. *Canadian Journal of Psychology* 28: 310–18.
- Casillas JV (2015) Production and perception of the /i/–/ɪ/ vowel contrast: The case of L2-dominant early learners of English. *Phonetica* 72: 182–205.
- Casillas JV and Simonet M (2018) Perceptual categorization and bilingual language modes: Assessing the double phonemic boundary in early and late bilinguals. *Journal of Phonetics* 71: 51–64.
- Chang CB (2011) *Systematic drift of L1 vowels in novice L2 learners*. In: Lee WS and Zee E (eds) *Proceedings of the 17th International Congress of Phonetic Sciences*. Hong Kong: The University of Hong Kong, pp. 428–31.
- Chang CB (2012) Rapid and multifaceted effects of second-language learning on first-language speech production. *Journal of Phonetics* 40: 249–68.
- Chang CB (2013) A novelty effect in phonetic drift of the native language. *Journal of Phonetics* 41: 520–33.
- Colantoni L, Steele J, and Escudero P (2015) *Second language speech: Theory and practice*. Cambridge: Cambridge University Press.
- Detey S and Nespoulous JL (2008) Can orthography influence second language syllabic segmentation? Japanese epenthetic vowels and French consonantal clusters. *Lingua* 118: 66–81.
- Elman JL, Diehl RL, and Buchwald SE (1977) Perceptual switching in bilinguals. *The Journal of the Acoustical Society of America* 62: 971–74.
- Escudero P (2005) Linguistic perception and second language acquisition: Explaining the attainment of optimal phonological categorization. Unpublished PhD thesis, Utrecht University, Utrecht, Netherlands.
- Escudero P (2009) The linguistic perception of SIMILAR L2 sounds. In: Boersma P and Hamann S (eds) *Phonology in perception*. Berlin: Mouton de Gruyter, pp. 151–90.
- Escudero P and Boersma P (2002) The subset problem in L2 perceptual development: Multiple-category assimilation by Dutch learners of Spanish. In: *Proceedings of the 26th Annual Boston University Conference on Language Development*. Somerville, MA: Cascadilla, pp. 208–19.
- Escudero P and Boersma P (2004) Bridging the gap between L2 speech perception research and phonological theory. *Studies in Second Language Acquisition* 26: 551–85.
- Escudero P and Chládková K (2010) Spanish listeners' perception of American and Southern British English vowels. *The Journal of the Acoustical Society of America* 128: EL254–60.
- Escudero P, Benders T, and Lipski SC (2009) Native, non-native and L2 perceptual cue weighting for Dutch vowels: The case of Dutch, German, and Spanish listeners. *Journal of Phonetics* 37: 452–65.

- Escudero P and Wanrooij K (2010) The effect of L1 orthography on non-native vowel perception. *Language and Speech* 53: 343–65.
- Escudero P, Kastelein J, Weiland K, and van Son RJH (2007) Formal modelling of L1 and L2 perceptual learning: Computational linguistics versus machine learning. In: *Proceedings of the 8th Annual Conference of the International Speech Communication Association*, pp. 1889–92.
- Flege JE (1995) Second language speech learning: Theory, findings, and problems. In: Strange W (ed.) *Speech perception and linguistic experience: Issues in cross-language research*. Timonium, MD: York Press, pp. 233–77.
- Flege JE, Munro MJ, and MacKay IR (1995) Factors affecting strength of perceived foreign accent in a second language. *The Journal of the Acoustical Society of America* 97: 3125–34.
- Fox MM and Maeda K (1999) Categorization of American English vowels by Japanese speakers. In: Ohala JJ, Hasegawa Y, Ohala M, Granville D, and Bailey AC (eds) *Proceedings of the 14th International Congress of Phonetic Sciences*. San Francisco, CA: University of California, pp. 1437–40.
- García-Sierra A, Diehl RL, and Champlin C (2009) Testing the double phonemic boundary in bilinguals. *Speech Communication* 51: 369–78.
- García-Sierra A, Ramírez-Esparza-Esparza N, Silva-Pereyra J, Siard J, and Champlin CA (2012) Assessing the double phonemic representation in bilingual speakers of Spanish and English: An electrophysiological study. *Brain and Language* 121: 194–205.
- Gonzales K and Lotto AJ (2013) ‘A bafri, un pafri’: Bilinguals’ pseudoword identifications support language-specific phonetic systems. *Psychological Science* 24: 2135–42.
- Grosjean F (2001) The bilingual’s language modes. In: Nicol J (ed.) *One mind, two languages: Bilingual Language Processing*. Oxford: Blackwell, pp. 1–22.
- Hattori K and 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: 469–79.
- Hillenbrand JM, Clark MJ, and Houde RA (2000) Some effects of duration on vowel recognition. *The Journal of the Acoustical Society of America* 108: 3013–22.
- Hillenbrand JM, Getty LA, Clark MJ, and Wheeler K (1995) Acoustic characteristics of American English vowels. *Journal of the Acoustical Society of America* 97: 3099–3111.
- Iverson P, Kuhl PK, Akahane-Yamada R, et al. (2003) A perceptual interference account of acquisition difficulties for non-native phonemes. *Cognition* 87: B47–57.
- Klatt DH and Klatt LC (1990) Analysis, synthesis, and perception of voice quality variations among female and male talkers. *The Journal of the Acoustical Society of America* 87: 820–57.
- Kriengwatana B, Terry J, Chládková K, and Escudero P (2016) Speaker and accent variation are handled differently: Evidence in native and non-native listeners. *PLOS ONE* 11: e0156870.
- Kuznetsova A, Brockhoff PB, and Christensen RHB (2017) lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software* 82: 1–26.
- MacKain KS, Best CT, and Strange W (1981) Categorical perception of English /r/ and /l/ by Japanese bilinguals. *Applied Psycholinguistics* 2: 369–90.
- Maekawa K (2003) Corpus of Spontaneous Japanese: Its design and evaluation. In: *Proceedings of the ISCA and IEEE Workshop on Spontaneous Speech Processing and Recognition*. Tokyo, pp. 7–12.
- Morrison GS (2002) Perception of English /i/ and /ɪ/ by Japanese and Spanish listeners: Longitudinal results. In: Morrison GS and Zsoldos L (eds) *Proceedings of the Northwest Linguistic Conference 2002*. Burnaby, BC: Simon Fraser University Linguistics Graduate Student Association, pp. 29–48.
- Morrison GS (2007) Logistic regression modelling for first- and second-language perception data. In: Solé MJ, Prieto P, and Mascaró J (eds) *Segmental and prosodic issues in Romance phonology*. Amsterdam: John Benjamins, pp. 219–36.

- Nishi K, Strange W, Akahane-Yamada R, Kubo R, and Trent-Brown SA (2008) Acoustic and perceptual similarity of Japanese and American English vowels. *The Journal of the Acoustical Society of America* 124: 576–88.
- Norris D, McQueen JM, and Cutler A (2003) Perceptual learning in speech. *Cognitive Psychology* 47: 204–38.
- Prince A and Smolensky P (1993) Optimality Theory: Constraint interaction in generative grammar. *Rutgers University Center for Cognitive Science Technical Report 2*. Rutgers University.
- Prince A and Smolensky P (2004) Optimality Theory: Constraint interaction in generative grammar. Oxford: Blackwell.
- R Core Team (2017) *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing. Available at: <https://www.R-project.org/> (accessed February 2019).
- Simonet M (2016) The phonetics and phonology of bilingualism. In: *Oxford Handbooks Online*. Oxford: Oxford University Press, pp. 1–25.
- Strange W, Hisagi M, Akahane-Yamada R, and Kubo R (2011) Cross-language perceptual similarity predicts categorical discrimination of American vowels by naïve Japanese listeners. *The Journal of the Acoustical Society of America* 130: EL226–31.
- Strange W, Akahane-Yamada R, Kubo R, et al. (1998) Perceptual assimilation of American English vowels by Japanese listeners. *Journal of Phonetics* 26: 311–44.
- ter Schure S, Junge C, and Boersma P (2016) Semantics guide infants' vowel learning: Computational and experimental evidence. *Infant Behavior and Development* 43: 44–57.
- van Leussen JW and Escudero P (2015) Learning to perceive and recognize a second language: The L2LP model revised. *Frontiers in Psychology* 6: 1000.
- Williams L (1977) The perception of stop consonant voicing by Spanish–English bilinguals. *Perception and Psychophysics* 21: 289–97.