Testing Second Language Linguistic Perception

A Case Study of Japanese, American English, and Australian English Vowels

Kakeru Yazawa

January 17, 2020

A doctoral dissertation submitted to
the Graduate School of International Culture and Communication Studies
Waseda University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy

ACKNOWLEDGMENTS

First of all, I would like to thank my principal supervisor, Mariko Kondo, for her advice and guidance throughout the course of this thesis. Not only has she supervised my Ph.D., but she has also guided my entire academic career since I was an undergraduate student, for which I am truly grateful. I would also like to thank Paola Escudero at the MARCS Institute, Western Sydney University, for being my external supervisor and supporting my research visit at the University. This thesis would not have been possible without her. I also express my gratitude to my supervisory panel, Tetsuo Harada and Yasuaki Shinohara, for their constructive comments and suggestions. I also thank my mentors, Takayuki Konishi and James Whang, for their continuous advice and friendship.

I owe thanks to Eline Smit, Jaydene Elvin, Jeff Moore, Mikey Elmers, Nicole Traynor, and Risa Matsubara for their help with running the experiments, and of course to those how kindly participated in the tedious experiments. I would like to thank the Japan Society for the Promotion of Science (JSPS) for financially supporting my research in Japan (DC2 Research Fellowship) and Australia (JSPS Overseas Challenge Program). I am also grateful to the Laboratory for Language Development, RIKEN Center for Brain Science for providing me with a rare opportunity to engage in a research work environment. I express my personal thanks to the following (former) members of the Lab: Reiko Mazuka, Ken’ya Nishikawa, Nobuyuki Jincho, and Kuniyoshi Tanaka.

I met many inspirational figures before and during my Ph.D., who directly or indirectly influenced my research. I am glad that I could see the following people in person: Benjamin Tucker at the University of Alberta; Anne Cutler and Catherine Best at the MARCS Institute; James Flege at the University of Alabama in Birmingham; Kimiko Tsukada at Macquarie University; Jason Shaw at Yale University; and Shigeto Kawahara at Keio University. Many thanks to all my colleagues, including Juqiang Chen, Sanae Matsui, Yumi Ozaki, Amanda Rysling, Erina Sawada, Greg Short, Masaki Takeuchi, Alba Tuninetti, and Marina Kalashnikova, among others.

My special thanks go to Christian Laubman for volunteering to proofread the entire thesis. Dear friend, it must have been a daunting task. All remaining errors are mine. I would like to thank my wife, Karin, for keeping me sane over the past few months with her constant support and encouragement. Finally, I thank my mother, Hisako, for her endless love, support, and understanding. I owe everything I have accomplished to her.

DECLARATION OF AUTHORSHIP

I, Kakeru Yazawa, hereby declare that this dissertation and the work described in it are my own. Where I have consulted the work of others, this is always clearly stated.

Signed: Kakeru Yazawa

Date: January 17, 2019

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF OPTIMALITY THEORY TABLEAUX

ABBREVIATIONS

AmE	American English
AOL	Age of learning
AusE	Australian English
BiPhon	Bidirectional Phonology and Phonetics
C	Consonant
CE	Canadian English
CF	Canadian French
CG	Category-Goodness
CPH	Critical Period Hypothesis
EDCD	Error-Driven Constraint Demotion
EEG	Electroencephalography
EFL	English as a second language
F0	Fundamental frequency
F1	First formant
F2	Second formant
F3	Third formant
fMRI	Functional magnetic resonance imaging
GLA	Gradual Learning Algorithm
HARK	Hypothesize after results are known
IPA	International Phonetic Alphabet
ISI	Inter-stimulus interval
L1	First language
L2	Second language
L2LP	Second Language Linguistic Perception
LME	Linear mixed effects
LP	Linguistic Perception
MCA	Multiple-category assimilation
MMN	Mismatch negativity
NA	Non-Assimilable
OT	Optimality Theory
PAM	Perceptual Assimilation Model
PAM-L2	Perceptual Assimilation Model for L2
SBE	Southern British English
SE	Scottish English
SLA	Second language acquisition
SLM	Speech Learning Model
SMG	Supramarginal gyrus
STG	Superior temporal gyrus
TC	Two-Category
TD	Trajectory direction
TL	Trajectory length
UC	Uncategorized-Categorized
UU	Uncategorized-Uncategorized
V	Vowel
VISC	Vowel inherent spectral change
VOT	Voice onset time

To my maternal family

Chapter 1: Introduction

1.1 Background

The acquisition of native-like competence in one's second language (L2) is typically challenging. It is common for adult L2 learners to have foreign accents, which may have a number of undesirable consequences. For example, foreign accents may cause communication difficulties and also lead to social stigmatization and discrimination (Gluszek & Dovidio, 2010; Hosoda & Stone-Romero, 2010). From early phonological theories, the difficulty of L2 speech production has been attributed to the perception of speech sounds that is 'attuned' to the first language (L1). As Trubetzkoy (1969, p. 52) stated, "(e)ach person acquires the system of his mother tongue. But when he hears another language spoken he intuitively uses the familiar "phonological sieve" of his mother tongue to analyze what has been said. However, since this sieve is not suited for the foreign language, numerous mistakes and misinterpretations are the result." This is the fundamental idea that underlies most current theories and models of L2 speech acquisition. Many studies support the proposition that perception precedes production and that accurate perception is a prerequisite for accurate production (Escudero, 2007), while relatively few studies claim that production might precede or be more accurate than perception in some occasions (Hattori, 2009; Sheldon & Strange, 1982).

Current models of L2 speech acquisition thus emphasize the role of perception. Two models have been widely used in the field: the Speech Learning Model (SLM; Flege, 1995) and the Perceptual Assimilation Model (PAM; Best, 1995). These models commonly propose that perceptual similarities and dissimilarities between L1 and L2 sounds, along with other factors such as the learner's age and the quality of input, predict success in L2 speech acquisition. For example, native Japanese listeners tend to perceive English /θ/ as similar to Japanese /s/, thus producing *three* /θ̲riː/ as [s̲ɾiː]. According to the models, Japanese learners of English may learn to accurately produce English /θ/ if they notice its phonetic dissimilarities from Japanese /s/, although they may also retain their accent if they fail to perceive the differences. Numerous studies have been conducted to test the predictions of SLM and PAM, some of which are reviewed in this thesis. These studies, in general, have provided support for the two models, confirming that perceptual similarities are a good predictor of L2 learners' performance in perception and production. However, the models have also received criticisms over the years that it is unclear how such perceptual similarities can be quantified objectively. For example, given that English /θ/ sounds more similar to /t/ for Russian listeners (Weinberger, 1987), how similar is /θ/ to /s/ or to /t/? The answer would be language-, speaker-, and listener- specific, which is too vague for making reliable predictions.

A more recent model of L2 speech perception, the Second Language Linguistic Perception (L2LP) model (Escudero, 2005), approaches this problem from a different perspective. While L2LP shares certain conceptual similarities with the previous two models, the model is unique in its use of computational simulations. The simulations are based on a probabilistic extension of Optimality Theory (OT; Prince & Smolensky, 1993, 2004) called Stochastic OT (Boersma, 1998) and the associated learning algorithm called the Gradual Learning Algorithm (Boersma & Hayes, 2001). By using these computational frameworks, L2LP is capable of making very specific and testable predictions concerning L2 listeners' perceptual acquisition. For example, Boersma and Escudero (2008) conducted Stochastic OT-based simulations and predicted that Dutch listeners' perceptual boundary between their native vowels /ɛ/ and /ɑ/ on the acoustic continuum of the first formant (F1) frequencies would be approximately 824 Hz. The study also predicted that, when the same Dutch listeners learn Spanish as L2, the perceptual boundary would shift toward 662 Hz to match the acoustic distribution of Spanish /e/ and /a/. Such concrete predictions can be empirically tested with real listeners' perception, which serves as an objective test of the model. However, despite its strength and potential, L2LP has been relatively underutilized in the field, presumably due to the difficulty of implementing computational simulations.

1.2 Purpose of the thesis

The primary aim of this thesis is to provide a comprehensive test of the L2LP model. According to L2LP, three types of learning scenarios can be distinguished depending on the number of sound categories in L1 and L2: the SIMILAR scenario, in which the number of L1 categories equates with that of L2 categories ("one-to-one"); the SUBSET scenario, in which the number of L1 categories exceeds that of L2 categories ("many-to-one"), and the NEW scenario, in which the number of L1 categories falls short of that of L2 categories ("one-to-many"). These scenarios are associated with different levels of relative difficulty: SIMILAR (least difficult) < SUBSET (medium difficulty) < NEW (most difficult). The present thesis thus conducts three separate case studies (Studies 1, 2, and 3), each of which corresponds to one of the proposed types of learning scenarios. This would allow testing not only the model's predictions per scenario but also the relative levels of difficulty across scenarios. Specifically, Study 1 investigates Japanese listeners' perception of American English (AmE) high front vowels, which are perceived as SIMILAR to Japanese high front vowels. Study 2 investigates naïve Australian English (AusE) listeners' perception of Japanese high front vowels, which constitute a SUBSET of their native high front vowels. Study 3 investigates Japanese listeners' perception of a NEW vowel in AmE that does not have an equivalent in Japanese.

The present thesis concerns the perception of vowels rather than other types of sounds such as consonants and prosody because the majority of previous studies under L2LP have investigated vowel perception. Out of several varieties of English, AmE was chosen as the target language for Japanese learners because it is widely used in formal English language education in Japan and therefore is most familiar to the learners. AusE is of interest because of its unique acoustic characteristics of the vowels. Also, Japanese is a popular foreign language in Australia, so the learning scenario is relevant to real AusE listeners.

1.3 Outline of the thesis

The overall structure of the present thesis is as follows. Chapters 2 and 3 are literature and theoretical reviews, Chapter 4 is empirical tests of the L2LP model, and Chapter 5 is a discussion of the empirical tests. Chapter 6 presents the conclusions.

Chapter 2 presents a literature review of the subject of the present thesis, namely cross-linguistic and L2 vowel perception across Japanese, AmE, and AusE. The general nature of L2 speech perception is reviewed first, followed by descriptions of the vowel systems of the languages of interest as well as previous cross-linguistic perception studies pertaining to these languages.

Chapter 3 presents a theoretical review of the aforementioned three models of L2 perception, namely SLM, PAM, and L2LP. A comparison is also made between the models to illustrate their commonalities and differences in theoretical principles. Detailed explanations of Stochastic OT and the GLA are also presented, which would be necessary for interpreting the results of the case studies.

Chapter 4 presents the three case studies (Studies 1, 2, and 3) corresponding to L2LP's three types of learning scenarios (SIMILAR, SUBSET, and NEW). In each study, computational simulations based on L2LP are first presented to provide specific predictions regarding the particular learning scenario. The predictions are then compared with the result of a perception experiment on real listeners in order to test whether and to what extent simulated and real perceptual patterns match. The predicted and attested levels of difficulty are also investigated within and across scenarios.

Chapter 5 presents a general discussion of the case studies. This includes a discussion of the overall results from both theoretical and pedagogical perspectives. The chapter also discusses potential limitations of the current L2LP model and how the limitations may be addressed in future research to extend the model, possibly beyond speech perception and toward speech production. Finally, Chapter 6 concludes the present thesis based on the general discussion.

Chapter 2: Literature review

2.1 Introduction

This chapter presents a concise review of the literature that is relevant to the subject of the present thesis, namely cross-linguistic and L2 vowel perception across Japanese, AmE, and AusE. The chapter covers three areas: the general nature of L2 speech perception (Section 2.2), vowel systems of the languages of interest (Section 2.3), and previous case studies on cross-linguistic vowel perception (Section 2.4). Section 2.2 explains the language-specific nature of speech perception and how the language-specificity of L1 perception affects L2 phonological acquisition. The possibility of a 'reverse' influence of the acquired L2 on L1 perception is also mentioned. Section 2.3 then describes the specific vowel systems of Japanese, AmE, and AusE. The section also explains which acoustic cues native listeners use and to what extent to identify their native vowels, which is expected to shape their cross-linguistic and L2 perception patterns. The subsequent Section 2.4 reviews previous cross-linguistic perception studies pertaining to the language combinations of interest, namely Japanese listeners' perception of English vowels and English listeners' perception of Japanese vowels. The relationship between native and cross-linguistic perceptual patterns is also discussed. Finally, Section 2.5 presents an overall summary of the chapter.

2.2 Nature of L2 speech perception

While speech perception is often assumed to be a general-auditory and universal capability, it is in fact a language-specific phenomenon, which underlies the difficulty of L2 perception. This section reviews the language-specificity of speech perception and how it relates to cross-linguistic and L2 perception.

2.2.1 Language-specificity of perception

A large body of empirical evidence suggests that speech perception is shaped by one's unique history of linguistic experience. Research on infant language development has demonstrated that infants' ability to discriminate speech segments undergoes a significant change during the first year of life (Kuhl, 2000, 2004; Werker & Yeung, 2005). Infants are initially equipped with general sensitivity to discriminate nearly all speech contrasts, including those that are not phonemic in their native language. The ability to discriminate nonnative speech sounds begins to decline by six months for vowels and eleven months for consonants, while the ability to discriminate native speech sounds is maintained or even enhanced (Mazuka et al., 2014). This pattern of development is called perceptual attunement or narrowing. For example, both Japanese- and English-learning infants initially discriminate [ɹ] and [l] in English equally well. However, Japanese-learning infants' sensitivity to the sounds gradually declines as these sounds belong to a single phonemic category /r/ ([ɾ]) in Japanese, while the English-learning infants' sensitivity remains as these sounds are contrastive in English. Thus, infants' perception is attuned to show selective sensitivity toward their native language.

The experience-based perceptual attuning in infancy is known to shape adult speech perception. Adult Japanese listeners' perception of English /r/ and /l/ provides a good example. In their seminal study, Miyawaki et al. (1975) compared native Japanese and AmE listeners' discrimination of synthetic "speech-like" /ra/ and /la/ stimuli, which varied in third formant (F3) frequencies (i.e., the single most important acoustic cue for distinguishing English /r/ and /l/) while the first and second formants (F1 and F2) were fixed. Discrimination tests of a comparable set of stimuli consisting of the isolated F3 components provided a "nonspeech" control. The result found poor discrimination of "speech-like" stimuli by Japanese participants, while AmE participants' discrimination was nearly categorical. However, performance on the "nonspeech" stimuli was virtually identical for Japanese and AmE participants. These seemingly contradictory results indicate that Japanese listeners are not auditorily 'deaf' to the F3 acoustic cue, but their perceptual system 'filtered out' the cue as irrelevant. Therefore, adult speech perception is considered to be language-specific and different from general auditory processing.

Another piece of evidence for a difference between speech perception and general audition comes from Werker and Logan (1985). In the study, monolingual English listeners were tested on the Hindi /ʈ/-/t̪/ contrast using a same-different (AX) discrimination procedure with different inter-stimulus intervals (ISIs). Since retroflex and dental stops are not contrastive in English, native English listeners usually cannot discriminate them reliably. As expected, subjects tested with long (1500 ms) ISIs had difficulty in discriminating the sound categories. Interestingly, however, when another set of monolingual English subjects were tested with short (250 ms) ISIs, they were better able to discriminate the contrast. The result can be interpreted as follows. When stimuli are closely adjacent to each other, adult listeners can make use of fine-grained acoustic information available from short-term memory to differentiate unfamiliar nonnative sound contrasts that are not phonologically differentiated in their native language. However, a longer period of silence between stimuli forces them to rely on more abstract, language-specific phonological representations stored in long-term memory.

Such acoustic-phonological distinctions in perception have been attested at a neurological level as well. Jacquemot, Pallier, LeBihan, Dehaene, and Dupoux (2003) used a fast event-related functional magnetic resonance imaging (fMRI) paradigm to investigate French and Japanese participants’ acoustic and phonological perception. They used pseudoword stimuli, which differed either acoustically or phonologically in each of the two languages. For example, vowel length is phonemic in Japanese but not in French. Thus, the difference between e.g., *ebuzo* and *ebuuzo* would be “acoustic” in French but “phonological” in Japanese. In contrast, consonant clusters are phonologically illegal and thus ‘repaired’ by the process of vowel epenthesis in Japanese, but not in French. Therefore, the difference between e.g., *ebzo* and *ebuzo* would be “acoustic” in Japanese but “phonological” in French. fMRI scanning data revealed that two regions in the left hemisphere that have been associated with speech processing (superior temporal gyrus, STG, and supramarginal gyrus, SMG) were more activated when the stimuli changed phonologically than when they changed acoustically, for both Japanese and French participants. This provides further evidence for the distinction between general auditory (i.e., acoustic-phonetic) processing and speech (i.e., phonological) perception.

In sum, behavioral and neurological studies on infant and adult speech perception suggest that speech perception is language-specific in nature. Such language-specificity is considered to benefit native listening, which is remarkably efficient even in adverse (e.g., noisy) listening environments (Cutler, 2012). However, the L1-attuned perception can, in turn, hinder L2 speech perception. This is illustrated in the following section.

2.2.2 Language transfer

The language-specificity of speech perception, which facilitates L1 listening, can cause difficulties in L2 perception and its acquisition. It is well-known that adult Japanese listeners have considerable difficulty in distinguishing English words such as *write* and *light*, which results from their L1-attuned perception where /r/ and /l/ are not phonologically distinguished. Although adult Japanese listeners may be able to improve their perceptual accuracy for this sound contrast through intensive training (MacKain et al., 1981; Shinohara & Iverson, 2018), their perception may remain nonnative-like even after several years of instruction as they tend to rely on irrelevant acoustic cues such as F2 (Iverson et al., 2003). This aligns with Trubetzkoy's analogy of phonological "sieve," which is "not suited for the foreign language" and causes "numerous mistakes and misinterpretations." In other words, adult listeners "perceptual foreign accents" (Strange, 1995, p. 39).

In the field of second language acquisition (SLA), the influence of one's L1 on L2 acquisition is called *language transfer.* The term *interference* was also used in the past, although this term has mostly been displaced by the former to avoid the unwanted implication that knowledge of the L1 always hinders L2 acquisition. In fact, language transfer can be either positive or negative (Bardovi-Harlig & Sprouse, 2017). Positive transfer occurs when the influence of the L1 leads to the immediate or rapid acquisition or use of the L2. For example, native Japanese listeners may find it relatively easy to acquire nonnative length contrasts because Japanese has both vowel and consonant length compared to those whose L1 does not have a length contrast (Tsukada, 2012; Tsukada et al., 2018). On the other hand, negative transfer occurs when the influence of the L1 leads to errors in the acquisition or use of the L2. Japanese listeners' inability to discriminate the English /r/-/l/ contrast is a good example of negative transfer. However, researchers do not fully agree on to what extent the knowledge of the L1 transfers to an L2. Many believe that all the properties of the L1 are transferred at the onset of L2 acquisition, which is commonly referred to as *full transfer* (Schwartz & Sprouse, 1996). The learner assumes that the L2 is fundamentally similar to the L1, and their task subsequently is to replace L1 properties with appropriate L2 properties. Conversely, some believe that L2 learners start with "universals of language" (cf. Universal Grammar; Chomsky, 1965), and do not transfer L1 properties at the onset. Proponents of the *no transfer* hypothesis claim that the initial state of L2 acquisition is essentially the same as that of L1 acquisition. Others believe that there is transfer but it is limited, which can be termed *partial transfer*. In contemporary SLA, it appears that full transfer is widely accepted and also has the most empirical support (VanPatten & Benati, 2015).

While the L1 properties may transfer to the L2, recent research has also found that experience with an L2 affects the L1, which is sometimes called *backward* or *reverse transfer* (Cook, 2003). Language transfer, as explained above, can be termed *forward transfer* to differentiate it from backward transfer. According to Cook, backward transfer can be positive, negative, or neutral. Knowing another language can benefit the use of the L1, which can be seen as positive backward transfer. For example, research on bilingual language development generally shows that L2-using children have more precocious metalinguistic knowledge than their monolingual peers (Bialystok, 2001). Also, being bilingual has a positive effect on later-life cognition, including in those who acquired their L2 in adulthood (Bak et al., 2014). Conversely, the knowledge of the L2 can be harmful because it can cause language loss or attrition. For example, Ventureyra, Pallier, and Yoo (2004) examined L1 attrition in native Koreans who were adopted by French-speaking families and have stopped using their F1 for many years. They found that the adoptees did not perceive the differences between Korean lenis, fortis, and aspirated consonants better than naïve French controls, indicating that they had lost their L1-specific perceptual sensitivity. However, L1 attrition is not necessarily ‘negative’ but could be ‘neutral’ because losing one’s L1 indicates that the L2 user has blended in the L2 community successfully. Thus, the evaluations rely on a value judgment.

Concerning L2 phonology, there are cases in which the L1 sound system is not entirely lost but is affected to some extent by L2 experience. This is called *phonetic drift* (Kartushina et al., 2016), in which L1 sound categories can drift toward (assimilation) or away from (dissimilation) the closest L2 sound. For example, Chang (2012) examined novice English learners of Korean and found that their L1 production changed toward similar Korean sounds even after brief exposure to L2 Korean. Chang (2013) further found that the magnitude of phonetic drift was more pronounced for novice L2 learners than experienced ones, presumably due to a novelty effect. Deflection of L1 categories from monolingual norms (and from L2 categories) has also be reported in early Spanish learners of English, who produced Spanish /p, t, k/ with shorter voice onset time (VOT) than monolingual Spanish speakers to increase the phonetic contrast with English /p, t, k/ with long VOTs (Flege & Eefting, 1987). While L1 phonetic drift has been studied mainly in production, a few studies also report perceptual assimilation of L1 sound categories toward similar L2 ones (Lev-Ari & Peperkamp, 2013; Mora & Nadeu, 2012). Thus, the language-specific aspects of speech perception interact dynamically between L1 and L2, in which the L1 sound properties may affect the L2 and vice versa.

2.3 Vowel systems of interest

Given the language-specific nature of speech perception and how it transfers to another language, investigation of cross-linguistic and L2 perception would require a detailed description of the specific sound systems of interest. For the purpose of the current study, this section reviews the vowel systems of Japanese, AmE, and AusE.

2.3.1 Japanese

The Japanese vowel system consists of five distinct qualities /i, e, a, o, u/, which form five short (one-mora) and long (two-mora) pairs (Keating & Huffman, 1984; Nishi et al., 2008).[1] Vowel length is contrastive in all short-long pairs, as summarized in Table 2-1. The short vowels contrast in height and backness: /i/ is high front, /e/ is mid front, /a/ is low central, and /o/ is mid back. /u/ has traditionally been described as high back unrounded [ɯ], but it is closer to central rounded [ʉ] (Nogita et al., 2013). Roundedness, therefore, is correlated with backness: back vowels /o, u/ are rounded while /i, e, a/ are not. Long vowels are spectrally very similar to their short counterparts and are approximately two or three times longer in duration (Hirata, 2004). The present thesis transcribes Japanese long vowels with double letters (i.e., /ii, ee, aa, oo, uu/) because they

[1] The present thesis focuses on Tokyo Japanese because the majority of the Japanese participants were from the greater Tokyo area.

are phonologically considered to be a sequence of identical vowels, which are phonetically realized as [iː, eː, aː, oː, ʉː].

Table 2-1. Japanese vowel inventory with example words.[2]

Long			Short		
/ii/	*kii*	‘strange’	/i/	*ki*	‘tree’
/ee/	*meesi*	‘name card’	/e/	*mesi*	‘rice’
/aa/	*maai*	‘interval’	/a/	*mai*	‘dance’
/oo/	*koodo*	‘altitude’	/o/	*kodo*	‘radian’
/uu/	*Suusi*	‘numeral’	/u/	*susi*	‘sushi’

While Japanese long and short vowels share very similar spectral qualities, a few studies have reported slight yet systematic differences in formant frequencies between them. Hirata and Tsukada (2009) found that all long vowels except /uu/ occupied a more peripheral position in the vowel space than short vowels. Yazawa and Kondo (2019) found a similar displacement effect but for all five long-short pairs (Figure 2-1). The study also found that vowel duration was correlated with vowel height regardless of phonological length (/a, aa/ > /e, ee/ > /o, oo/ > /u, uu/ > /i, ii/), presumably because lower vowels require a larger degree of jaw opening and thus are more time-consuming (S. Kawahara et al., 2017). Thus, acoustically speaking, the quality and quantity of these vowels are not entirely independent of each other.

[2] /ei/ and /ou/ are phonologically neutralized to /ee/ and /oo/, respectively.

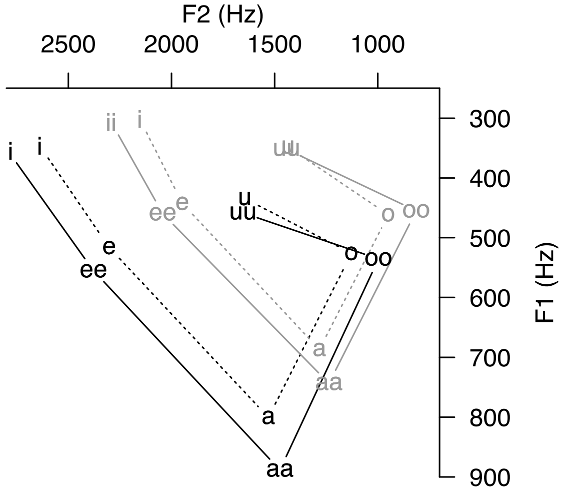


Figure 2-1. Average F1 and F2 of Japanese vowels (gray = male, black = female).

It is worth noting that native Japanese listeners do not seem to refer to vowel quality when perceptually judging vowel length (Arai et al., 1999). Since phonologically long vowels exhibit more peripheral qualities than short vowels and since vowel duration is systematically different at different heights (low > mid > high), in theory, listeners can use spectral information as a secondary cue for vowel length identification, but they seem not to do so. This implies that Japanese listeners have an invariant [+long] feature that is independent of spectral qualities and that the above spectral-temporal relations are caused by articulatory rather than phonological constraints. Therefore, it can be said that Japanese listeners use spectral and temporal acoustic cues independently in perception, where the former informs vowel quality (type) and the latter informs quantity (length) exclusively.

2.3.2 American English (AmE)

The AmE vowel system consists of nine monophthongs /iː, ɪ, ɛ, æ, ʌ, ɑ, ɔ, uː, ʊ/, three ‘true’ diphthongs /aɪ, aʊ, ɔɪ/, two ‘false’ diphthongs /eɪ, oʊ/, and one rhotic vowel /ɝ/, as summarized in Table 2-2 (Hillenbrand et al., 1995; Nishi et al., 2008; Peterson & Barney, 1952).[3] The AmE system is therefore much denser compared to that of Japanese (Figure 2-2). The monophthongs contrast in height, backness, and roundedness: the back vowels /ɔ, uː, ʊ/ are rounded while the others are unrounded. Among these, /iː, ɑ, ɔ, uː/ are tense and /ɪ, ɛ, æ, ʌ, ʊ/ are lax. In many dialects of AmE, /ɑ/ and /ɔ/ are neutralized to low back rounded [ɒ]. The high back vowels /uː, ʊ/ are both being fronted in many dialects of AmE as well (Fridland, 2008) and thus contrast mainly in vowel height. /ʌ/ is a conventional transcription and is more accurately near-low central [ɐ]. In some dialects, /æ/ is realized as a “tense æ” with a more fronted nucleus and centralizing offglide that could be rendered [eə] (Labov et al., 2006). While vowel length is not contrastive in AmE, there are systematic differences between the intrinsic duration of peripheral vowels /iː, æ, ɑ, ɔ, uː/ and their centralized counterparts /ɪ, ɛ, ʌ, ʊ/ (Peterson & Lehiste, 1960; Umeda, 1975).

Table 2-2. AmE vowel inventory with example words.

[3] Although the present thesis assumes ‘General American,’ it should be noted that there is a substantial regional variation in AmE vowel production (Clopper et al., 2005; R. A. Fox & Jacewicz, 2009; Hagiwara, 1997).

Monophthong				Diphthong		Rhotic	
Tense		Lax					
/iː/	*heed*	/ɪ/	*hid*	/aɪ/	*hide*	/ɝ/	*heard*
/ɑ/	*hod*	/ɛ/	*head*	/aʊ/	*how'd*		
/ɔ/	*hawed*	/æ/	*had*	/ɔɪ/	*hoyed*		
/uː/	*who'd*	/ʌ/	*hud*	/eɪ/	*hayed*		
		/ʊ/	*hood*	/oʊ/	*hoed*		

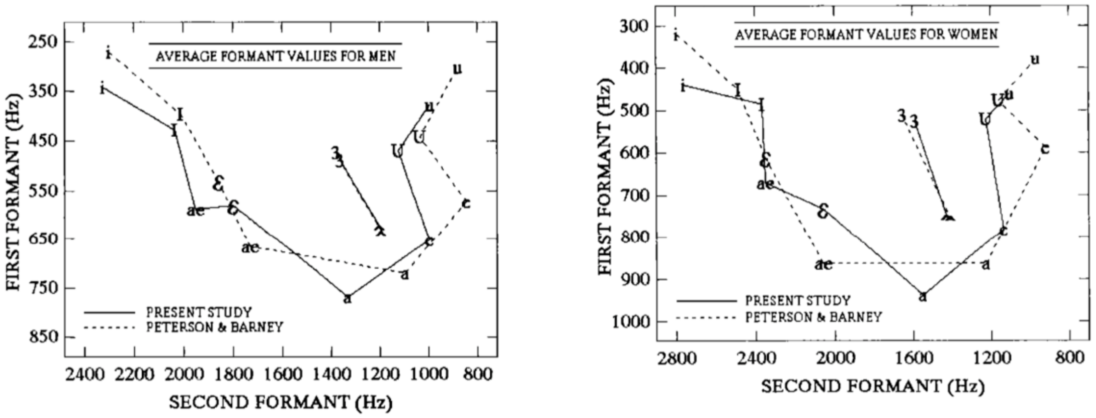


Figure 2-2. Average F1 and F2 of AmE vowels (Hillenbrand et al., 1995).

AmE listeners are known to rely primarily on spectral cues and use durational cues only to a limited extent in identifying their native vowels. Hillenbrand, Clark, and Houde (2000) examined the role of duration in AmE listeners' vowel perception by using synthesized /hVd/ stimuli. Fifteen phonetically trained subjects were tested on their identification of twelve AmE vowels (except true diphthongs /aɪ, aʊ, ɔɪ/) whose durations were manipulated to be original (matched to the original utterance), neutral (grand mean across all vowels), shortened (two standard deviations below the mean), or lengthened (two standard deviations above the mean). They found that duration had a small overall effect on vowel perception because the majority of vowels were identified correctly regardless of the durations. Vowel contrasts that differ consistently in duration such as /iː/-/ɪ/ and /uː/-/ʊ/ were minimally affected by duration. However, identification of other vowel contrasts such as /æ/-/ɛ/ and /ɑ/-/ɔ/-/ʌ/ were significantly affected by duration. While these results may seem contradictory, Hillenbrand et al. (2000) explained that non-duration-sensitive pairs such as /iː/-/ɪ/ and /uː/-/ʊ/ are quite distinct from one another based on their spectral qualities and, therefore, less dependent on duration for their distinction. In contrast, vowels such as /æ/-/ɛ/ and /ɑ/-/ɔ/-/ʌ/ show a greater degree of spectral overlap, thus resulting in a greater reliance on duration for their distinction. A statistical pattern classifier was used to test this hypothesis, which yielded a comparable result to the listener data. The result supports the idea that the role of duration in AmE vowel perception depends not only on the magnitude and consistency of observed durational differences among vowels but also on the degree to which vowels can be distinguished based on spectral cues. Therefore, spectral and temporal acoustic cues have complementary roles in AmE, where spectral cues are dominant and the relative significance of the duration cues depends on the informativity of the spectral cues.

2.3.3 Australian English (AusE)

The AusE vowel system consists of 12 monophthongs /iː, ɪ, eː, e, ɐː, ɐ, oː, ɔ, ʉː, ʊ, ɜː, æ/ and six diphthongs /ɪə, æɪ, ɑe, æɔ, əʉ, oɪ/ (Cox & Palethorpe, 2007; Harrington et al., 1997), as summarized in Table 2-3.[4] The monophthongs contrast in height, backness, and roundedness: the back vowels /oː, ɔ, ʉː/ are rounded while the others are unrounded. Length is also near-contrastive, where long vowels are approximately 1.5 times the length of their corresponding short vowels (Cox, 2006). The classification of monophthongs and diphthongs can be problematic for some vowels because they exhibit variable realizations. For example, /iː/ is phonologically a monophthong but is phonetically diphthongal as it typically shows onglide (i.e., [^{ə}iː]). /ʉː/ is sometimes realized with onglide, although it is more usually monophthongal. In contrast, /ɪə/ is phonologically considered a diphthong but is often monophthongized (i.e., [ɪː]), especially in closed syllables (Cox, 2006). AusE vowels are thus characterized by dynamic spectral qualities, which is sometimes referred to as vowel inherent spectral change (VISC). This can be seen in Figure 2-3, which shows the VISC of AusE monophthongs and diphthongs in various consonantal contexts produced by male (left) and female (right) speakers (Elvin et al., 2016).

[4] The present thesis focuses on Western Sydney English because all of the AusE participants were from the greater Sydney area.

Table 2-3. AusE vowel inventory with example words.

Monophthong				Diphthong	
Long		Short			
/iː/	*heed*	/ɪ/	*hid*	/ɪə/	*here'd*
/eː/	*haired*	/e/	*head*	/æɪ/	*hade*
/ɐː/	*hard*	/ɐ/	*hud*	/ɑe/	*hide*
/oː/	*horde*	/ɔ/	*hod*	/æɔ/	*how'd*
/ʉː/	*who'd*	/ʊ/	*hood*	/əʉ/	*hode*
/ɜː/	*heard*	/æ/	*had*	/oɪ/	*hoyd*

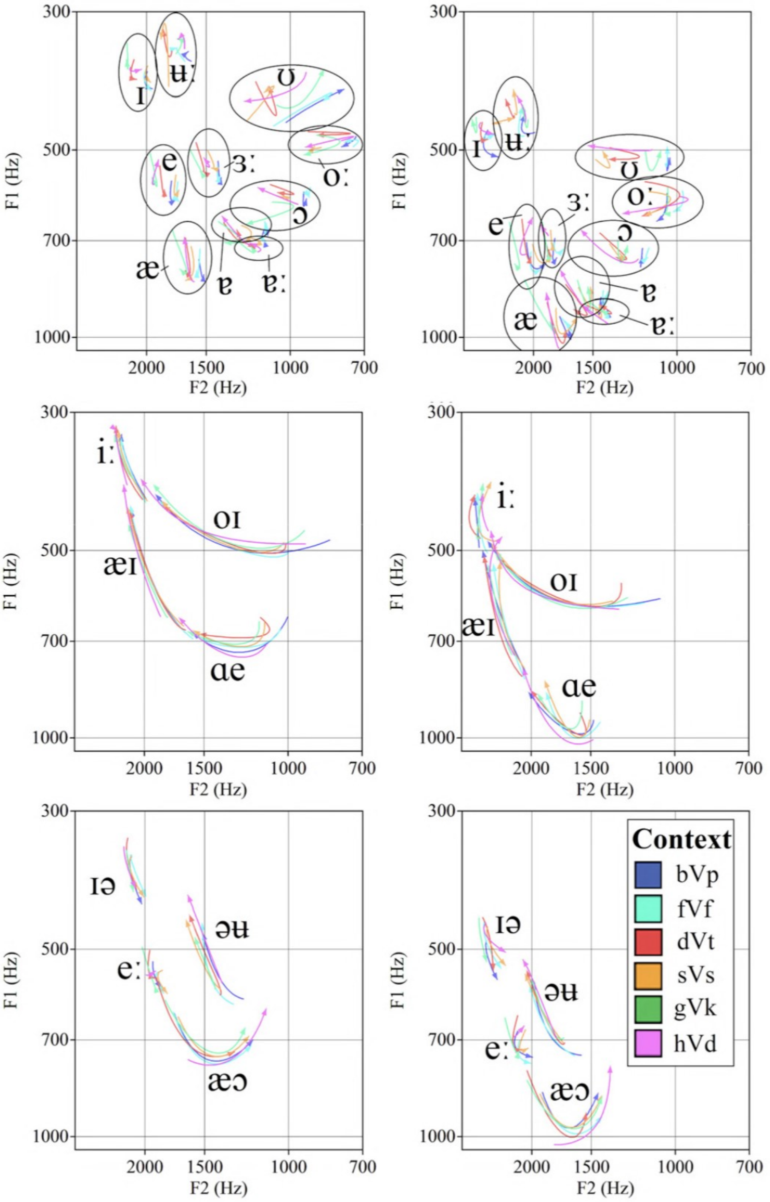


Figure 2-3. Average F1 and F2 trajectories of AusE vowels (Elvin et al., 2016).

Given the unique acoustic properties of AusE vowels, native AusE listeners are expected to utilize duration and VISC cues in vowel identification. To test this hypothesis, Williams, Esucudero, and Gafos (2018) investigated monolingual AusE listeners' perception of /iː, ɪə, ɪ/, which have almost identical mean or midpoint formants but differ in duration and VISC. Specifically, /iː/ is long and shows diverging VISC (i.e., onglide), /ɪə/ is long(er) and shows converging VISC (i.e., offglide), and /ɪ/ is short and shows very small, converging VISC. The study used synthetic vowel-like stimuli, which shared the same midpoint formant frequencies but varied in duration, trajectory direction (TD; diverging, converging, or zero) and trajectory length (TL). The result found that duration was by far the most important cue for distinguishing /iː/-/ɪ/ and /ɪə/-/ɪ/, suggesting that AusE listeners distinguish long /iː, ɪə/ from short /ɪ/ based on duration. As for the two VISC cues, TD was important for categorizing /iː/-/ɪ/ but not for /ɪə/-/ɪ/, reflecting that /iː/ has diverging VISC while /ɪə/ and /ɪ/ have converging VISC. TL was important for distinguishing both /iː/-/ɪ/ and /ɪə/-/ɪ/. The significant effect of TL for /ɪə/-/ɪ/ is worth noting because it indicates that listeners distinguish /ɪə/ from /ɪ/ not solely by duration but also by the magnitude of VISC. The overall result demonstrates that duration and VISC are indispensable acoustic cues for AusE /iː, ɪə, ɪ/, which may apply to other vowels in AusE.

2.4 Case studies of cross-linguistic vowel perception

The differences in vowel systems and perceptual cue usage between Japanese, AmE, and AusE, as seen above, are expected to shape the cross-linguistic and L2 perception patterns across these languages. This section reviews previous case studies on cross-linguistic vowel perception between Japanese and English.

2.4.1 Japanese listeners' perception of English vowels

Strange et al. (1998) conducted an extensive study on Japanese listeners' perception of AmE vowels, in which perceptual assimilation of 11 non-rhotic AmE vowels (/iː, ɪ, e^{ɪ}, ɛ, æ, ʌ, ɑ, ɔ, o^{ʊ}, uː, ʊ/) to ten Japanese vowel categories (/ii, i, ee, e, aa, a, oo, o, uu, u/) was tested. The AmE vowels were embedded in /hVbɑ/ disyllables ("citation" condition), and in a sentence *I say the hVb on the tape* ("sentence" context). The CVCV disyllable (rather than CVC or CV) was chosen for the citation condition so that it conformed to Japanese CV syllable structure as well as English phonotactics disallowing lax vowels in word-final open syllables. For the sentence condition, the word following the target CVC syllable was selected so that the sequence /hVbɑ/ was similar in both conditions, ignoring syllable and word boundaries. The vowels were produced by four male AmE speakers of Midwestern dialect who all maintained the /ɑ/-/ɔ/ distinction in their speech.

Twenty-four native Japanese listeners (11 male, 13 female, aged 18 – 23) participated in the perception experiment. Sixteen spoke the Kansai dialect, while the other eight were from other regions of Japan. All participants were undergraduate students who had received “standard” instruction in English, which consisted of six years of classes in secondary education and some English instruction in tertiary education. None of the listeners had spent an extended period of time in an English-speaking country. The participants categorized the /hV/ target syllable as most similar to one of 18 Japanese response alternatives by selecting one of 18 katakana characters displayed on the computer screen. The 18 characters represent /hV(V)/ syllables containing five one-mora vowels /ha, hi, hu, he, ho/, five two-mora vowels /haa, hii, huu, hee, hoo/, six palatalized CV(V) combinations /h^{j}a, h^{j}aa, h^{j}u, h^{j}uu, h^{j}o, h^{j}oo/, and two-mora vowel combinations /hei, hou/. After the response, the same stimulus was repeated, and the participant rated its “goodness” as an instance of the chosen response alternative on a scale from one (not Japanese-like) to seven (Japanese-like).

For each of the 11 AmE vowels, frequencies of the selection of each response category in percentage and overall median goodness ratings were computed over speakers and listeners. Table 2-4 shows the results in both citation and sentence conditions, where phonetically long /iː, e$^{\text{ɪ}}$, æ, ɑ, ɔ, o$^{\text{ʊ}}$, uː/ and short /ɪ, ɛ, ʌ, ʊ/ are presented separately.

Table 2-4. Assimilation of AmE to Japanese vowels in Strange et al. (1998).

	Citation			Sentence		
	Response	%	Goodness	Response	%	Goodness
/iː/	/i/	59	6	/ii/	83	6
/e^{ɪ}/	/ei/	65	5	/ei/	78	5
/æ/	/a/	31	2	/aa/	34	2
/ɑ/	/a/	79	6	/aa/	71	5
/ɔ/	/o/	31	3	/oo/	50	3
/o^{ʊ}/	/o/	54	5	/ou/	54	5
/uː/	/u/	61	5	/uu/	87	5
/ɪ/	/i/	58	3	/i/	77	4
/ɛ/	/e/	83	4	/e/	58	4
/ʌ/	/a/	64	4	/a/	65	4
/ʊ/	/u/	83	3	/u/	53	3

A comparison of the response categories in the citation and sentence conditions reveals that each of the 11 vowels tended to be assimilated to the same spectral category across conditions. The spectral assimilation patterns were predictable from cross-linguistic phonetic similarities to some extent. AmE long vowels /iː, e^{ɪ}, ɑ, o^{ʊ}, uː/ that are spectrally close to Japanese counterparts /i, e, a, o, u/ were assimilated with greater consistency and rated goodness than the short vowels /ɪ, ɛ, ʌ, ʊ/ and the other long vowels /æ, ɔ/. However, for the majority of the 11 vowels, assimilation patterns were more consistent in the sentence condition than in the citation condition, as shown by the percentages. This was due to the long vowels being more consistently assimilated to two-mora Japanese categories in the sentence condition, which indicates Japanese listeners' perceptual sensitivity to duration cues, but not in the citation condition.

To explain the differences in temporal assimilation patterns between conditions, Strange et al. (1998) speculated that the final /ɑ/ in the citation condition, which was stressed and thus quite long relative to the target vowels, might have led to a response bias for short vowels. Thus, they conducted a follow-up experiment in which the final /ɑ/ in the citation condition were truncated to sound more like the schwa /ə/ or left nontruncated as control. However, AmE long vowels were not assimilated to Japanese two-mora categories in either truncated or nontruncated conditions, suggesting that the differences in temporal assimilation patterns were not due to the duration of the immediately following vowel but to a broader prosodic context in the carrier sentence.

Acoustic analysis of the stimuli further indicated that some, but not all, of the variation in assimilation patterns could be accounted for by differences in specific phonetic realizations of the vowels. In particular, when listening to poor exemplars of Japanese categories such as /æ, ɔ/, listeners appeared to be responding based on differences in static and dynamic spectral properties. For example, some speakers produced /æ/ as "tense æ" with greater VISC (i.e., [eə]), which was in some cases interpreted by Japanese listeners as most similar to palatalized /h^{j}a/. Other speakers showed much smaller VISC, which were perceived as similar to e.g., /ha/ or /he/ depending on the specific phonetic realization of each vowel.

Following the above study, Strange, Akahane-Yamada, Kubo, Trent, and Nishi (2001) examined the same Japanese participants' perceptual assimilation of the 11 AmE vowels, but this time in six consonantal contexts /bVb, bVp, dVd, dVt, gVg, gVk/ in the sentence condition. The general procedure was the same as the previous study. While the overall assimilation patterns did not change, it was found that temporal assimilation patterns differed as a function of the voicing of the final consonant. Spectral assimilation patterns also varied with consonantal context and speakers. The result thus suggests that context-specific phonetic realizations determine cross-linguistic perception patterns.

Finally, a more recent study (Strange et al., 2011) revisited the same research topic. The participants were twenty-one adult native Japanese listeners of the Kansai dialect (aged 19 – 25) who had little English conversational exposure. Eight AmE monophthongal vowels /iː, ɪ, ɛ, æ, ʌ, ɑ, uː, ʊ/ embedded in [hVbə] disyllables, produced in citation form by a male AmE speaker of New York dialect, served as stimuli. The participants completed a categorical AXB discrimination test involving the eight vowels as well as a perceptual assimilation task with goodness ratings on a nine-point scale. Discrimination performance was compared with perceptual assimilation results to test whether assimilation patterns predicted discrimination accuracy.

Table 2-5. Assimilation of AmE to Japanese vowels in Strange et al. (2011).

	Response	%	Goodness
/iː/	/ii/	73	7
/ɪ/	/e/	97	7
/ɛ/	/a/	73	3
/æ/	/aa/	58	3.5
/ɑ/	/aa/	96	6
/ʌ/	/a/	91	6
/uː/	/uu/	57	4
/ʊ/	/u/	50	6

A slightly different assimilation pattern emerged (Table 2-5). Specifically, AmE /ɪ/ and /ɛ/, which were assimilated to Japanese /i/ and /e/ in the previous two studies, were categorized as Japanese /e/ and /a/, respectively. Given that the study used a native New York speaker instead of speakers of Midwestern dialect, these differences are likely due to dialect variations in the stimuli. However, the general tendency for AmE long /iː, æ, ɑ, uː/ and short /ɪ, ɛ, ʌ, ʊ/ to assimilate to Japanese two- and one-mora categories did not change, confirming Japanese listeners' sensitivity to duration. It was also found that assimilation patterns were highly predictive of discrimination accuracy.

In sum, the series of studies by Strange et al. suggest that Japanese listeners utilize both spectral and temporal cues in perceptually assimilating AmE vowels to Japanese vowels, as would be predicted from their native perception. However, assimilation patterns were shown to vary depending on the specific phonetic realizations of the target vowels, which are subject to factors such as consonantal context and dialectal variation.

2.4.2 English listeners' perception of Japanese vowels

The majority of studies on English listeners' perception of Japanese vowels have targeted AmE listeners. For example, Nishi, Strange, Akahane-Yamada, Kubo, and Trent-Brown (2008) conducted a 'reverse' study of Strange et al. (1998), which investigated AmE listeners' perception of Japanese vowels. For the stimulus materials, four adult male Tokyo Japanese speakers produced five long-short pairs of Japanese vowels /ii, i, ee, e, aa, a, oo, o, uu, u/ in nonsense /hVba/ disyllables for the citation condition and in a carrier sentence *kore wa /hVba/ desu ne* 'This is /hVba/, isn't it?' for the sentence condition. Twelve undergraduate students at the University of South Florida participated in the experiment (two male, ten female, mean age = 26.2). The majority of the listeners had lived in Florida for more than ten years, while a few had lived in the northeastern United States for more than ten years. All spoke only AmE fluently, and none of them had resided in a foreign country for an extended period of time. The participants first categorized each of the Japanese stimuli into 11 AmE vowels /iː, ɪ, e$^{\text{ɪ}}$, ɛ, æ, ʌ, ɑ, ɔ, o$^{\text{ʊ}}$, uː, ʊ/ by clicking one of the buttons with the International Phonetic Alphabet (IPA) symbols (participants were familiarized with IPA symbols prior to testing) and keywords in /hVd/ context (cf. Table 2-2). They then judged the category goodness of the stimulus vowel in the chosen AmE category on a seven-point scale from one (foreign) to seven (English).

Table 2-6. Assimilation of Japanese to AmE response vowels in Nishi et al. (2008).

	AmE response categories									
	/iː/	/ɪ/	/e^{ɪ}/	/ɛ/	/æ/	/ɑ-ɔ/[5]	/ʌ/	/o^{ʊ}/	/uː/	/ʊ/
Citation										
/ii/	99%									
/i/	95%									
/ee/			94%							
/e/		16%	76%							
/aa/						89%				
/a/						57%	39%			
/oo/								99%		
/o/								95%		
/uu/									92%	
/u/									91%	
Sentence										
/ii/	99%									
/i/	98%									
/ee/			97%							
/e/		23%	48%	28%						
/aa/						96%				
/a/						77%	21%			
/oo/								98%		
/o/								95%		
/uu/									96%	
/u/									89%	

Table 2-6 summarizes the result of the categorization patterns in both citation and sentence conditions, expressed as percentages of total responses summed over speakers and listeners. Most frequent response categories are underlined, and only responses above 10% are labeled. The overall median goodness ratings were generally high: six for /ii, i, uu, u, oo, o/ in both conditions and /ee/ in sentence condition, and five for /e, aa, a/ in both conditions and /ee/ in citation condition.

[5] These vowel categories were pooled in the study due to the prevalent /ɑ/-/ɔ/ merger.

It is evident from Table 2-6 that the majority of Japanese vowels, long or short, were consistently assimilated to tense AmE vowel categories. All Japanese long vowels /ii, ee, aa, oo, uu/ were consistently assimilated to tense AmE counterparts in both conditions (89 – 99%). Three out of five Japanese short vowels /i, u, o/ were also consistently assimilated to Japanese tense vowels in both conditions (89 – 98%). This suggests that AmE listeners disregarded the temporal differences between long and short Japanese vowels, perceiving them as equally good exemplars of AmE tense vowels. The observed perceptual patterns are thus comparable with their native perceptual cue usage, in which spectral cues are primary and duration cues play a marginal role.

The assimilation patterns of Japanese short /a/ and /o/ were not as straightforward. It was found that a few listeners consistently made non-modal responses for these vowels, which could be classified into either spectra-based or duration-based categorizations. For example, three listeners consisted of 83% of the /ɪ/ responses for /e/ in the citation condition, which is supposedly due to the spectral similarities between these vowels. In contrast, although /a/ in the sentence condition was spectrally most similar to AmE /ɑ-ɔ/, three listeners perceived it as most similar to AmE /ʌ/, suggesting an influence of duration. Thus, AmE listeners utilized duration as a secondary cue for identifying non-high vowels, which is again comparable to their native perception.

AusE listeners' perception of Japanese vowels is far less studied, though it is the focus of one of the case studies in the present thesis. Tsukada (2010) reported a preliminary result of vowel categorization experiments on AusE learners and non-learners of Japanese. The participants were asked to listen to the ten Japanese vowels /ii, i, ee, e, aa, a, oo, o, uu, u/ in monosyllabic words and categorize them into 12 AusE monophthongs /iː, ɪ, eː, e, ɐː, ɐ, oː, ɔ, ʉː, ʊ, ɜː, æ/ using 12 real words (*bead*, *bid*, *paired*, *bed*, *hard*, *bud*, *board*, *pod*, *booed*, *good*, *bird*, and *bad*, respectively). The learners and non-learners showed divergent patterns of perceptual assimilation. In general, non-learners tended to identify long Japanese vowels with short English vowels (e.g., Japanese /uu/ was identified as AusE /ʊ/ 47 % of the time). Learners, on the other hand, tended to assimilate short Japanese vowels to short English vowels (52 – 87%) and long Japanese vowels to long English vowels (57 – 75%). Thus, learners appeared to have developed sensitivity to Japanese vowel duration and learned to assign Japanese vowels to the appropriate phonological length. The result is consistent with the finding that AusE listeners utilize durational cues in identifying at least some of the native vowels. However, much remains to be known including how the static and dynamic spectral cues relate to their nonnative perception because Tsukada (2010) did not disclose the specific assimilation patterns.

2.5 Chapter summary

This chapter presented a brief and selective overview of the literature that is relevant to the topic of the present thesis. Section 2.2 presented empirical evidence for the language-specific nature of speech perception, together with how the language-specific properties may transfer from the L1 to the L2 (forward transfer) and vice versa (backward transfer). Section 2.3 described the vowel systems of Japanese, AmE, and AusE, including native listeners' perceptual cue usage for vowel identity in each of these languages. Section 2.4 then presented previous case studies on cross-linguistic perception across these languages, in which a firm relationship was found between native and cross-linguistic cue usage.

While the previous perception studies reviewed in Section 2.4 were quite extensive, it should be noted that they focused mainly on cross-linguistic rather than L2 perception. Thus, it remains to be investigated whether and how the observed perceptual patterns may change as a result of L2 experience, which the present thesis aims to explore. Models of L2 perception are useful in making predictions regarding the process of L2 perceptual acquisition, which is why a large number of L2 perception studies have adopted such models. The next chapter is dedicated to a theoretical review of L2 perception models, including the Second Language Linguistic Perception (L2LP) model.

Chapter 3: Models of L2 perception

3.1 Introduction

Models are a simplified representation of a system of interest, built for us to understand it from a particular perspective (Maria, 1997). In L2 phonology, various models have been proposed to explain the complex processes of L2 speech perception and to predict difficulties in perceptual acquisition. Among these, the Speech Learning Model (SLM; Flege, 1995) and the Perceptual Assimilation Model (PAM; Best, 1995), and its extension to L2 learning (PAM-L2; Best & Tyler, 2007) have been widely used in the literature. More recently, the Second Language Linguistic Perception (L2LP) model (Escudero, 2005; van Leussen & Escudero, 2015) has also been increasingly used, although not many studies have fully utilized the model. The present thesis focuses primarily on L2LP, while the other two models are also consulted wherever necessary, as they help explain the complex nature of L2 perception acquisition from a different angle.

In this chapter, I first review each of the three L2 perception models per section, namely SLM in Section 3.2, PAM(-L2) in Section 3.3, and L2LP in Section 3.4. The models are then compared with each other in Section 3.5 to highlight their commonalities and differences in theoretical principles. Section 3.6 provides a summary of the chapter.

3.2 The Speech Learning Model (SLM)

The Speech Learning Model (SLM; Flege, 1995) aims to account for age-related limits on the ability to produce L2 sounds in a native-like manner. It is primarily concerned with the ultimate attainment of L2 pronunciation, so studies carried out within the framework usually focus on very experienced L2 learners who have used their L2 for many years, such as immigrant populations. According to the model, the cause of foreign accents resides in learners' inaccurate perception of L2 sounds. During L1 acquisition, speech perception becomes attuned to the contrastive phonetic elements in the language. As speakers become highly skilled at classifying various phonetic realizations of L1 sounds, their perceptual system works as a phonological "sieve" (Trubetzkoy, 1939, 1969) to filter out properties of L2 sounds that are not phonologically important in the L1. For example, Japanese speakers typically hear and pronounce English /θ/ as /s/ while Russian speakers as /t/, even though both languages share the same phonemes /s/ and /t/ (Weinberger, 1987). This language-specific perceptual insensitivity, which is considered to exacerbate with age (Flege, 1981), causes nonnative speakers to perceive and produce L2 sounds differently from native speakers. However, this is not to say that no L2 learning occurs. Language-specific perception and production patterns can be amended, at least to some extent, during naturalistic L2 learning.

The original version of SLM[6] consists of several postulates and hypotheses, as summarized in Table 3-1. The first postulate (P1) states that the mechanisms used in acquiring the L1 sound system remain adaptive over the lifespan and can be applied to L2 learning. The language-specific perceptual patterns, hypothetically specified in long-term memory representations called phonetic categories (P2), can be modified through naturalistic L2 learning. More specifically, the phonetic categories established during L1 acquisition gradually evolve to reflect the properties of all the L1 and L2 sounds encountered throughout the lifetime (P3). An important assumption here is that both L1 and L2 phonetic categories exist in a common phonological space. Contrary to the widespread view in which only influence of the L1 on the L2 (forward transfer) is assumed, SLM considers cross-linguistic influences as bidirectional and also assumes an influence of the L2 on the L1 (backward transfer). The model thus predicts that bilinguals strive to maintain sufficient auditory contrast between L1 and L2 phonetic categories (P4), which may make both of their L1 and L2 categories different from those of monolinguals. Flege (1995) claims that P4 is consistent with Grosjean's (1989) view that a bilingual is not two monolinguals in one person but is instead a unique and specific speaker-hearer.

[6] A revised version of the model ("SLM-r") is in preparation according to my personal communication with Flege (July 21, 2018, at Sophia University).

Table 3-1. Postulates and hypotheses in SLM (Flege, 1995).

Postulates

P1 The mechanisms and processes used in learning the L1 sound system, including category formation, remain intact over the life span, and can be applied to L2 learning.

P2 Language-specific aspects of speech sounds are specified in long-term memory representations called phonetic categories.

P3 Phonetic categories established in childhood for L1 sounds evolve over the life span to reflect the properties of all L1 or L2 phones identified as a realization of each category.

P4 Bilinguals strive to maintain contrast between L1 and L2 phonetic categories, which exist in a common phonological space.

Hypotheses

H1 Sounds in the L1 and L2 are related perceptually to one another at a position-sensitive allophonic level, rather than at a more abstract phonemic level.

H2 A new phonetic category can be established for an L2 sound that differs phonetically from the closest L1 sound if bilinguals discern at least some of the phonetic differences between the L1 and L2 sounds.

H3 The greater the perceived phonetic dissimilarity between an L2 sound and the closest L1 sound, the more likely it is that phonetic differences between the sounds will be discerned.

H4 The likelihood of phonetic differences between L1 and L2 sounds, and between L2 sounds that are noncontrastive in the L1, being discerned decreases as AOL [(age of learning)] increases.

H5 Category formation for an L2 sound may be blocked by the mechanism of equivalence classification. When this happens, a single phonetic category will be used to process perceptually linked L1 and L2 sounds (diaphones). Eventually, the diaphones will resemble one another in production.

H6 The phonetic category established for L2 sounds by a bilingual may differ from a monolingual's if: 1) the bilingual's category is "deflected" away from an L1 category to maintain phonetic contrast between categories in a common L1-L2 phonological space; or 2) the bilingual's representation is based on different features, or feature weights, than a monolingual's.

H7 The production of a sound eventually corresponds to the properties represented in its phonetic category representation.

Based on the above four postulates, SLM proposes seven hypotheses (H1-H7) regarding the acquisition of L2 sound categories. First, the model hypothesizes that L2 learners perceptually relate L2 sounds that are closest to the L1 sounds at a position-sensitive allophonic level (H1). Studies have suggested that L2 learners have different degrees of difficulty in perceiving and producing certain allophones than others, depending on the position within a word or a syllable. For example, native Japanese speakers typically have difficulty in discriminating English /r/ and /l/, but Strange (1992) found that Japanese learners of English were more accurate at perceiving and producing them in the word-final than in the word-initial position. Takagi (1993) also reported that native Japanese speakers perceived English word-initial /r/ as Japanese /r/ ([ɾ]) while English word-final /r/ as Japanese /a/. These findings led Flege (1995) to consider that L1 and L2 sounds must be perceptually related at an allophonic rather than a more abstract phonemic level. Following Weinreich (1957), Flege (1995) calls the L1 and L2 sounds that are perceptually linked to each other *diaphones*.

However, not all L2 sounds end up being perceptually linked to an existing L1 category. According to H2 and H3, L2 learners may notice at least some of the phonetic differences between L1 and L2 sounds, especially when there is a substantial perceived phonetic dissimilarity between them. In this case, learners are expected to establish a new phonetic category for the L2 sound (Figure 3-1). H4 further states that the likelihood of phonetic differences being discerned (and thus, the likelihood of new category formation) decreases as AOL increases. Therefore, SLM attributes the well-attested negative effect of age on L2 acquisition to the learner's decreased phonetic sensitivity to L2 sounds. Flege (1995) argues that empirical evidence aligns better with this explanation than with the Critical Period Hypothesis (CPH), which asserts that native-like language acquisition becomes impossible after a certain age threshold such as puberty (Lenneberg, 1967; Patkowski, 1990; Penfield & Roberts, 1959). Specifically, he presents the results of his research (Flege et al., 1995), in which the relation between AOL and the degree of foreign accentedness were found to be quasi-linear, as evidence against the CPH. If L2 production ability were inhibited by neurological maturation as the CPH proposes, one would have seen a precipitous increase in the degree of foreign accentedness after a certain age, which was not attested in the study. Related to this, Flege (1995) also points out that foreign accents are apparently not inevitable. It has been reported that there are exceptionally successful late L2 learners whose L2 speech is indistinguishable from native speakers' (Bongaerts, 1999; Ioup et al., 1994), which also contradicts the CPH.

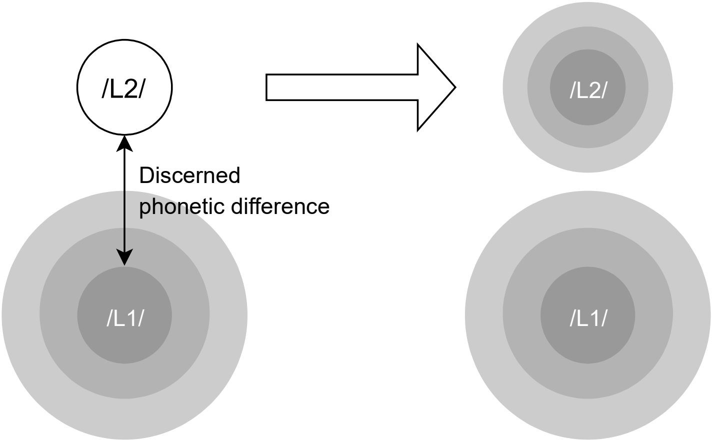


Figure 3-1. New category formation in SLM.

Due to an increased AOL as well as a small phonetic dissimilarity between L1 and L2 sounds, learners may fail to discern the cross-linguistic phonetic differences and perceive an L2 sound as a realization of an L1 category (*equivalence classification*). When this happens, the process of new category formation is blocked (H5), and a single phonetic category will be used to process the perceptually linked L1 and L2 sounds or diaphones (Figure 3-2). The diaphones are expected to eventually resemble one another, as they co-exist in a common phonological space. As stated in H7, the production of an L2 sound eventually reflects the properties of its phonetic category representation. Thus, bilinguals' production may be foreign-accented when a diaphone category is used, whereas native-like production may be achieved when a new phonetic category is formed. However, SLM considers two circumstances where a bilingual's newly established phonetic category may differ from a monolingual's (H6). First, since all phonetic categories exist in a common phonological space, a bilingual's L1 and L2 categories may disperse so as to maintain sufficient auditory contrast within the space. In such a case, a new category established for an L2 sound may be deflected away from an existing L1 category (which is analogous to historical sound change). Second, although L2 learners are capable of establishing new phonetic categories using L1-like learning mechanisms, they may do so by using different features or feature weights than a monolingual's. For example, Japanese learners of English can establish a categorical perception between English /r/ and /l/ through intensive perceptual training (MacKain et al., 1981). However, they may do so by using an irrelevant cue such as F2 that native English speakers hardly use (Iverson et al., 2003). In such a case, the L2 sound is not produced in precisely the same way as monolingual native speakers' production.

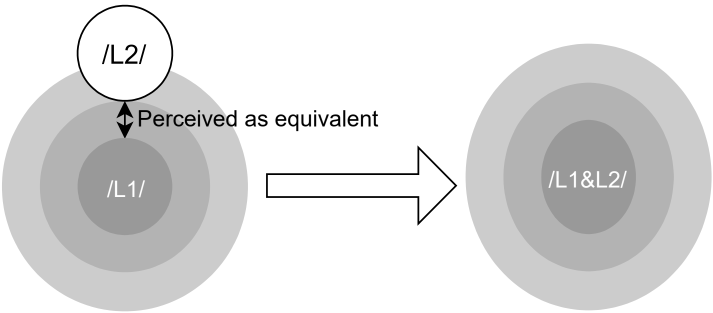


Figure 3-2. Diaphone category in SLM.

In sum, SLM claims that success in L2 speech acquisition depends largely on whether and how bilinguals discern the L1-L2 phonetic differences to establish a new phonetic category. However, it must be noted that an objective means for gauging the degree of perceived cross-language phonetic distance is yet to be defined (Flege, 1995, p. 264). Flege further notes that, in some instances, positionally defined allophones may be too coarse a unit of analysis, and smaller units such as features may be required. For example, he proposed that L2 features not used to signal a phonological contrast in the L1 will be difficult to perceive, and therefore learners will have difficulties producing the contrast based on this feature. McAllister, Flege, and Piske (2002) tested this “feature hypothesis” by investigating the perception and production of Swedish vowel quantity by native speakers of Estonian, AmE, and Spanish in Sweden. The three languages differ in their phonological status of vowel length; duration is highly informative in Estonian, only supplementary in AmE, and uninformative for segmental distinction in Spanish. The result found that the Estonian group performed much like native Swedish controls, whereas some English speakers and even more Spanish speakers differed from Swedish speakers. The result suggests that the role of the duration feature in L1 phonology is related to the learners’ success in acquiring the L2 quantity contrast, indicating a necessity of incorporating a fine-grained unit such as features into the model.

While there has been scarce research on the relevance of features on L2 speech acquisition, Flege (1995) states that several points can be made with some certainty. First, the features used to distinguish L1 sound contrasts can probably not be freely recombined to produce new L2 sounds. Flege and Port (1981) found that native Arabic speakers had difficulty in producing English /p/ (which is absent in Arabic) but not English /b, d, t, k/ (which are present in Arabic), suggesting that the ability to produce the [+labial] and [-voice] features respectively did not allow producing a new L2 sound /p/ comprising these features. Second, certain features may be more advantageous than others because of the nature of their acoustic specifications. Bohn (1995) found that native speakers of Spanish and Mandarin relied heavily on duration to distinguish English vowels /iː/-/ɪ/ and /ɛ/-/æ/ despite the lack of phonological vowel length in these languages, which suggests that the duration feature may be psychoacoustically more salient than spectral features. Finally, features may be evaluated differently as a function of position in the syllable and frequency of occurrence. For example, the [+ spread glottis] feature is consistently realized word-finally in French but not in English, leading French listeners to overuse it during the perception of word-final stops in English (Flege & Hillenbrand, 1987). Flege (1995) concludes that these questions, along with many others, must be answered in order to fully understand the nature of L2 perception and its contribution to L2 production.

3.3 The Perceptual Assimilation Model (PAM)

The Perceptual Assimilation Model (PAM; Best, 1995) is a cross-linguistic perception model that applies a direct realist approach to speech perception. The central premise of direct realism is that, in all aspects of perception, the perceiver directly apprehends the perceptual object itself and does not merely apprehend a representative or 'deputy' from which the existence of the object must be inferred. For speech perception, this means that the perceiver can directly detect distal articulatory gestures (e.g., "bilabial," "front," "fricative," "high") in the speech signal, which are not built up from the analysis of the acoustic waveform. The direct realist account of speech perception is in some respects similar to the motor theory of speech perception (Liberman et al., 1967; Liberman & Mattingly, 1985), which hypothesizes that listeners detect speakers' intended gestures through their innate knowledge of the vocal tract. However, the former differs from the latter in stating that the integrated perceptual systems to detect distal gestures gradually develop in reaction to the perceiver's ambient linguistic environment, rather than being innate. Languages differ in their selection of gestures or gestural constellations (Browman & Goldstein, 1992), and native perceivers of a language have attuned their perceptual systems through linguistic experience to detect the gestural constellations in the particular language effectively.

Under the PAM framework, nonnative speech segments are perceived according to their similarities to, and discrepancies from, the native gestural constellations that are closest to them in native phonological space. For example, native listeners of a language that has bilabial, alveolar, velar, but no dental stops may perceive the dental stop as similar to the alveolar stop because of their gestural proximity in constriction location. Similarities between nonnative speech segments and native gestural constellations are expected to determine listeners' perceptual assimilation of nonnative sounds into native ones. Specifically, a nonnative sound could be perceived as a good exemplar of a native category, an acceptable but not ideal exemplar of a native category, or a notably deviant exemplar of a native category. However, listeners are also expected to detect discrepancies from the native categories as well, especially when the discrepancies are large. In such a case, nonnative sound may not be categorized as a clear exemplar of any particular native category (i.e., it falls within native phonological space but between specific native categories). In an extreme case, a nonnative sound may not even be recognized as having speech-like properties, but instead be heard as some sort of non-speech sound (i.e., it falls outside native phonological space). Based on these premises, PAM proposes several possible perceptual assimilation patterns of nonnative sound contrasts, which are summarized in Figure 3-3 and Table 3-2 below.

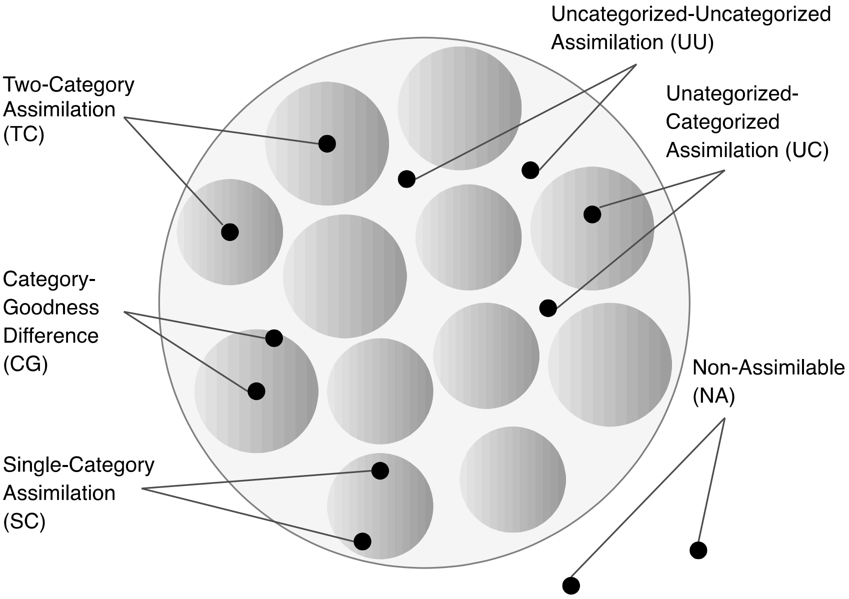

Uncategorized-Uncategorized Assimilation (UU)
Two-Category Assimilation (TC)
Unategorized-Categorized Assimilation (UC)
Category-Goodness Difference (CG)
Non-Assimilable (NA)
Single-Category Assimilation (SC)

Figure 3-3. Perceptual assimilation patterns in PAM.

Table 3-2. Assimilation patterns and discrimination difficulties in PAM (Best, 1995).

Assimilation pattern	Discrimination difficulty
Two-Category Assimilation (TC)	Excellent discrimination Each nonnative sound is assimilated to a different native category.
Category-Goodness Difference (CG)	Moderate to very good discrimination Both nonnative sounds are assimilated to the same native category, but they differ in the degree of discrepancy from the native "ideal."
Single-Category Assimilation (SC)	Poor discrimination Both nonnative sounds are assimilated to the same native category, but are equal in fit to the native "ideal."
Uncategorized-Uncategorized (UU)	Poor to very good discrimination Both nonnative sounds fall outside of native categories.
Uncategorized-Categorized (UC)	Very good discrimination One nonnative sound is assimilated to a native category, while the other falls outside of native categories.
Non-Assimilable (NA)	Good to very good discrimination Both nonnative sounds fall outside of speech domain and are heard as non-speech sounds.

Empirical evidence supports the proposed perceptual assimilation patterns. Best (1994) summarized her research on adult AmE listeners' perception of three Zulu contrasts: lateral fricative voicing distinction (/ɬ/-/ɮ/), velar voiceless aspirated vs. ejective stop distinction (/k/-/k'/), and voiced bilabial stop vs. implosive distinction (/b/-/ɓ/). The first contrast was expected to assimilate to two different categories in English (TC contrast), namely /ɬ/ as English /s, ʃ, θ/ and /ɮ/ as English /z, ʒ, ð/ or /l/, perhaps with a subsequent /l/ due to its /l/-like positioning of the tongue. The second contrast was expected to assimilate to English /k/ with a difference in the goodness of fit (CG contrast), in which the ejective /k'/ would be heard as a more deviant exemplar of English /k/ due to its nonnative glottal gesture. The third contrast was expected to assimilate to English /b/ equally well (SC contrast), although the glottal gesture of the implosive /ɓ/ is absent in English /b/ (and thus potentially a weak CG contrast). AXB discrimination tests and a post-test questionnaire gave support to the predictions. Participants discriminated the /ɬ/-/ɮ/ (TC) contrast with a near-ceiling level of performance, hearing the former as "s," "sh," or "thl" and the latter as "z," "zh," "zhl," or "l." They also discriminated the /k/-/k'/ (CG) contrast fairly easily, though not as well as the TC contrast, hearing /k'/ as a deviant ("choked" or "coughed") /k/. Finally, they had trouble discriminating the /b/-/ɓ/ (SC or weak CG) contrast, suggesting that both sounds were equally assimilated to English /b/.

The UU and UC assimilation patterns have also been investigated by Guion, Flege, Akahane-Yamada, and Pruitt (2000). The study ran a discrimination experiment on Japanese listeners' perception of English consonant contrasts: /r/-/l/ and /r/-/w/.[7] The hypothesis was that the former contrast would follow the UU assimilation pattern (both poor exemplars of Japanese /r/), and the latter contrast would follow the UC pattern (a poor exemplar of Japanese /r/ vs. a good exemplar of Japanese /w/). The results found poor discrimination of the /r/-/l/ (UU) contrast, while the /r/-/w/ (UC) contrast was discriminated with a moderate to high level of accuracy. Thus, PAM's predictions regarding the discriminability of UU and UC assimilation patterns were borne out. The remaining NA assimilation pattern was also examined in Best, McRoberts, and Sithole (1988). The study tested native AmE listeners' perception of a variety of Zulu non-nasalized clicks, which were expected to be non-assimilable as speech sounds to English listeners. The click contrasts were expected to be relatively easy to discriminate, despite the listeners' lack of prior exposure to click consonants in speech. The participants' discrimination performance was indeed excellent (80 to 95% correct), supporting PAM's prediction for the NA pattern.

[7] English /s/-/θ/ was also tested as a UC pattern, where /θ/ was hypothesized to be 'uncategorized.' However, this contrast may more accurately be a CG assimilation pattern, where English /θ/ is a more deviant exemplar of Japanese /s/.

PAM was initially developed to account for naïve listeners' perception of nonnative sound contrasts. Nevertheless, the model has been widely used in L2 research as well because it is "quite amenable to experience-dependent adjustments in adults' perception of previously unfamiliar contrasts" (Best, 1995, p. 198). However, PAM's predictions are limited to the discriminability of sound contrasts at the very onset of L2 learning, with no specific predictions as to how perceptual learning would proceed subsequently. Best and Tyler (2007) thus extended the principles of PAM to address the issues of L2 learning (PAM-L2). The aim of PAM-L2 was to reinterpret the perceptual assimilation patterns in PAM to predict the relative ease or difficulty of learning particular L2 sound contrasts. Importantly, the revised model emphasizes that perceptual assimilation occurs not only at the phonetic level but also at the phonological level. For example, French /r/ ([ʁ]) and English /r/ ([ɹ]) have little phonetic similarity. However, English learners of French nonetheless tend to equate the two sounds, presumably because French /r/ behaves very similarly to English /r/ in terms of syllable structure, phonotactic regularities, and phonological alternations (Ladefoged & Maddieson, 1996; Lindau, 1980). Thus, English learners of French can be said to assimilate French /r/ to English /r/ at the phonological level, but not at the phonetic level. The distinction between phonetic and phonological assimilations is an important revision to PAM.

To illustrate how PAM's framework can be extended to L2 learning, Best and Tyler (2007) outline the following four cases of acquiring L2 minimal contrasts:

(1) *Only one L2 phonological category is perceived as equivalent (perceptually assimilated) to a given L1 phonological category*. This scenario constitutes TC or UC assimilation patterns. The learner would have little difficulty in discriminating minimally contrasting words for these distinctions. Given that the L2 phone is perceived as a good exemplar of the L1 category, further perceptual learning is not very likely to occur for it or at least will be small in magnitude. Alternatively, it is also possible that the L2 phone is perceived as phonetically deviant from but phonologically equated with the L1 sound. For example, English learners of French may perceive the phonetic difference between French [ʁ] and English [ɹ], establishing two phonetic categories under the common L1-L2 phonological category /r/.

(2) *Both L2 phonological categories are perceived as equivalent to the same L1 phonological category, but one is perceived as being more deviant than the other.* This scenario constitutes a CG assimilation pattern. The learner is expected to be able to discriminate the L2 phones fairly easily, though not as well as TC or UC types. As the learner should be able to recognize the lexical-functional differences between the L2 phones, a new phonetic and phonological category is likely to be formed for the deviant

L2 phone, whereas no new category is likely to be learned for the better-fitting L2 phone. However, new category formation for the less deviant L2 phone is also possible in theory, of which likelihood depends on the degree of its perceived similarity to the L1 category.

(3) *Both L2 phonological categories are perceived as equivalent to the same L1 phonological category, but as equally good or poor instances of that category.* This scenario constitutes a SC assimilation pattern. The learner would initially have trouble discriminating the L2 phones, as they are phonetically and phonologically assimilated to the single L1 category. This would result in L2 minimal pairs contrasting in these sounds being perceived as homophones. Whether an L2 learner perceives the difference between the L2 phones depends on whether each phone is perceived as a better or poorer exemplar of the L1 phone, but perceptual learning is expected to be unlikely for most learners. However, the likelihood of perceptual learning may increase if many minimal pairs are contrasting in the L2 phones, as this would put more communicative pressure on the learner to perceptually learn the distinction.

(4) *No L1-L2 phonological assimilation.* This scenario constitutes an UU assimilation pattern. The learner does not perceive either of the L2 phones as belonging clearly to any L1 category, but rather as having a mixture of more modest similarities to multiple L1 categories. The difficulty of discriminating the L2 phones and the consequent

likelihood of new category formation depend on the similarities of the L2 phones to the existing L1 phones. If the uncategorized L2 phones are similar to different sets of L1 phones, they should be easily discriminated, and two new L2 phonological categories would be formed. However, if the uncategorized L2 phones are similar to the same set of L1 phones, then discrimination is expected to be difficult, and a single new phonological category encompassing the two L2 phones would be formed. This single category can theoretically split into different L2 categories, but it can also remain intact.

Finally, it remains unknown whether non-assimilable (NA) phones in the L2 that fall outside the L1 phonological space ever become integrated into the space as speech categories, as no study has empirically investigated this situation. Two possibilities can be considered. First, non-assimilable L2 sounds might eventually be incorporated into the L1 phonological space as uncategorized sounds, possibly resulting in one or two new categories being formed as in the UU pattern. Alternatively, learners may never incorporate the L2 sounds into their native phonological space and continue to ignore them as non-speech sounds in linguistic tasks such as word recognition. Best and Tyler (2007) state that the learning possibilities for NA phones suggest a particularly exciting line of future investigation on L2 perceptual learning, concluding that PAM-L2 raises a good range of empirical and theoretical issues to be investigated.

3.4 The Second Language Linguistic Perception (L2LP) model

The Second Language Linguistic Perception (L2LP) model aims to provide a formal linguistic account of L2 perceptual acquisition from the initial to end state (Escudero, 2005; van Leussen & Escudero, 2015). It grew out of and co-evolved with the Bidirectional Phonology and Phonetics (BiPhon) framework (Boersma, 1998, 2011), which itself is an extension of OT (Prince & Smolensky, 1993, 2004). The central tenet of L2LP is that speech perception is language-specific in nature (hence "Linguistic Perception"), as opposed to the traditional view of speech perception being extra-linguistic and general-auditory (Holt & Lotto, 2008; Hume & Johnson, 2001; Hyman, 2001). According to L2LP, listeners are equipped with a *perception grammar*, which is a formal linguistic grammar that maps the incoming acoustic signals onto abstract linguistic representations. Native listeners are *optimal perceivers*, whose perception grammars have been attuned to their L1 to process the sounds in the language efficiently. This language-specificity to facilitate native perception can, in turn, cause difficulty in nonnative perception, although listeners would attempt to achieve optimal perception in the L2 as well. L2LP concerns how L2 learners would undergo learning tasks to obtain the L2 optimal perception grammar in various learning scenarios, which are classified into the following three types: SIMILAR, SUBSET, and NEW.

A unique strength of L2LP is that the perception grammar and its acquisition can be implemented through computational simulations. Note that simulation is defined as the operation of a model, typically used for making predictions about a real system and for evaluating the model (Maria, 1997). Following BiPhon, L2LP utilizes two kinds of computational frameworks: Stochastic OT (Boersma, 1998, 1997), which is a probabilistic extension of OT, to model the perception grammar; the Gradual Learning Algorithm (GLA; Boersma & Hayes, 2001), which is an error-driven algorithm for learning optimal constraint rankings in Stochastic OT, to model the acquisition of the grammar.[8] The incorporation of computational simulations allows L2LP to make specific and detailed predictions as to how linguistic experience shapes one's perception, which can be compared with real listeners' perception to serve as a self-test of the model.

In what follows, I will first present Escudero's definition of speech perception in Section 3.4.1, followed by her proposal of modeling L1 and general speech perception as Linguistic Perception (LP) in Section 3.4.2. Section 3.4.3 then introduces the extension of LP to L2 acquisition, namely the L2LP model. Finally, Section 3.4.4 explains how L2LP can be computationally implemented under Stochastic OT and the GLA.

[8] A recent revision to L2LP (van Leussen & Escudero, 2015) employed a connectionist-inspired implementation of the two computational frameworks. However, the present thesis does not adopt this approach for reasons discussed later in Section 5.4.2.1.

3.4.1 Definition of speech perception

Escudero (2005, p. 7) defines speech perception as "the act by which listeners map continuous and variable speech onto linguistic targets." The listener's task is to decode the incoming variable speech signal onto discrete and abstract linguistic representations such as phonological features, segments, and prosody, to understand the message intended by the speaker. This is illustrated in Figure 3-4, in which an auditory continuum is mapped to discrete linguistic representations via the act of speech perception. For example, given an auditory continuum of F1, listeners need to map the variable acoustic values to a meaningful linguistic feature, e.g., vowel height (e.g., "low," "mid," "high"). Importantly, the mapping patterns and linguistic representations are language-specific; the boundary between "high" and "mid" differs from language to language, and a language such as Arabic may even lack a "mid" feature (Salameh & Abu-Melhim, 2014).

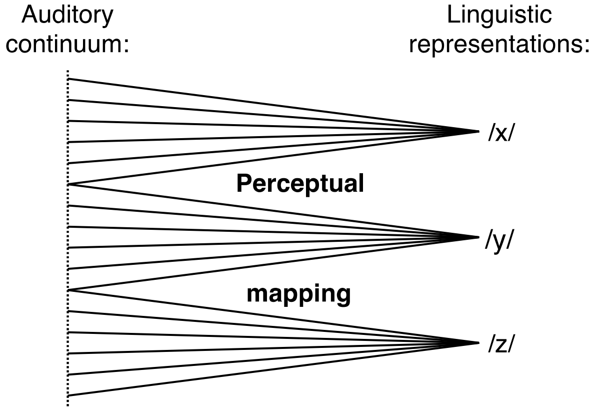


Figure 3-4. Perceptual mapping in L2LP.

The L2LP model considers speech perception as part of the whole process of speech comprehension, which involves multiple levels of representations. Figure 3-5 shows an overview of the levels of representations and connections between them, which is inspired by the BiPhon model (Boersma, 1998, 2011). The first representation at the bottom, the [auditory] form, refers to the incoming speech sounds as they arrive in the peripheral auditory system. The variable [auditory] form is then mapped to the following /surface/ form, which encodes the listener's language-specific and invariant representations of speech sounds, including context-specific allophonic details. The /surface/ form is connected to the third, |underlying| form, which encodes only contrasts that can change the meaning of a word. Finally, the |underlying| form connects to the <lexical> level where words and morphemes are stored.

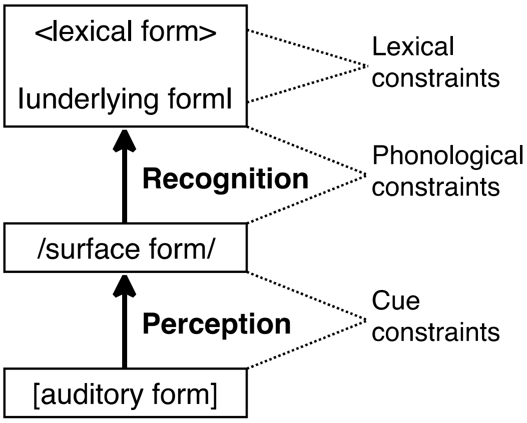


Figure 3-5. Levels of representations in L2LP.

As can be seen in Figure 3-5, L2LP makes a distinction between pre-lexical perception and lexical recognition, which is consistent with many psycholinguistic models of speech perception. However, it is still a matter of debate whether the two processes are sequential (i.e., bottom-up) or interactive (i.e., bottom-up and top-down). The original L2LP model (Escudero, 2005) held a sequential view, focusing primarily on the [auditory] to /surface/ mappings only. In this view, lexical influences on perception are explained as a result of offline (i.e., post hoc) learning rather than online (i.e., ad hoc) feedback from the lexicon (Norris et al., 2000, 2003). In contrast, the revised L2LP (van Leussen & Escudero, 2015) allows for testing the interactive view as well, in which higher-level (|underlying| and <lexical>) representations can influence lower-level ([auditory] and /surface/) representations during online perception (cf. the TRACE model; McClelland & Elman, 1986). However, following the original L2LP, and given a lack of convincing evidence for a direct influence of higher-level information on lower-level processes (Cutler, 2012, pp. 443–445), the present thesis adopts the sequential view. The thesis thus defines perception as the mapping of [auditory] forms to /surface/ forms, which is not affected by higher-level processes such as phonological operations (i.e., /surface/ → |underlying| mapping) and word recognition (i.e., |underlying| → <lexical> mapping).

3.4.2 Linguistic Perception (LP)

Escudero (2005) proposes that the language-specific mapping of [auditory] forms to /surface/ forms (henceforth "Linguistic Perception" or "LP") is handled by a formal *perception grammar*, which can be represented by Optimality-Theoretic negatively formulated constraints.[9] These are called *cue constraints* (cf. Figure 3-5) because listeners seek perceptual cues in the auditory continua (e.g., low F1 in vowels) to extract meaningful linguistic representations (e.g., /high/ feature). The simplest kind of cue constraints can be represented as in Figure 3-6, which maps a single auditory dimension onto a single kind of phonological feature: "a value of *x* on the auditory continuum *y* should not be perceived as the phonological feature *z*." Examples of such cue constraints are "a value of [300 Hz] on the auditory continuum F1 should not be perceived as the phonological feature /low/" and " [F2 = 2000 Hz] is not /back/."

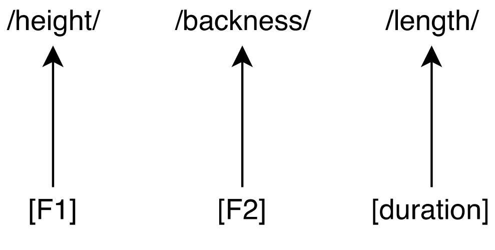


Figure 3-6. One-dimensional auditory-to-feature constraint.

[9] The reader is directed to McCarthy (2007) for a concise review of OT. See Boersma and Escudero (2008) for a discussion of why negatively formulated constraints rather than positively formulated rules are necessary for modeling perception.

However, one-dimensional constraints as in Figure 3-6 do not adequately explain speech perception because listeners are known to combine several auditory dimensions as perceptual cues to identify speech sounds. For example, F1 serves as an important perceptual cue not only for vowel height but also for the place of articulation and voicing of stops (Benkí, 2001; Lisker, 1999). Therefore, it is necessary to assume multi-dimensional constraints as in Figure 3-7: "a value of *x* on the auditory continuum *y* should not be perceived as the phonological feature *z*." Here, any value on any auditory continuum can in principle map to any phonological feature. In other words, the relationship between auditory dimensions and phonological features are arbitrary. Examples of cue constraints now include such ones as "[F1 = 300 Hz] is not /long/"and "[F2 = 2000 Hz] is not /short/," which seems rather odd. However, such constraints would be necessary to account for e.g., formant frequency values affecting the perception of phonological length in Swedish vowels (Behne et al., 1997).

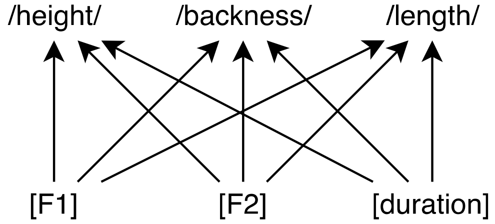


Figure 3-7. Multi-dimensional auditory-to-feature constraint.

Although the multi-dimensional auditory-to-feature constraints as in Figure 3-7 can express cue integration, phonological features alone are not abstract enough to model adult-like sound categorization, as adult listeners are known to integrate features into more abstract categories for efficient language use. Escudero and Boersma (2004) thus proposed multi-dimensional auditory-to-segment constraints as in Figure 3-8, which refer to highly arbitrary phonological categories such as vowels and consonants: "a value of *x* on the auditory continuum *y* should not be perceived as the phonological segment *z*." Examples of such constraints are "[F1 = 300 Hz] is not /e/," "[F2 = 2000 Hz] is not /a/" and "[duration = 120 ms] is not /i/." The relationship between auditory dimensions and phonological categories is again arbitrary, so any value on any auditory continuum can, in principle, map to any phonological category. Escudero (2005) argues that this type of multi-dimensional auditory-to-segment constraints can adequately explain adult speech perception.

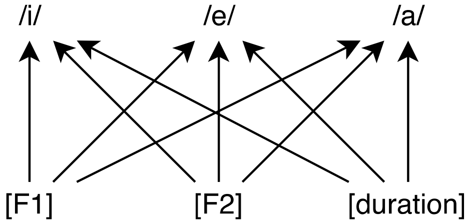


Figure 3-8. Multi-dimensional auditory-to-segment constraint.

Now that cue constraints are determined, the question then is how these constraints are ranked in the perception grammar to achieve language-specific perception. Escudero (2005) proposes that the ranking of the constraints depends on the acoustic distributions of the sounds in the listener's ambient production environment, which results in optimal perception for categorizing the sounds in the particular language (*optimal perception hypothesis*). Figure 3-9 illustrates this point. In this example, the F1 of the vowels /i/ and /e/ is distributed evenly around 300 Hz and 450 Hz. In order to maximize the possibilities of correctly perceiving these vowels, the constraints prohibiting the perception of /i/ should be ranked low when the acoustic value is likely to be that of /i/ (e.g., 300 Hz) and ranked high when the acoustic value is unlikely to be that of /i/ (e.g., 450 Hz). In contrast, the /e/-prohibiting constraints should have 'reversed' rankings, i.e., low when /e/ is likely and high when /e/ is unlikely.

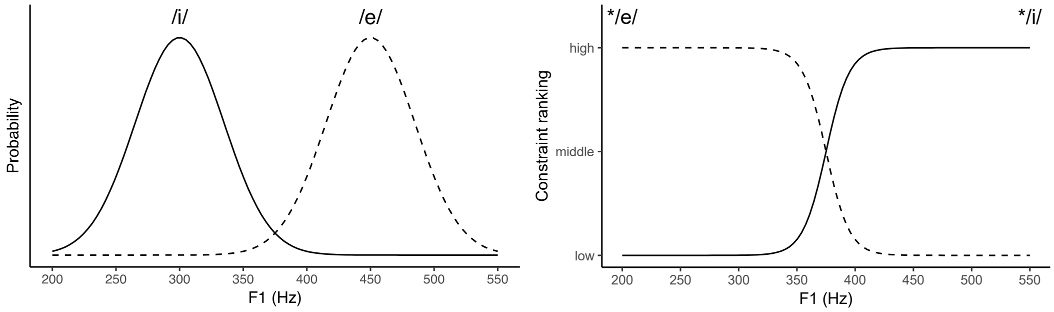


Figure 3-9. Acoustic distributions (left) and optimal rankings (right).

Such perceptual patterns can be formally represented in OT grammars as in Tableau 3-1 and Tableau 3-2. At the top of the left-most column is the auditory input, followed by candidates for the perceptual output. The ranking of the constraints determines which sound category is perceived. In Tableau 3-1, [F1 = 300 Hz] is perceived as /i/ because the constraint "[F1 = 300 Hz] is not /e/" is ranked higher than the constraint "[F1 = 300 Hz] is not /i/." Likewise, [F1 = 450 Hz] is perceived as /e/ in Tableau 3-2 because "[F1 = 450 Hz] is not /i/" is ranked higher than "[F1 = 450 Hz] is not /e/." LP can thus be represented by a number of cue constraints involving an auditory dimension (e.g., [F1 = 300 Hz], [F1 = 301 Hz], ... [F1 = 700 Hz]) and language-specific sound categories (e.g., /i/, /ɪ/, /e/) whose rankings are appropriate for the particular language.

Tableau 3-1. Perception of [F1 = 300 Hz] as /i/.

[F1=300 Hz]	[F1=300 Hz] not /e/	[F1=300 Hz] not /i/
☞ /i/		*
/e/	*!	

Tableau 3-2. Perception of [F1 = 450 Hz] as /e/.

[F1=450Hz]	[F1=450 Hz] not /i/	[F1=450 Hz] not /e/
/i/	*!	
☞ /e/		*

The above example focused on the mapping of only one auditory dimension, but constraint rankings should be defined across multiple auditory dimensions because listeners usually utilize more than one dimension during perception. Let us consider a case of Escudero and Boersma (2004), which examined the perception of /i/ and /ɪ/ by Scottish English (SE) and Southern British English (SBE) listeners. The two dialects differ in their relative use of acoustic dimensions that signal the vowel contrast. While the contrast is signaled mainly by the F1 in SE, both F1 and duration are used to distinguish them in SBE. This is shown in Figure 3-10, where darker color indicates more probability of /i/ perception. Consequently, a vowel with e.g., [F1 = 349 Hz, duration = 74 ms] (the diamond in Figure 3-10) is expected to be perceived as different vowels by SE and SBE listeners, namely /i/ by the former and /ɪ/ by the latter. In other words, listeners of the two dialects of English put different weights to the same acoustic dimensions (*cue weighting*).

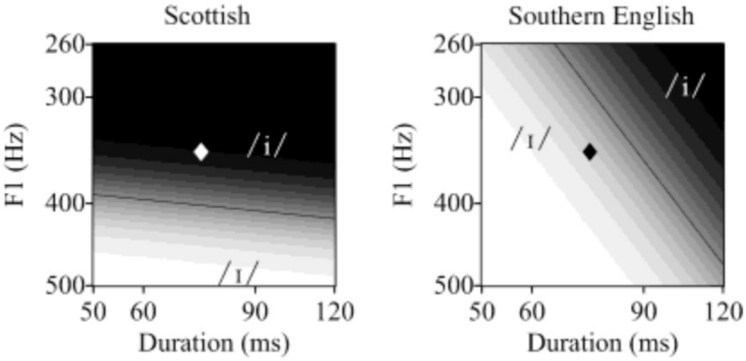


Figure 3-10. Perception of /i/ and /ɪ/ in SE and SBE (Escudero & Boersma, 2004).

Perceptual cue weighting across multiple auditory dimensions can be represented in OT grammars as in Tableau 3-3 and Tableau 3-4, which show the perception of [F1 = 349 Hz, duration = 74 ms] (i.e., the diamond in Figure 3-10) by SE and SBE listeners, respectively. In Tableau 3-3, the SE listener perceives the token as /i/ because the constraint "[F1 = 349 Hz] is not /ɪ/" is ranked the highest, i.e., an F1 of 349 Hz is too low for a vowel to be /ɪ/ in SE. The durational constraints are ranked lower because duration is a less informative cue in the dialect. In contrast, the SBE listener perceives the same token as /ɪ/ because the constraint rankings are different, as shown in Tableau 3.5. In the grammar, the highest-ranked constraint is "[duration = 74 ms] is not /ɪ/," reflecting the importance of duration for the vowel contrast in SBE.

Tableau 3-3. Perception of [F1 = 349 Hz, duration = 74 ms] by SE listener.

[F1=349 Hz, dur=74 ms]	[F1=349 Hz] not /ɪ/	[dur=74 ms] not /i/	[dur=74 ms] not /ɪ/	[F1=349 Hz] not /i/
☞ /i/		*		*
/ɪ/	*!		*	

Tableau 3-4. Perception of [F1 = 349 Hz, duration = 74 ms] by SBE listener.

[F1=349 Hz, dur=74 ms]	[dur=74 ms] not /i/	[F1=349 Hz] not /i/	[F1=349 Hz] not /ɪ/	[dur=74 ms] not /ɪ/
/i/	*!	*		
☞ /ɪ/			*	*

To summarize, speech perception is considered a language-specific phenomenon ("Linguistic Perception"), which is handled by a formal perception grammar as represented by Optimality-Theoretic cue constraints. Native listeners are optimal perceivers, whose perception grammar is optimized for perceiving sound contrasts in the particular language (optimal perception hypothesis). This language-specificity of LP may hinder adequate perception in another language because the sound system differs from language to language. The next section presents how the theoretical components of LP can be extended to L2 acquisition, namely the Second Language Linguistic Perception (L2LP) model, followed by its OT-based computational implementation in the subsequent section.

3.4.3 Second Language Linguistic Perception

The L2LP model is an extension of LP to L2 acquisition, which consists of five theoretical ingredients (Figure 3-11). In the figure, the straight arrows represent the sequential nature of the ingredients, and the curved arrows represent the relation between them. Before explaining each ingredient, it is important to note that the model strictly distinguishes perceptual mappings and sound representations, which are both language-specific. For example, Japanese and Spanish both have five vowels /i, e, a, o, u/, but the acoustic distributions of the vowels are not identical across the two languages, and neither are their perceptual mapping patterns (i.e., same representations but different mappings). Besides, some languages such as English have more categories than others, such as Arabic (i.e., different representations). Escudero (2005) argues that a strict separation of perceptual mappings and sound representations leads to an adequate comparison of L1 and L2 sound systems, which is crucial for modeling L2 speech acquisition.

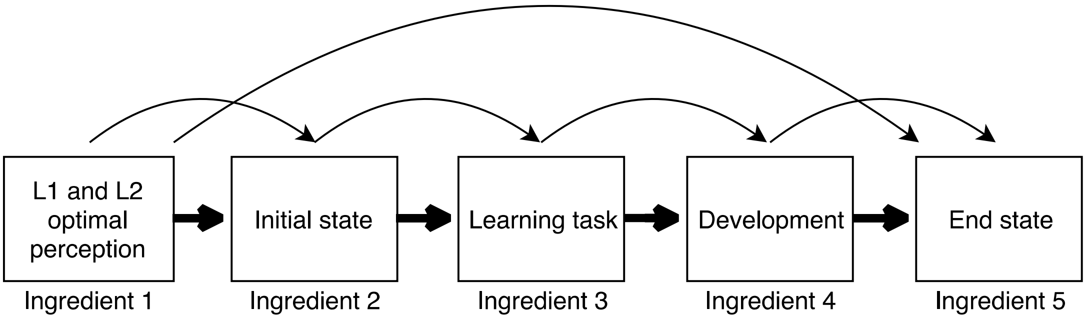


Figure 3-11. Five ingredients composing L2LP.

The first ingredient is optimal perception in the L1 and the target L2. Following the *optimal perception hypothesis*, L2LP defines L1 optimal perception as the best possible way of perceiving sound categories in the learner's L1. Likewise, the target L2 optimal perception is defined as the best possible way of perceiving sound categories in the learner's target language, which is predicted to be found in native listeners of the language. Note that optimal perception here involves both optimal perceptual mappings and optimal sound categories. The description of L1 optimal perception leads to the prediction of the initial state of L2 learning (Ingredient 2), i.e., L1-like perceptual behavior the learner would initially exhibit at the onset of L2 acquisition. On the other hand, the description of L2 optimal perception predicts the L2 end state (Ingredient 5), which the learner aims to attain ultimately. This is why the curved arrows connect Ingredient 1 with Ingredients 2 and 5. The mismatch between the initial and end states then determines what type of learning tasks (Ingredient 3) and development (Ingredient 4) the learner needs to undergo in order to arrive at L2 optimal perception.

The second ingredient is the L2 initial state. It is hypothesized that L2 learners transfer their L1 optimal perception to L2 perception at the very onset of L2 acquisition, i.e., the absolute beginner stage. This stage can be seen as nonnative rather than L2 perception, in which listeners perceive only L1 sound categories because L2 categories are yet to be formed. L2LP proposes that L1 transfer results in the formation of a *copy* or *duplicate* of their L1 perception grammar for perceiving L2 speech (*Full Copying* hypothesis). L2 sounds are equated with the copied L1 sound categories in the duplicated L2 perception grammar, which is likely to result in non-optimal perception because of the cross-linguistic mismatch in perceptual mappings (i.e., mismatch in categorical boundaries) and sound representations (i.e., mismatch in the number of sound categories) between the two languages.

The third ingredient is L2 learning tasks. Since the initial L2 grammar is typically not optimal for perceiving L2 sounds, learners need to bridge the gap between their initial state and the target L2 to attain L2 optimal perception. There are two types of learning tasks specified in L2LP: *perceptual* and *representational*. The perceptual task refers to adjusting and creating perceptual mappings, which usually involves redistribution or splitting of already-acquired L1 mappings in the duplicated L2 grammar. Also, L2LP considers another situation that involves an auditory dimension that has not previously been used in the learner's L1. For example, duration is a non-previously categorized dimension for native Spanish listeners because length is not phonologically contrastive in Spanish. In such a case, learners would need to create completely new mappings along the 'blank slate' or 'uncategorized' dimension. Perceptual tasks are considered *distributional* because learners need to utilize the acoustic distributions of L2 sounds in order to optimize their perceptual mappings. On the other hand, the representational task refers to changing the number of sound representations. For example, Spanish learners of Japanese would perceive Japanese long /ii/ and short /i/ as a single vowel representation /i/, which would result in their confusion of e.g., *biiru* 'beer' and *biru* 'building.' The learners' representational task, then, is to learn the semantic-lexical distinction between /ii/ and /i/ to create new sound representations such as "long /i/" and "short /i/" for optimal sound categorization in the L2. Representational tasks, therefore, are *meaning-driven*.

The fourth ingredient is L2 development. L2LP proposes that L2 learners have *Full Access* (Schwartz & Sprouse, 1996) to an L1-like learning device that enabled category formation and perceptual boundary adjustment in L1 acquisition. The model proposes that the device is available for L2 acquisition as well so that L2 learners gradually update their L2 perception grammar to become optimal perceivers in the L2. More specifically, the device creates new categories by splitting an existing category on an already-categorized dimension or by exploiting a non-previously-categorized dimension. The device also performs perceptual boundary shift, i.e., redistribution of existing categories. The GLA, which will be explained in Section 3.4.4, is a computational representation of the learning device.

The fifth ingredient is the L2 end state. Escudero (2005) claims that all L2 learners are capable of achieving L2 optimal perception if they are given appropriate kinds of perceptual input, such as enhanced acoustic cues and extensive listening experience. Although adult L2 learners are usually less successful than children in acquiring the L2, which L2LP attributes to cognitive plasticity, the model argues that the role of input overrules plasticity. The model also proposes that learners would maintain L1 optimal perception without being affected by the acquired L2 since the L2 grammar is a separate copy of the L1 grammar. This hypothesis of separate perception grammars may raise questions because L1 and L2 sound systems are known to interact with each other. However, L2LP explains such interactions as a result of the two grammars being activated at the same time. This notion 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, p. 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, and the instructions, which is expected to affect bilinguals' speech production and perception at any point in time.

To summarize, the L2LP model proposes that L2 learners start with a *Full Copy* of their L1 optimal perception grammar, which is typically not optimal for L2 perception. As L2 learners have *Full Access* to the L1-like learning mechanism, they will engage in perceptual and representational learning tasks to achieve L2 optimal perception. The model's separate perception grammars hypothesis combined with Grosjean's language mode hypothesis predicts that L2 learners will eventually attain L2 optimal perception while maintaining L1 optimal perception, which can be simultaneously activated to different extents depending on the given language context.

L2LP further predicts that there will be different kinds of mismatches between the L2 initial state and L2 optimal perception, resulting in different kinds of learning scenarios. Three types of scenarios are distinguished according to the model: SIMILAR, SUBSET, and NEW (Figure 3-12). First, the learner is faced with a SIMILAR scenario if the L1 perception grammar outputs the same number of sound categories as the L2 perception grammar (i.e., "one-to-one" relationship) because the L1 and L2 categories are perceived as equivalent. Second, if the L1 grammar outputs more categories than those required for optimally perceiving L2 sound categories (i.e., "many-to-one" relationship), the learner faces a SUBSET scenario because the L2 categories constitute a subset of the L1 categories. Finally, the NEW scenario refers to when the L1 grammar outputs fewer perceptual categories than required for optimal L2 perception (i.e., "one-to-many" relationship) because certain L2 sounds do not exist in the L1 and are therefore new. The NEW scenario comprises two types of sub-scenarios: the one involving already-acquired dimensions and the one involving non-previously-categorized (i.e., blank-slate) dimensions.

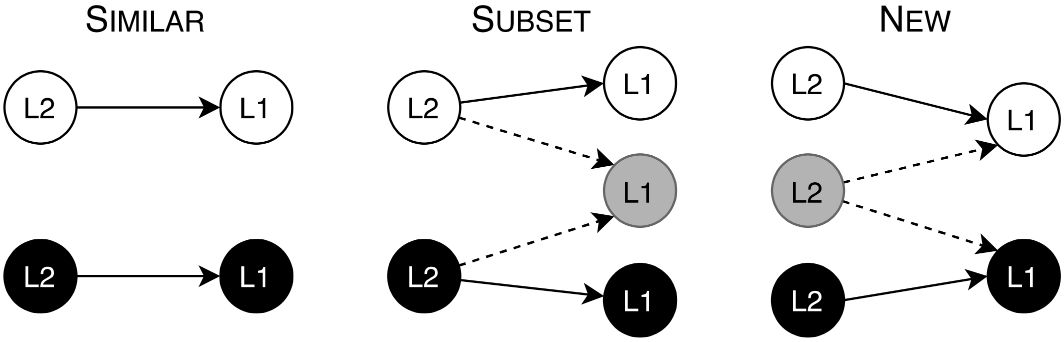


Figure 3-12. Three learning scenarios in L2LP.

The three types of scenarios are associated with a different number of tasks and different relative difficulties of acquisition, as summarized in Table 3-3. L2LP predicts that the number and nature of the learning tasks determine the relative difficulty. Specifically, the SIMILAR scenario is the least difficult because there is only one perceptual task of mapping adjustment, as there are already an appropriate number of sound representations in the L1 though their properties are not the same as the corresponding L2 representations. The SUBSET scenario is more difficult than the SIMILAR scenario because there is an additional representational task of reducing the number of lexical and perceived categories, since there are too many sound representations in the L1 for L2 optimal perception. The NEW scenario is expected to be the most difficult because it involves not only the creation of new mappings (perceptual task) and new categories (representational task) that are absent in the L1 but also the integration of non-previously categorized auditory dimensions. However, it should be noted that the exact level of difficulty of a NEW scenario may be reliant on whether the relevant auditory dimensions are already-categorized or non-previously categorized (i.e., the two sub-scenarios), which has not been tested yet within the model (Escudero, 2005, p. 317).

Table 3-3. Task and relative difficulty of L2LP learning scenarios.

	SIMILAR	SUBSET	NEW
Number of categories	L1 = L2	L1 > L2	L1 < L2
Perceptual task	Boundary shift	Boundary shift	Create mappings Cue integration
Representational task	None	Reduce categories	Create categories
Difficulty	Least difficult	Intermediate	Most difficult

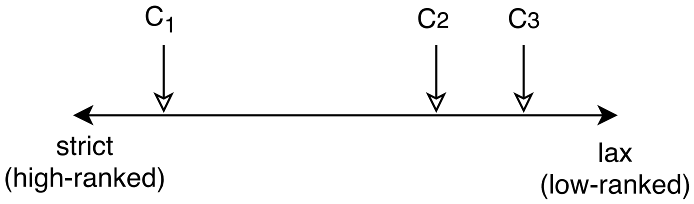
3.4.4 Stochastic Optimality Theory and the Gradual Learning Algorithm

This final section introduces Stochastic OT and the GLA, two computational frameworks associated with L2LP, to illustrate how the perception grammar and its acquisition in L2LP can be computationally implemented. Stochastic OT was proposed by Boersma (1998, 1997) as a probabilistic extension of OT (Prince & Smolensky, 1993, 2004). Stochastic OT differs from traditional OT in two ways. First, it assumes a continuous scale of constraint strictness rather than a set of discrete rankings. Second, its grammar is stochastic in that, at every time of evaluation, a small noise component is temporarily added to the ranking value of each constraint so that the grammar can produce variable outputs (Boersma & Hayes, 2001).

In traditional OT, constraints are ranked in a discrete and ordinal manner, e.g., $C_1 >> C_2 >> C_3$ (i.e., C_1 is stricter than C_2, and C_2 is stricter than C_3). Given the fixed ranking, the grammar will always choose the same candidate as the winner, i.e., there is no variation in the output. However, many linguistic phenomena are known to be gradient and variable in nature, which poses a challenge for traditional OT. Examples of such phenomena include optional phonological processes, free variation, and most importantly to the present thesis, speech perception. However, fixed constraint rankings in traditional OT cannot handle variation because they yield a single output given an input.

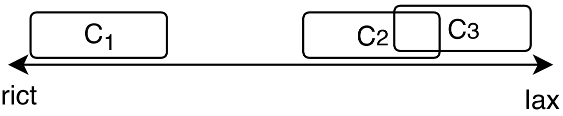
In Stochastic OT, on the other hand, constraints are ranked on a continuous scale, as illustrated in (3-1). Each constraint is assigned a continuous *ranking value* where higher values correspond to higher strictness.

(3-1) Constraint ranking along a continuous scale.



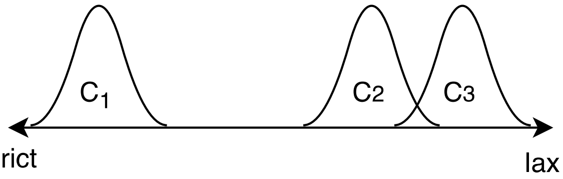
Constraint rankings are not only continuous but also stochastic in Stochastic OT. At each time of evaluation, the ranking values are temporarily perturbed by a random positive or negative value called *evaluation noise*. For example, a ranking value of 100.0 may become 100.8 at one time of evaluation and 99.6 at another. The temporary value used at evaluation time is called a *selection point*. The constraints are thus associated with ranges of values instead of single points, as illustrated in (3-2). Here, notice the overlap between the ranges of C_2 and C_3. This would most often result in C_2 outranking C_3, but if the selection point of C_2 is lower than that of C_3 at an evaluation time, then C_3 would outrank C_2, possibly changing the output of this particular evaluation. In this way, stochastic constraint rankings can yield multiple outputs for a single input.

(3-2) Constraint ranking with ranges.



Boersma (1998, 1997) further proposes that selection points follow the normal probability distribution, with the mean μ being the ranking value (e.g., 100) and the standard deviation σ being the evaluation noise (e.g., 2.0). This is because many noisy events in the real world occur with probabilities that are described with a normal distribution rather than being completely random. Therefore, selection points that are closer to the center are more probable to occur than those that are farther away, as illustrated in (3-3). As Boersma and Hayes (2001, p. 4) put it, by using probability distributions such as the normal distribution, "one can not only enumerate the set of outputs generated by a grammar but also make predictions about their relative frequencies."

(3-3) Constraint ranking in normal probability distributions.



For example, suppose that the constraints C_1, C_2, and C_3 in (3-3) have ranking values (i.e., μ) of 105.0, 98.0, and 95.0, respectively. If the evaluation noise σ is 2.0, then it can be calculated that C_2 would outrank C_3 with an approximate probability of 85.6% while the opposite ranking would occur approximately 14.4% of the time. In contrast, there is only a 0.7% chance of C_2 outranking C_1, and the chance of C_3 outranking C_1 is so small (0.02%) that it is virtually negligible. If the two distributions are dramatically far apart, the constraint 'reversing' seldom occurs, which essentially expresses an obligatory constraint ranking as assumed in traditional OT. Therefore, Stochastic OT is a flexible alternative to traditional OT that can represent both obligatory and variable constraint rankings. In fact, traditional OT can be seen as a special case of Stochastic OT with integer ranking values and zero evaluation noise.

I now turn to the GLA, which is an error-driven algorithm for learning optimal constraint rankings from the input in Stochastic OT. Teser and Smolensky (1998) devised an online learning algorithm called Error-Driven Constraint Demotion (EDCD) for traditional OT grammars, which changes the ranking order whenever the form produced by the learner is different from the correct form. The EDCD algorithm is fast and convergent, and sometimes leads to a significant change in the behavior of the grammar. However, EDCD is insufficient as a model of language acquisition because it is extremely sensitive to errors in the learning data and it does not show realistic gradual learning curves. For these reasons, Boersma (1997) proposed the GLA for Stochastic OT, which is, in some respects, a development of EDCD. The GLA and EDCD are similar in that they directly alter constraint rankings in response to the input data. They are also both error-driven; that is, constraint rankings are altered when the input data conflict with the current optimal output. However, the GLA differs from EDCD in that learning is moderate and gradual; the GLA executes only small perturbations to the constraints' ranking values rather than a complete reranking as EDCD applies. This allows the GLA to be robust to occasional errors in the input and to show gradual learning curves seen in real humans. More specifically, ranking values of constraints are adjusted by a small number called *plasticity*, which simulates the listener's neural or cognitive plasticity. Plasticity is set to gradually decrease over time, making learning fast but imprecise at an early stage (i.e., infancy) and slow but accurate at a later stage (i.e., adulthood). This plasticity scheme enables age-related modeling of language acquisition.

Below I demonstrate how perceptual acquisition of speech sounds can be computationally modeled with Stochastic OT and the GLA. At the initial state, the constraints begin with ranking values that are hypothesized by the modeler. In most of Boersma's works, all ranking values start at the same height of 100.0 (which I follow throughout the present thesis). The grammar is then presented with a learning datum, which includes the information of the [auditory] form (e.g., [F1 = 300 Hz]) and the intended /surface/ form (e.g., /i/). While the assumption that the listener has access to the intended form may be an idealization, it has been empirically shown that semantic feedback (i.e., an abstracted form) guides speech acquisition in both L1 (ter Schure et al., 2016) and L2 (Kriengwatana et al., 2016). The grammar then generates the output in the following way. For each constraint, a noise value (evaluation noise) is randomly drawn from the normal distribution and is added to the constraint's ranking value to obtain a selection point. The standard deviation of the normal distribution is usually 2.0 in Boersma's work (which I also follow throughout the thesis). Once a selection point has been picked for every constraint, the constraints are sorted in descending order of their selection points, which yields a strict constraint ranking for this particular evaluation. The remaining generation process follows the standard mechanisms of OT. If the form generated by the grammar is identical to the learning datum, no learning takes place. However, if the output does not match the intended form, the GLA notices the mismatch and proceeds to learning. Tableau 3-5 illustrates such a case in which the grammar incorrectly chooses /Candidate 1/ as the winner (marked with "☞") whereas the intended form is /Candidate 2/ (marked with "✓"). This situation can be interpreted as the listener noticing a mismatch between what they perceived and what the speaker must have intended through the use of their lexical knowledge and the semantic context (e.g., *sheep* /ʃip/ was perceived but context indicates *ship* /ʃɪp/). Alternatively, the learner may receive explicit feedback as to their perceptual mismatch from other speakers (e.g., L2 teachers in classroom settings).

Tableau 3-5. Mismatch between perceived form and intended form.

[auditory form] /surface form/	C_1	C_2	C_3	C_4	C_5	C_6
☞ /Candidate 1/		*	*			*
✓ /Candidate 2/	*!		*	*		

In such a situation, the GLA attempts to adjust the current grammar by raising the ranking values of all the constraints that would lead to the incorrect perception of /Candidate 1/ ("←" in Tableau 3-6) and by lowering the ranking values of all the constraints that would lead to the incorrect perception of /Candidate 2/ ("→") to increase the probability of correctly perceiving the same input in the next evaluation. The ranking values are adjusted by the current plasticity value (e.g., 1.0). Note that violations that match in the two candidates (e.g., C_3) are ignored since they make no difference to the outcome (*cancellation*). With repeated exposure to learning data, the grammar gradually learns to generate the correct /Candidate 1/ given the same [auditory] form (Tableau 3-7).

Tableau 3-6. Adjustment of ranking values by GLA.

[auditory form] /surface form/	C_1	C_2	C_3	C_4	C_5	C_6
☞ /Candidate 1/		←*	*			←*
✓ /Candidate 2/	*!→		*	*→		

Tableau 3-7. Adjusted grammar where perceived and intended forms match.

[auditory form] /surface form/	C_2	C_6	C_3	C_5	C_1	C_4
/Candidate 1/	*!	*	*			
☞✓ /Candidate 2/			*		*	*

To summarize, Stochastic OT, combined with the GLA, provides an ecologically valid proposal for simulating speech perception and its acquisition. The strength of Stochastic OT is that it can represent the categorical yet variable nature of perception. The GLA's error-driven adjustment of the perception grammar is also realistic because it approximates the gradual and meaning-driven perceptual learning in real listeners. Also, the decreasing plasticity scheme in the algorithm enables the modeling of the well-attested negative effect of age. Computational implementations of L2LP based on these two frameworks help make the model's predictions very explicit and specific, which can then be compared with real listeners as a self-test of the model.

3.5 Model comparison

Having outlined the three models of L2 perception, namely SLM, PAM(-L2), and L2LP, I now compare the theoretical principles of these models to illustrate their commonalities and differences. Shown in Table 3-4 is an overview of this comparison.

Table 3-4. Comparison of SLM, PAM(-L2), and L2LP.

	SLM	**PAM(-L2)**	**L2LP**
Target	Ultimate attainment of L2 pronunciation	Nonnative (and L2) speech perception	Entire L2 perceptual acquisition
Definition of perception	Phonetic categorization	Detection of articulatory gestures	Linguistic Perception (LP)
L2 acquisition mechanism	L1-like & accessible throughout lifetime	L1-like & accessible throughout lifetime	*Full GLA Access* throughout lifetime
Age factor	Phonetic sensitivity decreases with age; non-critical period	Age is not decisive per se; input is important	Cognitive plasticity decreases with age; input overrules
L1/L2 systems	Common	Common	Separate
Unit of analysis	Single category	Categorical contrast	Categorical contrast
L2 initial state	L1 phonetic categories	L1 phonetic/phonological categories	*Full Copying* of L1 perception grammar
L2 development	Category formation, category merging	Category formation, category merging	Category formation, boundary shift, category reduction
L2 end state	L1-L2 common space	L1-L2 common space	L1 & L2 optimal perception grammars

First of all, each of the three models was developed to achieve different kinds of purposes. SLM is primarily concerned with explaining the ultimate attainment of L2 pronunciation as a function of perception and age. The model states explicitly that foreign accents derive from learners' inaccurate perception of L2 sounds, which is subject to perceived cross-linguistic phonetic similarities and the learner's AOL. In contrast, PAM was initially developed to explain nonnative perceptual assimilation patterns by naïve listeners, i.e., those who have not started learning the language as L2 yet. PAM-L2 was then proposed to extend the principles of PAM to L2 learning, highlighting how the perceptual assimilations patterns relate to the difficulty of L2 perceptual acquisition. Conversely, L2LP aims to provide a comprehensive description, explanation, and prediction of L2 perceptual acquisition from the initial to end state. The model provides a formal linguistic account of L2 speech perception and its acquisition, based on the computational-phonological theory and the associated learning algorithm, namely Stochastic OT and the GLA.

The models also differ in their definitions of speech perception. SLM considers that language-specific aspects of speech sounds are specified in long-term memory representations called phonetic categories. Speech sounds are identified as a realization of each phonetic category at a position-sensitive allophonic level. L2 sounds are thus perceived according to the similarities to or differences from the closest L1 phonetic category. In contrast, PAM(-L2) proposes that the listener can directly detect the articulatory gestures of the speaker. Perceivers have not developed abstract categories in long-term memory as SLM proposes, but instead have become tuned for perceiving invariants in the speakers' vocal tract gestures. Nonnative sounds are perceptually assimilated to native gestural constellations to a different extent (or fall outside the speech domain) according to the similarities and discrepancies between them, which occurs at both phonetic and phonological levels. On the other hand, L2LP defines speech perception as LP, whereby the variable speech signal is mapped onto discrete and abstract linguistic representations. The mapping is handled by the listener's language-specific perception grammar as formally represented in OT. L2LP is the only model that strictly distinguishes between perceptual mappings and sound representations, which seem to be conflated in SLM (Escudero, 2005, p. 131) as well as PAM(-L2).

While the three models differ in their goals and premises, they share a common assumption that L2 perceptual learning is possible throughout the learner's lifetime. This is explicitly stated in SLM's P1: "The mechanisms and processes used in learning the L1 sound system, including category formation, remain intact over the life span, and can be applied to L2 learning." PAM's direct realist view of speech perception is also compatible with P1, according to Best and Tyler (2007). In the direct realist view of speech perception, listeners continue to refine their perception of speech throughout the lifespan even in their native language (e.g., dialect accommodation), and L2 learning is seen as a functional extension of this. Likewise, L2LP considers that L2 learners have *Full Access* to an L1-like learning device for their lifetime. The device is modeled by the GLA, which makes error-driven amendments to the L2 perception grammar through distributional and meaning-driven learning.

Related to the issue of learnability, all three models agree that age has an adverse effect on the development of L2 speech perception, but they offer different explanations of why it is so. SLM claims that the likelihood of discerning the phonetic differences between L1 and L2 sounds decreases as AOL increases. Consequently, those with increased AOL are more likely to reuse their existent L1 categories without creating new ones, resulting in nonnative-like perception and hence the production of the L2 sounds. Flege (1995) explicitly denies the CPH, which attributes age-related limitations to neurological maturation. PAM(-L2) agrees with SLM that the influence of age on L2 development is not due to biological or maturational reasons. Best and Tyler (2007) claim that the contributions of age to L2 perceptual learning occurs through interactions with the length of residence, relative usage of L1 and L2, and relative quantity and quality of input from native speakers. In their view, adult L2 learners are different from children in the quantity and quality of input they have received in their whole perceptual history, and therefore age per se is not a decisive factor. In contrast, L2LP does ascribe adults' difficulty in acquiring native-like competence to their reduced cognitive or neural plasticity. This notion is also present the GLA, in which the plasticity value that alters constraint ranking values is set to decrease over time gradually. The decreasing plasticity scheme is based on the observation that grammar appears to stabilize in adulthood, as non-lexical learning slows or halts (Boersma & Hayes, 2001). Crucially, however, Escudero (2005) argues that the role of input overweighs plasticity and that adult L2 learners can achieve optimal perception in the target language, provided that the learners have access to a sufficient amount of appropriate input. Thus, both PAM(-L2) and L2LP emphasize the importance of the role of input in L2 perceptual learning.

One of the most important theoretical differences between L2LP and the other two models is whether they consider L1 and L2 perceptual systems as shared or separate. SLM claims that both L1 and L2 phonetic categories exist in a common phonological space, in which cross-linguistic influence occurs bidirectionally. That is, L1 phonetic categories can influence L2 phonetic categories (forward transfer) and vice versa (backward transfer). This indicates that L1 phonetic categories can assimilate to or dissimilate from

L2 categories (phonetic drift), which, in the most extensive case, leads to L1 phonological attrition. PAM generally agrees with the notion of an L1-L2 common space, although it considers that L1-L2 interactions occur at both phonetic and phonological levels. A stark contrast to these models is L2LP, which hypothesizes that learners develop two utterly separate perception grammars for L1 and L2. Interactions between L1 and L2 sounds are explained as a result of parallel activation of the two separate grammars rather than a common space, following Grosjean's language mode hypothesis. Since the two grammars do not directly influence each other, no L1 phonetic drift or attrition is expected to be attested (Escudero 2005 p. 121).

The differences in theoretical principles among the models, including the above crucial difference between L2LP and the other two, result in different predictions regarding the development and outcome of L2 perceptual learning. Presented below is each model's predicted course of L2 learning and development, although a comparison across the models is not straightforward because SLM focuses on individual sounds (e.g., English /r/) while PAM(-L2) and L2LP focus on sound contrasts (e.g., English /r/-/l/). SLM claims that, at the initial state of L2 acquisition, an L2 sound is identified as a realization of, or perceived as different from, an L1 phonetic category. Category formation can occur for an L2 sound that is perceived as different from the closest L1 sound. However, this process may be blocked when the L2 sound is equated with an L1 phonetic category (equivalence classification), due to their phonetic similarities or the learners' increased AOL, or both. In such a case, the L1 and L2 categories will be linked as diaphones. The resultant end state of L2 learning is a common phonological space, in which L1 phonetic categories, newly formed L2 phonetic categories, and L1-L2 diaphone categories co-exist, continually influencing one another. Turning to PAM(-L2), L2 sound contrasts are perceptually assimilated to L1 sound contrasts (or fall outside of speech domain) at both phonetic and phonological levels at the initial state of L2 acquisition. There are a number of possible assimilation patterns, which determine the likelihood of new category formation. For example, as for the CG assimilation type, a new phonetic and phonological category is reasonably likely to be formed for the deviant L2 phone, whereas the better-fitting L2 phone may end up being assimilated to the L1 category. As for the TC assimilation type, no new category is likely to be formed for both L2 sounds unless they are assimilated at the phonological but not at the phonetic level. The end state of L2 learning is a common phonetic and phonological space, in which all L1 and L2 sounds can affect each other. Thus, SLM and PAM(-L2) are similar in their proposal that phonetic and phonological (dis)similarities between L1 and L2 sounds result in either formation or merging of sound categories. They also commonly propose that the end state of L2 learning is a shared space encompassing all the L1 and L2 sound categories encountered before, of which organization is ever-changing.

L2LP provides quite a different prediction from the above two models. The model states that the initial state of L2 learning is a copy of the L1 perception grammar (*Full Copying* hypothesis), through which L2 sounds are perceived. Three scenarios can be distinguished from this stage, depending on the numbers of L1 and L2 sound categories required for optimal perception in each language. First, the learner is faced with a SIMILAR scenario when the numbers of sound categories in L1 and L2 match. The learner has a perceptual task of adjusting categorical boundaries, whereas there is no representational task. This is the least difficult scenario of all. Second, the learner faces a SUBSET scenario when the number of L1 categories exceeds that of L2 categories. In addition to the perceptual task to adjust the categorical boundaries, the learner also has a representational task to remove the 'extraneous' vowel category, which is unnecessary for L2 optimal perception. The SUBSET scenario is thus more difficult than the SIMILAR scenario. It is worth noting that this scenario has not been considered in either SLM or PAM(-L2). Last, the NEW scenario occurs when the number of L1 categories falls short of that of L2 categories. The learner has to not only create new mappings and sound representations but also integrate non-previously-categorized auditory dimensions. The scenario is thus considered to be the most difficult of all. The end state of L2 learning is separate L1 and L2 optimal perception grammars (optimal perception hypothesis), which are independent of each other but can be simultaneously activated to a different degree depending on the current language mode.

Finally, a unique strength of L2LP is its incorporation of computational simulations. SLM and PAM(-L2) have been widely used in the field of L2 phonological acquisition for more than two decades, but they have also been a subject of criticism for the lack of concreteness in their predictions. This is because the predictions of the models are mostly dependent on perceived phonetic and phonological (dis)similarities between L1 and L2 sounds, which are difficult to define quantitively and objectively. Flege (1995, p. 264) indeed notes that there are no objective means for gauging the degree of perceived cross-language phonetic distance. Best and Tyler (2007, p. 26) also note that how listeners identify nonnative phones as equivalent to L1 phones "has not received adequate treatment in any model of nonnative or L2 speech perception." L2LP tackles this issue by adopting computational simulations based on Stochastic OT and the GLA, which not only makes its predictions specific but also serves as an objective self-test of the model.

3.6 Chapter summary

This chapter reviewed three models of L2 speech perception, namely SLM, PAM(-L2), and L2LP, all of which were designed to help explain the complex nature of L2 speech perception. While the models have different theoretical orientations, they share certain conceptual similarities. For example, they all assume that one's unique history of linguistic experience shapes their perception, that the relationship between L1 and L2 sounds results in different learning scenarios with different levels of difficulty, and that L2 learning is possible throughout the lifetime despite the adverse effect of age. However, the models also differ on a number of points. The most notable difference between L2LP and the other two models is that the former considers L1 and L2 perception grammars as separate, while the latter commonly assume a shared L1-L2 phonological space. L2LP is also the only model that adopts a strict distinction between perceptual mappings and sound representations, which seem to be conflated in the other two models. Furthermore, the incorporation of computational simulations in L2LP allows the model to make detailed and testable predictions regarding L2 perception, unlike the previous models. These differences make L2LP a unique model of L2 speech perception, which is worth further use and testing. The next chapter presents three case studies that were designed to test the predictions of L2LP for the proposed three types of learning scenarios.

Chapter 4: Empirical tests of L2LP

4.1 Introduction

This chapter presents three case studies (Study 1, Study 2, and Study 3), each of which corresponds to one of the three learning scenarios in L2LP (SIMILAR, SUBSET, and NEW), as part of a thorough empirical test of the model. The scenarios are in the order of proposed levels of difficulty: SIMILAR < SUBSET < NEW. While the primary focus is on L2LP, other models such as SLM and PAM(-L2) are also discussed whenever relevant.

Study 1 (Section 4.2) examines Japanese listeners' perception of L1 Japanese /ii/-/i/ and L2 AmE /iː/-/ɪ/. This follows a SIMILAR ("one-to-one") scenario, in which the L2 sound contrast is perceived as similar to the L1 contrast. The study focuses on how Japanese listeners' perceptual cue weighting may change as a function of language-specific perception modes, which serves as a test of L2LP's separate grammars hypothesis and the language mode hypothesis. The study is published in Second Language Research (Yazawa et al., 2019).

Study 2 (Section 4.3) examines monolingual AusE listeners' perception of native AusE /iː, ɪ, ɪə/ and nonnative Japanese /ii, i/. This follows a SUBSET ("many-to-one") scenario, in which the L2 categories constitute a subset of the L1 categories. The study focuses on how AusE listeners' cue usage in native vowel categorization is 'copied' to nonnative perception, which serves as a test of the *Full Copying* hypothesis in L2LP. A preliminary result of the experiment is available in Whang, Yawawa, and Escudero (2019).

Study 3 (Section 4.4) examines Japanese listeners' perception of L1 Japanese /e, a/ and L2 AmE /ɛ, æ, ʌ, ɑ/. This follows a NEW ("one-to-many") scenario, in which an L2 sound does not have an equivalent in the L1 and therefore is new to the learner. This is a NEW scenario with already-categorized auditory dimensions (F1 and F2), which is a sub-scenario of the NEW scenario that has not previously been investigated under L2LP. The study focuses on the process of new L2 category formation for AmE /æ/ by using segment- and feature-based implementations of L2LP.

Each case study is organized as follows. First, the Background section describes the learning scenario of interest in detail as well as L2LP's predictions regarding the particular scenario. The following Simulation section presents a computational implementation of L2LP based on Stochastic OT and the GLA in order to help make the model's predictions more specific and explicit. The Experiment section then presents a perceptual experiment on real listeners to empirically test the simulated predictions. The computational and experimental results are then discussed together in the Discussion section. Finally, the summary section presents an overall summary of the case study.

4.2 Study 1: SIMILAR scenario

4.2.1 Background

The first study investigates Japanese listeners' perception of high front vowels in L1 Japanese (i.e., /ii/-/i/) and L2 AmE (i.e., /iː/-/ɪ/) as a function of language-specific perception modes. This follows a SIMILAR scenario in L2LP, in which the L2 contrast is perceived as similar to the L1 contrast. While this type of learning scenario has been tested extensively, perceptual effects of language mode have not received much attention regarding L2 models despite the theoretical relevance. The current study thus investigates whether and how Japanese listeners' perceptual cue weighting is affected by language contexts, through the use of L2LP that incorporates Grosjean's language mode hypothesis.

In AmE, the tense /iː/ is more peripheral in the vowel space and also longer in duration than the lax /ɪ/ (Hillenbrand et al., 1995). Native AmE listeners are known to distinguish this contrast primarily by vowel spectra, with their perception "hardly affected at all by duration" (Hillenbrand et al., 2000, p. 3020). In other words, vowel length is a phonologically redundant feature for this contrast. However, studies have found that Japanese learners of English rely heavily on duration to distinguish /iː/ 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 /iː/ 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, 2001). 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 will learn to use spectral cues as they become proficient in L2 English. Morrison (2002) conducted a longitudinal study in which native Japanese and Spanish listeners were tested on Canadian English (CE) /iː/ and /ɪ/ (which share very similar acoustic properties with AmE /iː/ and /ɪ/), one and six month(s) 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 /iː/ 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.

L2LP's predictions for this particular learning scenario is as follows. First, optimal perception in L1 and L2 equates with primarily spectra- and duration-based perception, respectively (Ingredient 1). Japanese learners of English initially show predominantly duration-based perception because the initial state of L2 perception is a copy of the L1 perception grammar (Ingredient 2). The learners' perceptual task is to adjust this nonnative cue weighting, while there is no representational task because the number of sound representations in the L1 grammar matches that in the L2 grammar (Ingredient 3). With increased exposure to the L2 input, the learners' cue weighting in L2 perception gradually shifts from duration to spectra (Ingredient 4), ultimately resulting in native-like, spectral perception (Ingredient 5). On the other hand, their L1 perception will remain unaffected because a shift in cue weighting in the L1 grammar would result in an inaccurate perception of L1 sound contrasts, which the listeners would not favor. L2LP considers that the attainment of L2 optimal perception in a SIMILAR scenario is relatively easy because there is only one learning task, namely the perceptual task.

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 cues are used. As mentioned earlier, this notion of separate L1 and L2 perception grammars that can be simultaneously activated is based on the language mode hypothesis (Grosjean, 2001). More specifically, L2LP 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 (Figure 4-1). For the current learning scenario, 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.

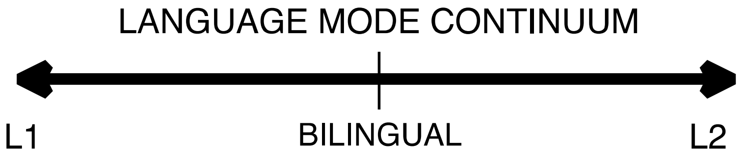


Figure 4-1. Language mode continuum in L2LP.

Studies suggest that language mode affects how speech sounds are perceived (Simonet, 2016). One of the earliest studies demonstrating such effects is Elman, Diehl, and Buchwald (1977), in which Spanish-English bilinguals were tested on a series of natural stimuli differing in 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, Diehl, and Champlin (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 a more sophisticated methodology to elicit different language modes also found similar effects (García-Sierra et al., 2012; Gonzales & 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 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 a 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 learners of Spanish perceived the same vowel tokens differently when they were told to classify 'Dutch' vowels into Dutch vowel categories (L1 mode), 'Spanish' vowels into Dutch categories (L1-L2 mode) and 'Spanish' vowels into Spanish categories (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.

The current study aims to investigate whether Japanese listeners' cue weighting (duration vs. spectra) for perceiving high front vowels is affected by the given language context (L1 Japanese or L2 AmE) to test L2LP's separate grammars hypothesis and the language mode hypothesis. In what follows, I will first present a computational implementation of L2LP to help make the theoretical predictions more specific and explicit (Section 4.2.2). A perception experiment is then presented in Section 4.2.3, which manipulates durational and spectral cues to investigate whether the reliance on either cue changes depending on their language modes ('Japanese' and 'English'). The computational and experimental results are then discussed, together with implications for other L2 models, in Section 4.2.4. Finally, Section 4.2.5 provides a summary of the study.

4.2.2 Simulation

In this section, I present a computational implementation of L2LP to simulate Japanese listeners' perceptual acquisition of L1 Japanese /ii/-/i/ and L2 AmE /iː/-/ɪ/ based on Stochastic OT and the GLA. In order to make the simulations as realistic as possible, detailed acoustic information of the target sounds was first collected. Table 4-1 shows the mean F1, F2, F2/F1 ratio, and duration of high front vowels in the two languages as produced by male native speakers. F2/F1 ratio was used to represent vowel tenseness in a single value (larger = tenser). The acoustic values for AmE were taken from Hillenbrand et al. (1995), in which 45 male monolingual AmE speakers pronounced a randomized list of isolated /hVd/ syllables (e.g., *heed* and *hid*), three times each. With the same procedures, 20 male native Japanese speakers' production of /hVda/ (i.e., / hiida/ and /hida/) in Japanese was recorded. 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. 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 from the table, the AmE vowels differ in both spectral and duration values, whereas the Japanese vowels differ predominantly in duration.

Table 4-1. Mean F1, F2, F2/F1, and duration of target vowels in AmE and Japanese.

Language	Vowel	F1 (Hz)	F2 (Hz)	F2/F1	Duration (ms)
AmE	/iː/	342	2322	6.79	243
AmE	/ɪ/	427	2034	4.76	192
Japanese	/ii/	294	2206	7.50	188
Japanese	/i/	302	2091	6.92	63

The vowels also differ in their frequency distributions. According to the CMU Pronouncing Dictionary, /ɪ/ appears approximately 1.5 times more frequently than /iː/ 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). It is thus assumed that short /i/ is nine times more frequent than long /ii/ in Japanese. The simulations make use of the mean acoustic values in Table 4-1 as well as the above frequency distributions to train the model.

In order to precisely model the perceptual space, a range of F2/F1 ratio from 4.65 (F1 = 430Hz and F2 = 2000Hz; most /ɪ/-like) to 6.76 (F1 = 340Hz and F2 = 2300Hz; most /iː/-like) and a range of duration from 70ms to 240ms were chosen. The spectral range was chosen so that a monolingual Japanese listener would perceive only /i/-like vowel qualities, while a monolingual AmE listener would hear the spectral difference between /iː/ and /ɪ/. The duration range was chosen to cover the entire durational variability of high front vowels in both languages. Each range was then divided into 21 logarithmically equal ‘bins’ (in $\log_2$, 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 /iː/, e.g., “[duration = 120 ms] is not /ii/ or /iː/”) and the other prohibiting the perception of the short or lax category (/i/ or /ɪ/, e.g., “[F2/F1 = 7.5] is not /i/ or /ɪ/”). Here, it is assumed that Japanese /ii/ and AmE /iː/, as well as Japanese /i/ and AmE /ɪ/, are representationally equal in Japanese listeners’ perception grammar. 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 listeners 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 & Nespoulous, 2008; Escudero & Wanrooij, 2010). Also, Japanese loanwords from English words containing /iː/ and /ɪ/ are usually transcribed with /ii/ and /i/ in Japanese orthography, respectively, which may further reinforce the connection. Therefore, the same set of 84 (2 auditory continua × 21 bins × 2 vowel categories) constraints were used to model the perception of high front vowels in both L1 Japanese and L2 AmE.

4.2.2.1 L1 Japanese perception

The procedure of the simulation of L1 Japanese perception is as follows. Initially, the virtual learner has a 'blank' perception grammar in which all 84 constraints have the same ranking values of 100.0. The evaluation noise is fixed at 2.0. 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 a token are randomly drawn from normal distributions, with the mean F2/F1 and duration being those in Table 4-1 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 the Japanese 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 GLA 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 × 0.7). The parameter settings for evaluation noise and plasticity are based on previous studies (Boersma & Escudero, 2008; Escudero & Boersma, 2004).

Figure 4-2 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 × 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. The duration-based perception is represented in Tableau 4-1, in which [F2/F1 = 4.65, duration = 240 ms] (i.e., bottom-right corner in Figure 4-2) is perceived as long /ii/.

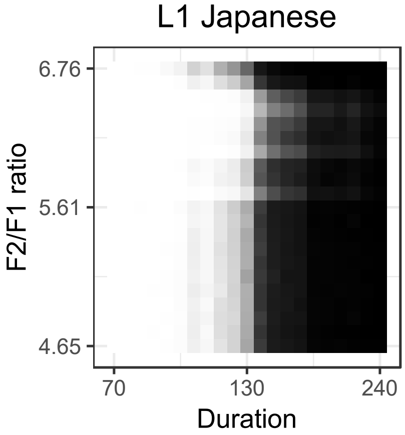


Figure 4-2. Model's perception of L1 Japanese /ii/ (black) and /i/ (white).

Tableau 4-1. Model's perception of [F2/F1 = 4.65, duration = 240 ms] as /ii/.

[F2/F1=4.65, dur=240ms]	[dur=240ms] not /i/	[F2/F1=4.65] not /ii/	[F2/F1=4.65] not /i/	[dur=240ms] not /ii/
☞ /ii/		*		*
/i/	*!		*	

4.2.2.2 L2 AmE perception

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/-/i/ and AmE /iː/-/ɪ/. That is, for example, the constraint "[duration = 240 ms] is not /ii/" in the L1 perception grammar was copied as "[duration = 240 ms] is not /iː/" to the L2 perception grammar with the same ranking value. In the same way as L1 acquisition, the learner received tokens of AmE /iː/ and /ɪ/ occurring randomly at different frequencies (40% and 60%, respectively). The acoustic values were randomly drawn from normal distributions with the means from Table 4-1. The standard deviations of the normal distributions were again 0.1 for F2/F1, but 0.8 for duration. The standard 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, 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. Thus, it was ensured 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. Evaluation noise was 2.0. For comparison, the perception of AmE /iː/ and /ɪ/ by a native AmE listener was also modeled with the same parameters (except that the plasticity was initialized to 1.0).

Figure 4-3 shows the outcome of the learner's acquisition of /iː/ 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 4-2. Darker shades indicate the likelihood of tense AmE /iː/ 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 in Figure 4-4, although the learner is slightly more likely to perceive tokens with short durations as /ɪ/. The acquired spectral perception is represented in Tableau 4-2, in which [F2/F1 = 4.65, duration = 240 ms] (which was perceived as Japanese long /ii/) is perceived as lax /ɪ/.

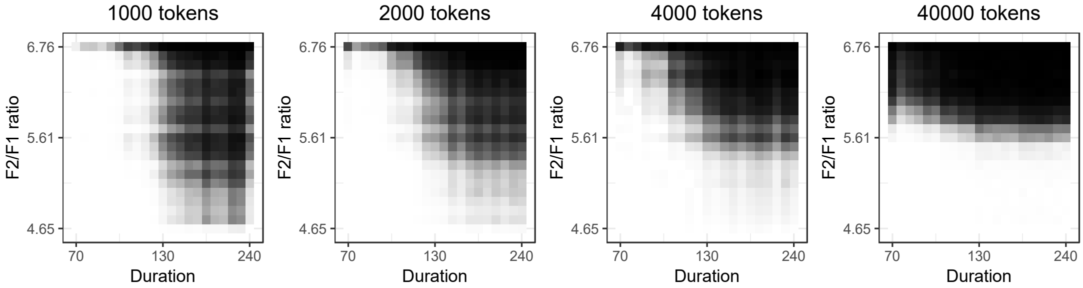


Figure 4-3. Virtual Japanese listener's perception of AmE /iː/ (black) and /ɪ/ (white).

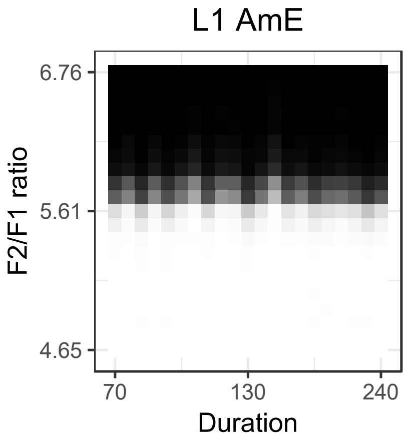


Figure 4-4. Model's perception of L1 AmE /iː/ (black) and /ɪ/ (white).

Tableau 4-2. Model's perception of [F2/F1 = 4.65, duration = 240 ms] as /ɪ/.

[F2/F1=4.65, dur=240ms]	[F2/F1=4.65] not /iː/	[dur=240ms] not /ɪ/	[dur=240ms] not /iː/	[F2/F1=4.65] not /ɪ/
/iː/	*!		*	
☞ /ɪ/		*		*

Note that, in principle, the shift in perceptual cue weighting occurs in the copied L2 grammar only; the L1 grammar is considered to remain intact. Therefore, when the learner is in L1 Japanese mode, predominantly duration-based responses as in Figure 4-2 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 4-3. The magnitude of the cue weighting shift is expected to be more pronounced for more proficient learners. These predictions were tested by comparing the simulation results to real listeners' perception, which is presented in the next section.

4.2.3 Experiment

4.2.3.1 Participants

Thirty-two Japanese learners of English who had received formal English language education in Japanese secondary schools participated in the experiment (20 female, 12 male, mean age = 21.5). Twenty-seven of them were graduate or undergraduate students at Waseda University, while others were graduates of the University or other universities in Japan. 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.

4.2.3.2 Stimuli

The stimuli were 49 synthetic vowels differing in spectral and duration values (Figure 4-5), created using the Klatt synthesizer (Klatt & Klatt, 1990) in Praat (Boersma & Weenink, 2019). 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 /iː/-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 4-5). 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 4-5). Fundamental frequency (F0) was fixed at 140 Hz, following Hillenbrand et al.'s measured F0 values for /iː/ (130Hz) and /ɪ/ (130Hz). Intensity was fixed at 70 dB.

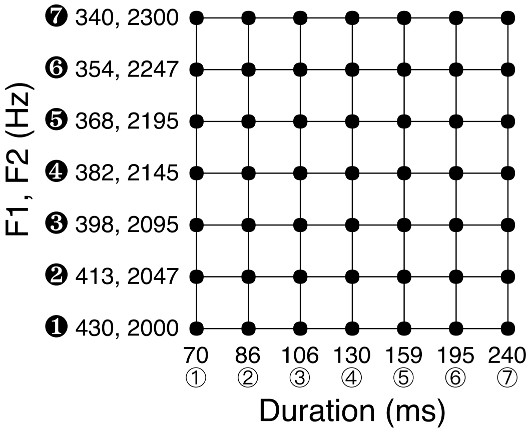


Figure 4-5. Acoustic values of the 49 stimuli used in experiment.

4.2.3.3 Procedure

To test participants' perception in L1 Japanese and L2 AmE, the experiment included Japanese (JP) and English (EN) sessions. Participants were informed that they would hear sounds from different languages (i.e., Japanese or English) between sessions, although they in fact heard an identical set of stimuli as in Figure 4-5 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 (a total of 245 trials 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 *oziisan* /oziisaɴ/ 'elderly man' or that of *ozisan* /ozisaɴ/ 'middle-aged man' displayed on a computer screen. Illustrations were used to avoid orthographic influences. The EN session followed a similar procedure, where the task was to identify each of the stimuli as either /iː/ or /ɪ/ in English by clicking either an illustration of *sheep* /ʃiːp/ or that of *ship* /ʃɪp/.

The two sessions were consecutive, and session order was counterbalanced 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 – 40 minutes to complete.

4.2.3.4 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 the current study is:

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

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 (βs) 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 their reliance on vowel duration is high. As explained earlier, 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, which represents relative weighting of spectral cues over durational cues (Casillas, 2015; Escudero et al., 2009) where a value above 0.5 means that vowel spectra is weighted heavier than duration:

$$\text{Cue weighting} = \frac{\beta_{spec}}{\beta_{spec} + \beta_{dur}}$$

Visual inspection of the data 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 /iː/ 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 it is not sure 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, 'participated in the experiment' where the 49 stimuli in Figure 4-5 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.

4.2.3.5 Result

Table 4-2 provides by-participant and -session results of logistic regression analyses. Participants have been sorted in the order of cue weighting in the EN session. When aggregated (Figure 4-6), 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. Linear mixed effects (LME) models were fitted to the response data (except for the three excluded participants) using the *lme4* (Bates et al., 2015) and *lmerTest* (Kuznetsova et al., 2017) packages in R (R Core Team, 2019), 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. 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 spectra and less on duration when they thought they were listening to English as opposed to Japanese.

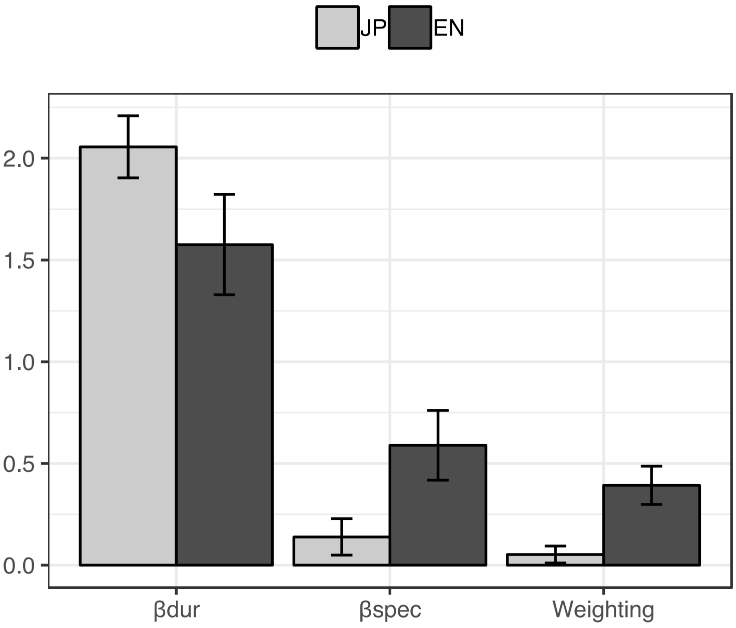
Table 4-2. Result of logistic regression analysis for each participant per session.

	β_{spec}		β_{dur}		Weighting	
ID	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

*: excluded from statistical analysis

Table 4-3. Result of logistic regression analysis for the virtual learner.

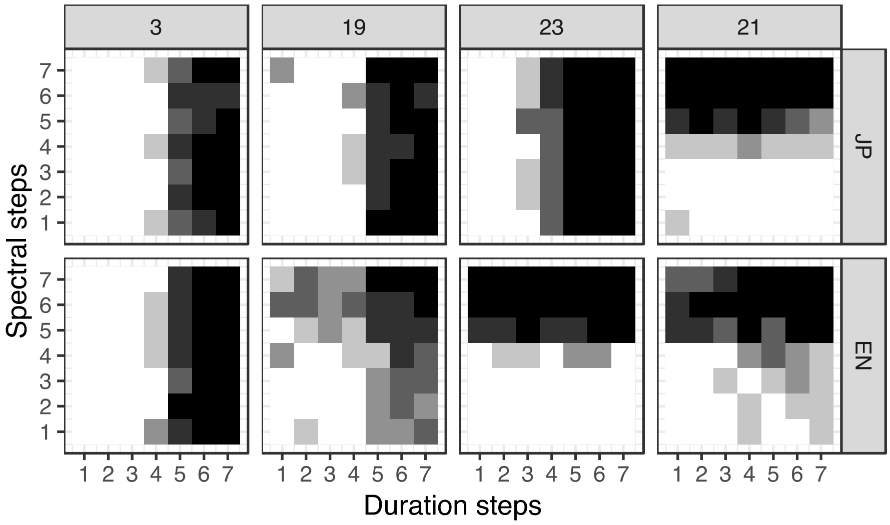
Grammar	Input	β_{spec}	β_{dur}	Weighting
L1 Japanese	40,000	0.07	1.98	0.03
L2 AmE	1,000	1.28	1.21	0.52
L2 AmE	2,000	1.37	1.17	0.54
L2 AmE	4,000	1.74	0.84	0.67
L2 AmE	40,000	2.65	0.53	0.83
L1 AmE	40,000	4.50	-0.04	1.01



(*22, *25, *31 excluded)

Figure 4-6. Mean β_{dur}, β_{spec}, and cue weighting ±1 standard errors.

Figure 4-7. Response patterns of Participants 3, 19, 23, and 21.



Additional LME models were run to test whether the participants' experience of having lived in the United States affected β_{spec}, β_{dur}, and cue weighting as a fixed effect, which yielded non-significant results. Although Participant 32, who had lived in the United States for four years, showed a drastic shift in cue weighting between sessions, other participants (Participants 16, 17, and 18) who had spent more than a year in the United States showed only a subtle shift. While no effect of participants' L2 proficiency or experience was found, substantial individual variability was found in the participants' response patterns. This is illustrated in Figure 4-7, in which darker color indicates a more frequent perception of /iː/ 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, 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. 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 4.2.4.

The participants' responses are directly comparable with those of the virtual learner, which is presented in Table 4-3. 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, 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 simulated L2 AmE (e.g., 40,000 tokens) and L1 AmE grammars.

4.2.4 Discussion

4.2.4.1 Interim summary

This study examined whether Japanese listeners' perceptual cue weighting for high front vowels change according to the language context (i.e., L1 Japanese or L2 AmE). The 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 identical. However, the degree of such a 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. Furthermore, no effect of L2 proficiency was found.

4.2.4.2 Effects of language mode

The experimental findings are compatible with predictions made by the L2LP model and Grosjean's language mode hypothesis. As mentioned earlier, L2LP interprets language modes as a selective or parallel activation of separate L1 and L2 grammars during speech perception. The observed shift in cue weighting in the experimental results can be interpreted as a result of different activation levels of the two grammars. Specifically, the participants' primarily duration-based perception in the JP session is attributable to strong activation of the L1 perception grammar in monolingual L1 Japanese mode, whereas their more spectra-based and less duration-based perception in the EN session is attributable to more activation of the L2 grammar in L1-L2 bilingual or possibly in L2 monolingual mode. In other words, the participants positioned themselves at different points along the language mode continuum (cf. Figure 4-1) across the experimental sessions, which affected their perceptual behavior. The results are also comparable to the 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 4-7.

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 strong activation of their monolingual L1 Japanese mode.

Although the experiment in the current study 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; L. 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 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 conducted in English). And yet, the bilinguals did show responses that were appropriate for each context. For testing the perception of Japanese and English high front vowels, phonological palatalization of certain consonants preceding high front vowels in Japanese (e.g., /si/ becoming [ɕi]) and the lack thereof in English (e.g., [siː]) could be useful in preparing stimuli where the consonant signals the language context.

Finally, it is worth noting that no effect of L2 proficiency (or experience) on the participants' perception patterns was found. As mentioned earlier, 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 & Simonet, 2018). Chang's (2012, 2013) studies on L1 phonetic drift have also shown 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. Vowel duration remained as the stronger cue for most of the listeners despite the shift in cue weighting, and three out of the four participants who had spent a relatively long time (more than one year) in the United States exhibited duration-based perception. This indicates that achieving optimal perception for SIMILAR L2 sounds can be quite challenging, contrary to L2LP's prediction that it is the least difficult scenario. A follow-up study with L1 Japanese listeners who have resided in the United States for an extended period of time could help further test the effects of proficiency.

4.2.4.3 Implications for SLM and PAM(-L2)

While L2LP in conjunction with the language mode hypothesis can straightforwardly explain the experimental results, SLM and PAM(-L2) provide alternative interpretations. SLM would explain Japanese listeners' persistent use of duration in perceiving AmE /iː/ 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, in which case Japanese listeners may establish a spectral distinction between AmE /iː/ and /ɪ/. PAM(-L2) makes somewhat similar predictions to SLM. According to the model, Japanese listeners' perception of AmE /iː/-/ɪ/ falls into a Two-Category (TC) assimilation pattern, in which each L2 sound is assimilated to a different L1 category 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 /iː/-/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 a case, a new category for the deviant L2 sound is reasonably likely to be formed, possibly leading to a spectral distinction between AmE /iː/ and /ɪ/. To summarize, SLM and PAM(-L2) predict that Japanese listeners are likely to maintain duration-based perception for AmE /iː/-/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.

Whereas both models predicted that Japanese listeners would likely continue to use duration 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 /iː/ or /ɪ/, or perhaps both. SLM would explain that, for some learners, equivalence classification resulted in the 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. The results thus are compatible with the two models. However, this flexibility of the models that 'predicts' both cases of assimilation and category formation at the same time indicates that the models are indecisive. Using these models, the researcher can hypothesize whether a particular learning scenario is subject to category assimilation or new category formation, or both as in the current case, even after the results are known. This problem is known as HARKing (Hypothesize After the Results are Known; Kerr, 1998). Besides, it is not very clear for which L2 sound new category formation might have occurred in the current scenario. One possibility is that the lax vowel /ɪ/ is prone to new category formation. As reviewed in Section 2.4.1, Strange et al. (2011) reports that Japanese listeners can perceive AmE /ɪ/ as Japanese /e/ instead of /i/ contrary to Strange et al. (Strange et al., 1998, 2001). This is likely due to the dialectal variation in the stimuli between the former study (New York dialect) and the latter two studies (Midwestern dialect). 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 is expected to be unlikely for AmE /iː/ as it was most likely perceived as Japanese /ii/ in all of the three studies. However, again, the researcher could HARK which sound is subject to new category formation.

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 & 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. The incorporation of such an idea would reinforce the theoretical principles of SLM and PAM(-L2).

4.2.4.4 Limitations and future directions

Although the computational implementation of L2LP provided explicit predictions for the current study's findings, a few limitations need to be addressed. First, it was assumed in the simulations that spectral and durational cues are equally used, which may not hold true. Bohn (1995) found that native listeners of Spanish and Mandarin, neither of which uses duration to differentiate vowel contrasts, relied heavily or exclusively on vowel duration to distinguish English /iː/-/ɪ/ and /ɛ/-/æ/ that are mainly distinguished by vowel spectra by native English listeners. Likewise, Escudero et al. (2009) found that Spanish learners of Dutch favored vowel duration over vowel spectra to categorize Dutch /aː/-/ɑ/, which is also distinguished chiefly by vowel spectra by native Dutch listeners. 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. Another property of the current simulations to consider is that they assumed perfect correspondence between AmE /iː/-/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 to an L1 category. This could be modeled by a perception grammar that contains multiple levels of representations rather than the simple [auditory] to /surface/ grammar employed in the current study.

An important avenue 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 baselines. 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 perception modes.

4.2.5 Summary

Study 1 provides evidence that Japanese learners of English employ different cue weighting (duration vs. spectra) to differentiate perceptually SIMILAR L1 and L2 sound contrasts (i.e., Japanese /ii/-/i/ and AmE /iː/-/ɪ/) according to the current 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. The computational implementation of L2LP, which incorporates the language mode hypothesis, predicted and explains the experimental results well. However, there is a caveat that the model’s predicted difficulty for a SIMILAR scenario (i.e., least difficult of all) was not entirely accurate, as most of the learners showed persistent reliance on duration in L2 perception despite the observed shift in cue weighting.

4.3 Study 2: Subset scenario

4.3.1 Background

The second study examines monolingual AusE listeners' perception of native /iː, ɪ, ɪə/ and nonnative Japanese /ii, i/. This follows a Subset scenario in L2LP, in which the L2 contrast constitutes a subset of the L1 sound inventory. Unlike the Similar scenario as in Study 1, the Subset scenario has received little attention because it is commonly considered to be an 'easy' scenario. However, L2LP proposes that the scenario can pose distinct perceptual and representational difficulties to L2 learners, which the current study aims to investigate.

The AusE vowels /iː, ɪ, ɪə/ are characterized by unique acoustic properties not seen in any other variety of English. They share almost identical average or midpoint formant frequencies but differ in duration and VISC (Elvin et al., 2016; Harrington et al., 1997). Specifically, AusE /iː/ is phonologically considered a monophthong but typically shows onglide or diverging VISC. AusE /ɪə/ is phonologically a diphthong but its offglide or converging VISC is not always phonetically realized, especially in closed syllables (Cox, 2006; Cox & Palethorpe, 2007). AusE /ɪ/ is shorter than the other two vowels, and its VISC is converging but very small in magnitude. AusE listeners' perception of these vowels reflects these acoustic properties. As mentioned in Section 2.3.3, Williams, Escudero, and Gafos (2018) conducted a detailed analysis of monolingual AusE listeners' perceptual cue usage in categorizing /iː, ɪ, ɪə/. The study used a series of synthetic vowel-like stimuli, which had the same midpoint formant frequencies but different trajectory direction (TD; diverging, converging or zero), trajectory length (TL; 0.00, 0.30, 0.90, 1.50, 2.10 and 3.90 ERB), and duration (94, 129, 177 and 244 ms). The experiment found that duration was by far the strongest cue for distinguishing /iː, ɪə/ from /ɪ/. TD and TL were important for categorizing /iː/ vs. /ɪ/, while only TL was important for categorizing /ɪə/ vs. /ɪ/. The observed cue usage can be straightforwardly explained by the acoustic characteristics of the vowels. Specifically, phonetically long /iː/ and /ɪə/ are distinguished from phonetically short /ɪ/ based on vowel duration. /iː/ and /ɪ/ are distinguished based on TD and TL because the former exhibits a large and diverging VISC, whereas the latter exhibits a small and converging VISC, thus differing in direction and magnitude of VISC. /ɪ/ and /ɪə/ are distinguished by TL (in addition to duration) but not TD because both vowels exhibit a converging VISC with different magnitudes of VISC. Thus, AusE listeners' perceptual patterns are predictable from the acoustic properties of the vowels.

The perceptual relevance of VISC has received increasing attention in recent years, not only for AusE but also for other languages (Morrison & Assmann, 2013). According to Morrison and Nearey (2007), three main hypotheses have been proposed as to how VISC perception can be represented: *onset* + *offset*, *onset* + *slope*, and *onset* + *direction*. All three hypotheses agree that the initial formant frequencies are perceptually relevant but disagree on what additional cues are relevant to vowel identification. The *onset* + *offset* hypothesis states that the formant frequencies at the end of the vowel are relevant for perception. The *onset* + *slope* hypothesis states that the relevant perceptual cue is the rate of change of formants over time. The *onset* + *direction* hypothesis states that the only relevant cue is the direction of formant movements. Although all three accounts can lead to correct vowel identification (Nearey & Assmann, 1986), Morrison and Nearey (2007) concluded that the *onset* + *offset* hypothesis is the most superior in terms of higher correct-classification rates and higher correlation with human-listeners' responses. The above experimental result of AusE /iː, ɪ, ɪə/ perception in Williams et al. (2018) also supports the *onset* + *offset* hypothesis. The finding that /ɪə/ and /ɪ/ were distinguished by TL but not by TD suggests that the *onset* + *direction* hypothesis is insufficient because listeners perceived different vowels out of tokens sharing the same converging direction. The *onset* + *slope* hypothesis is not supported either because the slope of formant trajectories was kept constant across the stimuli. The *onset* + *offset* hypothesis is valid because the direction and length of formant trajectories essentially express the relation between onset and offset formants.

Thus, recent research provides a coherent explanation of how AusE listeners perceive their native vowels /iː, ɪ, ɪə/ based on temporal and dynamic spectral cues. However, little is known as to how they utilize these cues in nonnative speech perception. The current study aims to unveil this by investigating AusE listeners' perception of Japanese /ii/ and /i/ in relation to the native vowels. According to L2LP, this case follows a SUBSET scenario in which the L1 has more sound categories than the L2. Here, L1 optimal perception equals accurate categorization of the three vowel categories /iː, ɪ, ɪə/ based on VISC and duration, whereas L2 optimal perception equals accurate categorization of the two vowel categories /ii, i/ based on duration only since VISC is irrelevant to this contrast (Ingredient 1). AusE listeners are expected to use the copy of their L1 perception grammar at the initial state of L2 perception, which is non-optimal because the VISC cue is unnecessary and also because the listener perceives too many categories than required (Ingredient 2). The learner's perceptual task then is to adjust their non-optimal use of the acoustic cues, while the representational task is to reduce the number of categories (Ingredient 3). As the learner receives more input in L2 Japanese, they would learn that duration is the only relevant cue for the L2 contrast and that only two vowel categories should be perceived (Ingredient 4). The expected end state is the perception of two vowel categories (/ii/ and /i/) based solely on duration (Ingredient 5).

The acquisition of this particular scenario is considered to be of medium difficulty because there are both perceptual and representational tasks to be completed.

As mentioned earlier, the SUBSET scenario has been largely neglected in L2 perception research because it is often considered to pose no perceptual problems for L2 learners. That is, learners may simply reuse existing categories without adjusting them and still be able to perceive the lexical contrasts in the L2 correctly. However, Escudero and Boersma (2002) argue that the SUBSET scenario results in multiple-category assimilation (MCA) where an L2 contrast is mapped to more than two categories in the L1 (Figure 4-8), which is problematic for two reasons. First, learners may create spurious lexical contrasts based on too many categories, with possible repercussions in production. Second, even if the learner acknowledges the appropriate number of categories in the L2, they cannot help perceiving an extraneous vowel category. In order to achieve optimal perception in the L2, the learner needs to solve these problems by disposing of the extraneous category, at least under the assumption that the L2 perceptual system is autonomous. Importantly, the possibility of such category reduction is not discussed in either SLM or PAM(-L2) because an L1 category is considered to remain in the common L1-L2 space as long as the category is used in L1 perception. On the contrary, L2LP assumes that an extraneous category should be disposed of from the L2 perception grammar because it is unnecessary for L2 optimal perception, while the category should remain intact in the L1 grammar.

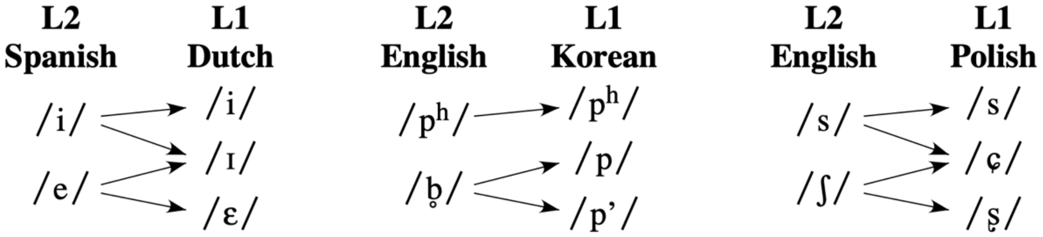


Figure 4-8. Examples of MCA (Escudero & Boersma, 2002).

The current study aims to shed more light on the potentially problematic nature of the SUBSET scenario by investigating AusE listeners' perception of Japanese /ii/ and /i/. The target population is monolingual AusE listeners without any previous knowledge of Japanese, so the current study focuses on nonnative rather than L2 perception. In what follows, I first present simulations of AusE listeners' perception of native /iː, ɪ, ɪə/ and nonnative Japanese /ii, i/ based on L2LP, in Section 4.3.2. Section 4.3.3 then presents a perception experiment to test the predictions, which examined AusE listeners' categorization of naturally produced all Japanese long and short vowels (including /ii/ and /i/) into their native vowel categories. The computational and experimental results are discussed in Section 4.3.4, focusing on the difficulty of acquiring L2 optimal perception in a SUBSET scenario. Section 4.3.5 provides a summary of the study.

4.3.2 Simulation

This section presents a series of L2LP-based simulations to model monolingual AusE listeners' perception of their native vowels as well as nonnative Japanese vowels. To this end, detailed acoustic information of the target vowels was first collected. As in Study 1, the current study utilizes F2/F1 ratios to represent vowel spectra in a single value (larger value = more peripheral). Table 4-4 shows mean F2/F1 ratios and duration of the target vowels in AusE and Japanese, as produced by male native speakers. The F2/F1 ratios were measured at the onset, midpoint, and offset of the vowels to characterize VISC. The AusE data was taken from Elvin, Williams, and Escudero (2016), in which seven male AusE speakers from Western Sydney (aged 18 – 30) produced 18 AusE vowels /iː, ɪ, eː, e, ɐː, ɐ, oː, ɔ, ʉː, ʊ, ɜː, æ, ɪə, æɪ, ɑe, æɔ, əʉ, oɪ/ embedded in six consonantal contexts: /bVp/, /dVt/, /fVf/, /gVk/, /hVd/, /sVs/). An example of a data collection format is: *See – ee – beep – beepa – In beep and beepa we have ee*. On the other hand, the Japanese data was taken from Yazawa and Kondo (2019), which adopted a very similar recording procedure. In the study, eight male Japanese speakers from greater Tokyo area (aged 21 – 30) produced ten Japanese vowels (/ii, i, ee, e, aa, a, oo, o, uu, u/) in disyllabic nonsense words $C_1V_1C_2V_2$ (where V_1 is the target vowel) with five consonantal context: /bV_1pV_2/, /dV_1tV_2/, /gV_1kV_2/, /zV_1sV_2/, /hV_1dV_2/. The V_2 was either /e/ or /o/ to minimize its influence on V_1, and the lexical pitch accent was placed on the first mora. These vowels were embedded in the carrier sentence *CVCe – CVCo – CVCe to CVCo ni wa V ga aru* 'CVCe – CVCo – In CVCe and CVCo there is V.' The utterances were recorded with a Sony F-780 microphone (sampling rate 44.1 kHz, 16-bit quantization) in an anechoic chamber at Waseda University.

Table 4-4. Mean F2/F1 and duration of target vowels in AusE and Japanese.

Language	Vowel	F2/F1$^{\text{ONSET}}$	F2/F1$^{\text{CENTER}}$	F2/F1$^{\text{OFFSET}}$	Duration (ms)
AusE	/iː/	5.04	6.15	7.06	168
AusE	/ɪ/	5.74	5.50	5.31	101
AusE	/ɪə/	6.22	5.82	5.11	205
Japanese	/ii/	7.53	7.49	7.54	147
Japanese	/i/	7.30	7.16	7.37	68

It can be seen from the table that the AusE vowels show a large degree of VISC (as represented by the change from F2/F1$^{\text{ONSET}}$ via F2/F1$^{\text{CENTER}}$ to F2/F1$^{\text{OFFSET}}$) and also differ in duration, whereas the Japanese vowels show little or no VISC and differ primarily in duration. Note that an increasing F2/F1 ratio represents diverging VISC while a decreasing F2/F1 ratio represents converging VISC. The dynamic properties of the AusE vowels are further illustrated in Figure 4-9, which shows the VISC trajectories of the vowels based on Elvin et al.'s (2016) data.

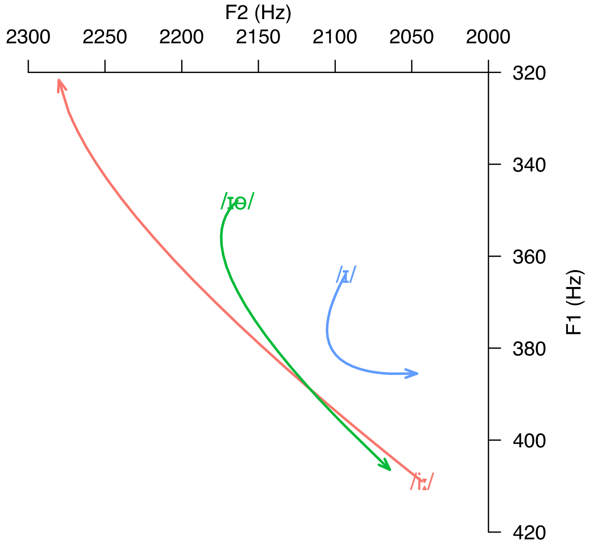


Figure 4-9. Average VISCs of AusE /iː, ɪ, ɪə/ based on Elvin et al. (2016)

In order to precisely model AusE listeners' perceptual space, the simulations focus on a range of F2/F1 ratio from 5.0 (F1 = 400 Hz and F2 = 2000 Hz; most central) to 7.67 (F1 = 300 Hz and F2 = 2300 Hz; most peripheral) and a range of duration from 70 ms to 210 ms. These ranges were meant to cover all the spectral and durational variability in Table 4-4. Each range then divided into 100 logarithmically equal bins (in natural log). The spectral range was further duplicated to create three ranges for the auditory dimensions of F2/F1$^{\mathrm{ONSET}}$, F2/F1$^{\mathrm{CENTER}}$, and F2/F1$^{\mathrm{OFFSET}}$. Each bin had three constraints prohibiting the perception of AusE /iː/, /ɪ/, and /ɪə/ respectively, giving a total of 1200 constraints (4 auditory continua × 100 bins × 3 vowel categories).

4.3.2.1 Native perception (AusE)

Based on the above acoustic data, a number of simulations were conducted to represent AusE listeners' perception of native vowels /iː, ɪ, ɪə/. Each simulation used a different set of acoustic cues as input to determine which cue is used for identifying which vowel. The general procedure of the simulations is as follows. Initially, the virtual learner has a blank perception grammar, in which all 600 constraints are ranked at the same height of 100.0. Evaluation noise is fixed at 2.0. The learner then starts learning AusE, receiving 40,000 randomly presented tokens of AusE /iː/, /ɪ/, and /ɪə/ occurring at the same frequency. The acoustic values of each token, namely F2/F1$^{\text{ONSET}}$, F2/F1$^{\text{CENTER}}$, F2/F1$^{\text{OFFSET}}$, and duration, are randomly drawn from normal distributions, with the means being those in Table 4-4 and the standard deviations being 0.1 for F2/F1 and 0.2 for duration (in natural log). 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 GLA adjusts 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. Once the training is complete, the learner is tested on their categorization of 1,000 randomly generated tokens. The parameter settings are thus identical to Study 1 except the standard deviations of the acoustic input.

Simulation 1: F2/F1$^{\text{CENTER}}$. The first model had access to the central formant ratio (F2/F1$^{\text{CENTER}}$) only. Figure 4-10 shows the ranking values of the model after being trained on 40,000 AusE vowel tokens. Note that a higher ranking value (i.e., constraint strictness) for a vowel category indicates a lower probability of the category being perceived. It can be seen from the figure that the ranking values are generally overlapping for all three vowel categories. This suggests that it is difficult to discriminate the vowel categories from one another. The model's categorization performance was very poor (Table 4-5), with their accuracy being slightly above the chance level (33.3%) for all three categories. The result indicates that central formant ratios are not an informative perceptual cue for identifying AusE /iː, ɪ, ɪə/.

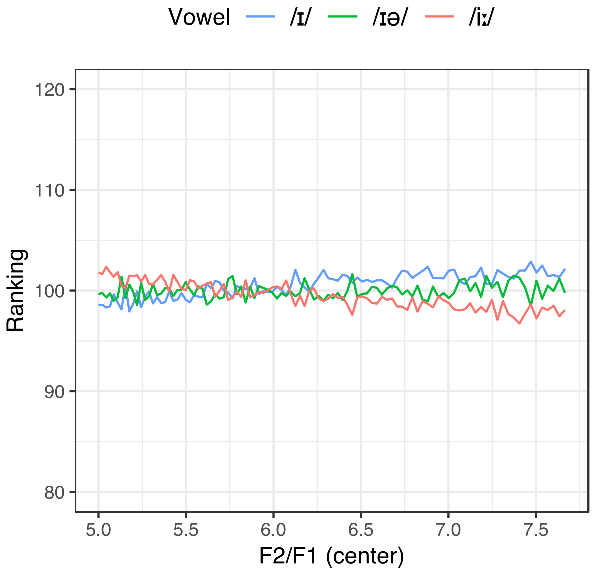


Figure 4-10. Model's ranking values (cue = F2/F1^{CENTER}).

Table 4-5. Model's classification accuracy (F2/F1^{CENTER}).

		Perceived		
		/iː/	/ɪ/	/ɪə/
Intended	/iː/	**44.1%**	23.9%	32.0%
	/ɪ/	27.7%	**39.0%**	33.3%
	/ɪə/	33.1%	31.9%	**35.0%**

Simulation 2: duration. The second model had access to vowel duration only. Figure 4-11 shows the ranking values of the trained model. Unlike Simulation 1, there is a noticeable difference in raking values between the three categories. The ranking value for /ɪ/ tends to be low (i.e., the vowel is likely to be perceived) when duration is short, and tends to be high (i.e., the vowel is unlikely to be perceived) when duration is long. In contrast, the ranking values of /iː/ and /ɪə/ show the opposite pattern, i.e., high when duration is short and low when duration is long. Their ranking values are also overlapping. These results indicate that, based on duration, short /ɪ/ can be distinguished from long /iː/ and /ɪə/, but long /iː/ and /ɪə/ cannot be discriminated from each other very well. This can also be seen in the model's classification rates (Table 4-6), where /ɪ/ is categorized fairly accurately whereas /iː/ and /ɪə/ seem to be confused, particularly when the intended form is /iː/. Therefore, vowel duration seems to be an important perceptual cue for AusE /ɪ/ vs. /iː, ɪə/, although other types of cues would be necessary to distinguish /iː/ and /ɪə/.

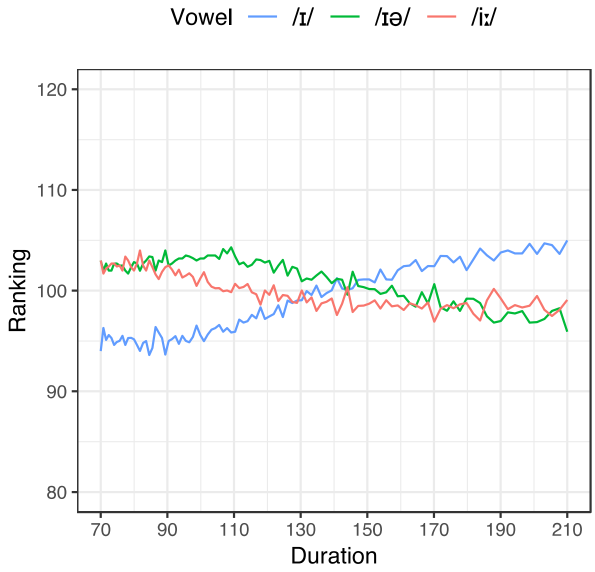


Figure 4-11. Model's ranking values (cue = duration).

Table 4-6. Model's classification accuracy (duration).

		Perceived		
		/iː/	/ɪ/	/ɪə/
Intended	/iː/	**44.2%**	14.3%	41.5%
	/ɪ/	12.1%	**<u>85.6%</u>**	2.3%
	/ɪə/	32.5%	3.0%	**64.5%**

Simulation 3: F2/F1$^{\text{ONSET}}$ + F2/F1$^{\text{OFFSET}}$. The third model had access to formant ratios at the onset and offset of vowel tokens (F2/F1$^{\text{ONSET}}$ and F2/F1$^{\text{OFFSET}}$). According to Morrison and Nearey (2007), onset and offset formant frequencies should be sufficient to represent the perception of VISC. Figure 4-12 shows the ranking values of the model trained on F2/F1$^{\text{ONSET}}$ and F2/F1$^{\text{OFFSET}}$. It can be seen that the ranking values for F2/F1$^{\text{ONSET}}$ are generally overlapping, while there is a clear distinction between /iː/ and

/ɪ, ɪə/ for F2/F1$^{\mathrm{OFFSET}}$. This indicates that F2/F1$^{\mathrm{ONSET}}$ is not a very informative cue for identifying AusE /iː, ɪ, ɪə/ and that /iː/ with diverging VISC can be distinguished from /ɪ/ and /ɪə/ with converging VISC based on F2/F1$^{\mathrm{OFFSET}}$. The results support the notion of "delayed target" in /iː/ (Harrington et al., 1997) because the target portion, namely F2/F1$^{\mathrm{OFFSET}}$, should be the most informative part of the vowel in terms of spectral quality. More specifically, /iː/ is likely to be perceived when F2/F1$^{\mathrm{OFFSET}}$ is high (e.g., [F2/F1$^{\mathrm{OFFSET}}$ = 7.5]) because its target is typically very peripheral. In contrast, /ɪ/ or /ɪə/ are likely to be perceived when F2/F1$^{\mathrm{OFFSET}}$ is low (e.g., [F2/F1$^{\mathrm{OFFSET}}$ = 5.0]) because their targets are typically central. The two vowels do not seem to be distinguished from each other based on either F2/F1$^{\mathrm{ONSET}}$ or F2/F1$^{\mathrm{OFFSET}}$ because they share similar converging VISC. The classification rates of the model (Table 4-7) confirm that /iː/ can be distinguished from /ɪ/ and /ɪə/ whereas the latter two vowels cannot be discriminated well based on the two acoustic cues. In sum, VISC as represented by F2/F1$^{\mathrm{ONSET}}$ and F2/F1$^{\mathrm{OFFSET}}$ (F2/F1$^{\mathrm{OFFSET}}$ in particular) can be used to distinguish diverging /iː/ from converging /ɪ/ and /ɪə/, although VISC alone does not inform the difference between /ɪ/ and /ɪə/ because they share similar movements.

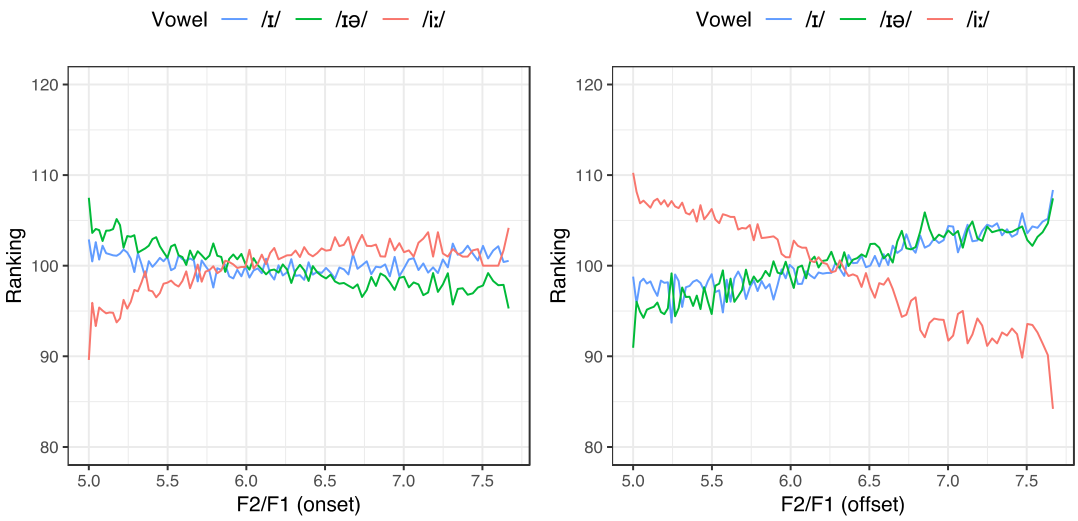


Figure 4-12. Model's ranking values (cue = F2/F1$^{ONSET \& OFFSET}$, F2/F1^{OFFSET}).

Table 4-7. Model's classification accuracy (F2/F1^{ONSET}, F2/F1^{OFFSET}).

		Perceived		
		/iː/	/ɪ/	/ɪə/
Intended	/iː/	**88.6%**	9.0%	2.4%
	/ɪ/	8.1%	**58.1%**	33.8%
	/ɪə/	2.5%	45.4%	**52.1%**

Simulation 4: F2/F1^{ONSET} + F2/F1^{OFFSET} + duration. The fourth model had access to F2/F1^{ONSET}, F2/F1^{OFFSET}, and duration of vowel tokens. Simulations 2 and 3 revealed that duration and VISC may play a complementary role: /ɪ/ can be distinguished from /iː/ and /ɪə/ based on duration, and /iː/ can be distinguished from /ɪ/ and /ɪə/ based on VISC (mostly by F2/F1^{OFFSET}). The current model examines whether combining the two types of cues results in an accurate classification of all three vowels. The prediction

was upheld. When trained on F2/F1$^{\mathrm{ONSET}}$, F2/F1$^{\mathrm{OFFSET}}$, and duration, the model successfully learned to classify all three vowels with very high accuracy rates (Table 4-9). Figure 4-13 shows the ranking values of the model, which reveals some additional points. First, although the constraint ranking patterns have not generally changed from the previous simulations, the overall amplitude of the ranking values seems to be enhanced. This is in line with the hypothesis that duration and VISC play a complementary role because it indicates that both types of cues need to be present for successful learning of the AusE vowel contrasts. Second, a comparison of the ranking values across the three types of cues reveal the relative strength of these cues. For example, duration seems to play a major role in identifying all three categories. F2/F1$^{\mathrm{OFFSET}}$ plays a large role for /iː/ vs. /ɪ, ɪə/. F2/F1$^{\mathrm{ONSET}}$ seems to play a relatively minor role in general.

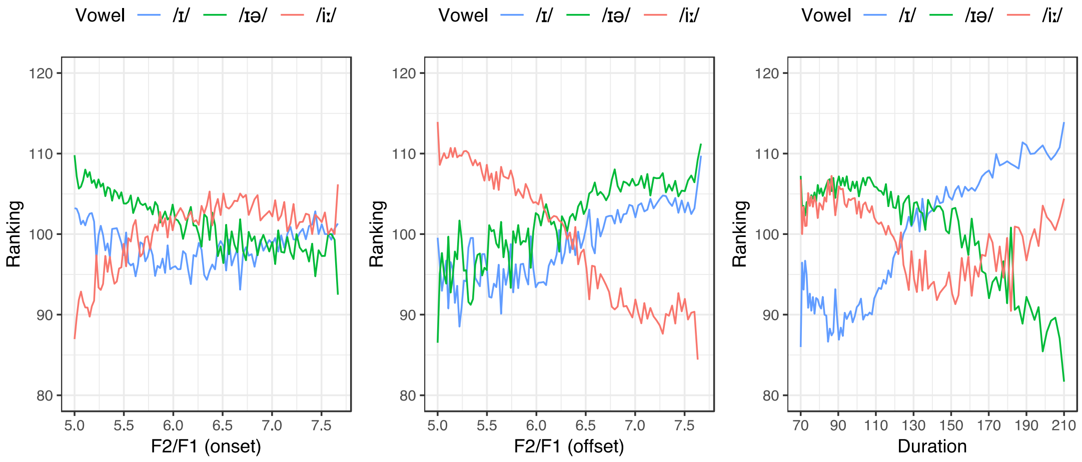


Figure 4-13. Model's ranking values (cue = F2/F1^{ONSET}, F2/F1^{OFFSET}, duration).

Table 4-8. Model's classification accuracy (F2/F1^{ONSET}, F2/F1^{OFFSET}, duration).

		Perceived		
		/iː/	/ɪ/	/ɪə/
Intended	/iː/	**93.7**	2.8	3.5
	/ɪ/	4.1	**91.0**	4.9
	/ɪə/	3.0	4.3	**92.7**

Simulation 5: F2/F1$^{\text{ONSET}}$ + F2/F1$^{\text{CENTER}}$ + F2/F1$^{\text{OFFSET}}$ + duration. The final model was meant to test whether adding the central formant ratio, which turned out to be uninformative on its own, to the previous model improves classification accuracy. The result found the opposite. The addition of F2/F1$^{\text{CENTER}}$ slightly lowered the classification accuracy for all three vowel categories, indicating that the acoustic cue is not informative even when combined with other acoustic cues. This indicates that the model with F2/F1$^{\text{ONSET}}$, F2/F1$^{\text{OFFSET}}$, and duration should suffice to represent AusE listeners' perception of /iː, ɪ, ɪə/.

Table 4-9. Model's classification accuracy (all cues) compared to Table 4-8.

		Perceived		
		/iː/	/ɪ/	/ɪə/
Intended	/iː/	**91.9% (-1.8)**	4.0%	4.1%
	/ɪ/	4.3%	**88.0% (-3.0%)**	7.7%
	/ɪə/	4.2%	3.6%	**92.2% (-0.5%)**

In sum, the above series of simulations of AusE listeners' perception found the following: (1) /ɪ/ can be distinguished from /iː/ and /ɪə/ based on duration, (2) /iː/ can be distinguished from /ɪ/ and /ɪə/ based on VISC as represented by F2/F1$^{\mathrm{ONSET}}$ and F2/F1$^{\mathrm{OFFSET}}$, (3) the relative strength of the acoustic cues is in the order of duration > F2/F1$^{\mathrm{OFFSET}}$ > F2/F1$^{\mathrm{ONSET}}$, and (4) F2/F1$^{\mathrm{CENTER}}$ is an uninformative cue for the contrast. The results are mostly compatible with the perceptual behavior of real AusE listeners found in Williams et al. (2018), in which (i) duration was important for categorizing /ɪ/ vs. /iː, ɪə/, (ii) VISC as represented by TD and TL were important for categorizing /iː/ vs. /ɪ/, (iii) duration was a stronger cue than VISC in general, and (iv) midpoint formant frequencies were not manipulated and listeners nonetheless perceived different vowel categories. The simulations also add to the result of Williams et al. (2018). For example, it was found in the simulations that F2/F1$^{\mathrm{OFFSET}}$ is more important than F2/F1$^{\mathrm{ONSET}}$, which could not be tested by Williams et al. (2018) because the onset and offset formant frequencies were manipulated at the same time. The simulations also found that F2/F1$^{\mathrm{CENTER}}$ plays little or no perceptual role, which was assumed but not directly tested by Williams et al. (2018). Based on the overall result, it can be concluded that the model with F2/F1$^{\mathrm{ONSET}}$, F2/F1$^{\mathrm{OFFSET}}$, and duration (i.e., Simulation 4) adequately represents AusE listeners' perception of native vowels /iː, ɪ, ɪə/.

4.3.2.2 Nonnative perception (Japanese)

This section examines how the above simulated AusE grammar would perceive unfamiliar nonnative Japanese vowels /ii, i/. L2LP asserts that, at the initial state of L2 acquisition (i.e., nonnative perception), listeners can only perceive L1 sound categories by applying L1-like cue usage (*Full Copying* hypothesis). Thus, it is expected that naïve AusE listeners would classify nonnative Japanese high front vowels as one of their native vowel categories /iː, ɪ, ɪə/ based on duration and VISC cues. Under this assumption, the model that was trained on F2/F1$^{\mathrm{ONSET}}$, F2/F1$^{\mathrm{OFFSET}}$, and duration of 40,000 AusE tokens (i.e., Simulation 4) was tested on 1,000 randomly generated Japanese tokens. The acoustic values of each test token were randomly drawn from normal distributions, with the means being those of Table 4-4 and the standard deviations being 0.1 for F2/F1 and 0.2 for duration in natural log (i.e., same as native AusE simulations). Given that the virtual listener is not learning Japanese, no error-driven learning occurred during testing.

Table 4-10 shows the model's classification of Japanese vowels into AusE vowel categories. It can be seen that Japanese long /ii/ is most likely perceived as AusE /iː/. Japanese short /i/ seems to be ambiguous between AusE /iː/ and /ɪ/, although AusE /ɪ/ is slightly more likely to be perceived. AusE /ɪə/ is an unlikely choice for both Japanese /ii/ and /i/, which implies that it is the extraneous category the listener wants to dispose of eventually. The constraint ranking values of the model in Figure 4-13 as presented earlier provide a detailed account of why the above perceptual pattern was observed. Here, note that Japanese /ii/ and /i/ typically show very peripheral spectral qualities throughout the vowel (e.g., [F2/F1 = 7.5]) and also that the mean durations of these vowels are approximately 150 ms and 70 ms respectively. First, the F2/F1$^{\text{ONSET}}$ constraints for [F2/F1$^{\text{ONSET}}$ = 7.5] do not seem to play much of a role, as they congregate at the default value of 100. This is presumably because the Japanese vowels are very peripheral at the onset, which does not resemble any of the AusE vowel categories. As for F2/F1$^{\text{OFFSET}}$ constraints, AusE /iː/ is very likely to be perceived when e.g., [F2/F1$^{\text{OFFSET}}$ = 7.5], reflecting the spectral proximity between AusE /iː/ and Japanese /ii, i/ at vowel offset or 'target'. However, the duration constraints compete with the F2/F1$^{\text{OFFSET}}$ constraints, particularly when vowel duration is short (e.g., [duration = 70 ms]). In other words, both Japanese /ii/ and /i/ resemble AusE /iː/ in terms of offset formant ratios, but Japanese short /i/ also resembles AusE /ɪ/ in terms of duration. Therefore, Japanese /ii/ is likely to be perceived as AusE /iː/ because they share similar spectral and temporal properties, whereas Japanese /i/ is ambiguous because it is close to AusE /iː/ in the spectral domain but close to AusE /ɪ/ in the temporal domain. AusE /ɪə/ is unlikely to be perceived at all because none of the acoustic characteristics of the vowel, namely intermediate

$\mathrm{F2/F1}^{\mathrm{ONSET}}$ (e.g., [$\mathrm{F2/F1}^{\mathrm{ONSET}}$ = 6.0]), low $\mathrm{F2/F1}^{\mathrm{OFFSET}}$ (e.g., [$\mathrm{F2/F1}^{\mathrm{OFFSET}}$ = 5.0]) and very long duration (e.g., [duration = 200 ms]), is present in either Japanese /ii/ or /i/.

Table 4-10. AusE model's classification of unfamiliar Japanese vowels.

		Perceived		
		/iː/	/ɪ/	/ɪə/
Presented	/ii/	**78.1**	11.6	10.3
	/i/	46.4	**51.5**	2.1

To summarize, the simulations predicted that AusE listeners would perceive Japanese /ii/ as AusE /iː/, but Japanese /i/ would be ambiguous between AusE /iː/ and /ɪ/ (/ɪ/ is slightly more likely to be perceived) because the vowel shares spectral similarities with the former but temporal similarities with the latter. Importantly, these predictions are based on the assumption that monolingual listeners would fully transfer their native cue usage to nonnative perception, following L2LP's *Full Copying* hypothesis. In order to test the predictions, the following experiment was conducted on real monolingual AusE listeners in Sydney, Australia.

4.3.3 Experiment

4.3.3.1 Participants

Twenty female AusE listeners were recruited for the experiment at the MARCS Institute, Western Sydney University. They were undergraduate or graduate students of the University between the ages of 17 – 35 (mean age = 21.4), born and raised in the greater Sydney area. All participants reported normal hearing and little to very basic knowledge of any foreign language. They were compensated for their time in the form of class credit or monetary compensation.

4.3.3.2 Stimuli

The stimuli were ten Japanese vowels – five short /i, e, a, o, u/ and five long /ii, ee, aa, oo, uu/ – embedded in three phonetic contexts (/bVp, dVt, gVk/) spoken by ten native Japanese speakers (five male, five female). This results in a total of 300 (10 vowels × 3 contexts × 100 speakers) tokens. The stimuli were created from a subset of the production data reported in Section 4.3.2 by clipping the /e/ from the /bVpe, dVte, gVke/ tokens in the sentence condition. The ten speakers were chosen out of 16 on the basis of minimal exposure to other languages and good recorded sound quality. The volume of the stimuli was then adjusted to have a peak intensity of 70 dB in Praat.

4.3.3.3 Procedure

The experiment was a forced-choice task, where the participants had to categorize the vowel in the CVC stimuli presented in isolation. The participants had to choose from a list the word that contains the vowel that best matches the vowel they heard in the stimuli. The words in the list all had the shape /hVd/ with the exception of two words, which had a /fVd/ shape (Table 4-11). Participants were asked to choose as quickly as possible. The stimuli were presented in random order through noise-isolating headphones, and participants selected their answer choices by clicking the word choice with a mouse. A break was programmed to occur after 150 tokens (midpoint of the experiment), which ended when participants clicked the mouse. The experiment was conducted in a sound-attenuated room at Western Sydney University, using PsychoPy2 v1.90.2 (Peirce, 2007).

Table 4-11. List of response words used in experiment.

Long		Short	
Word	Vowel	Word	Vowel
heed	/iː/	*hid*	/ɪ/
haired	/eː/	*head*	/e/
hard	/ɐː/	*hud*	/ɐ/
hoard	/oː/	*hod*	/ɔ/
food	/ʉː/	*hood*	/ʊ/
heard	/ɜː/	*had*	/æ/
feared	/ɪə/		

4.3.3.4 Analysis

In order to analyze how the participants' responses were affected by the actual phonetic properties of the stimuli, the acoustic values of each stimulus were measured in Praat. These included F1 and F2 frequencies at the onset, midpoint, and offset of the vowel as well as duration. Vowel duration was defined as the time between the first and last positive zero-crossings of the quasi-periodic waveform associated with the vowel. The formant values were obtained from the central portion (20%, 50%, and 80%, respectively), from which formant ratios (F2/F1$^{\mathrm{ONSET}}$, F2/F1$^{\mathrm{CENTER}}$, F2/F1$^{\mathrm{OFFSET}}$) were calculated.

Based on the acoustic values, the following logistic regression analysis was applied to a subset of the response data where presented stimuli were Japanese /ii/ and /i/:

$$Ln\left(\frac{P}{1-P}\right) = \alpha + \beta_{dur} \times \text{duration} + \beta_{ONSET} \times \frac{F2^{ONSET}}{F1^{ONSET}} + \beta_{OFFSET} \times \frac{F2^{OFFSET}}{F1^{OFFSET}}$$

where *P* is the probability that a particular response category is chosen, α is the intercept of the model, and βs represent to what extent the acoustic values (i.e., duration, F2/F1$^{\mathrm{ONSET}}$, and F2/F1$^{\mathrm{OFFSET}}$) contribute to the probability of the response category being chosen (in log odds). The model thus evaluated participants' reliance on the acoustic cues in choosing a particular response category when given tokens of Japanese /ii/ and /i/.

4.3.3.5 Result

Table 4-12 summarizes the categorization of Japanese vowels into AusE response alternatives, summed over speakers and participants. Most frequent responses are in bold. Only responses above 15% are labeled. It can be seen that Japanese long /ii/ was most frequently categorized as AusE /iː/. Japanese short /i/ was categorized as either /iː/ or /ɪ/, where /ɪ/ was slightly more likely to be perceived. AusE /ɪə/ was seldom chosen for either Japanese /ii/ or /i/. The result thus shows a close resemblance to the simulation, although the classification rates are not directly comparable because the experiment included more response categories than the simulation.

Table 4-12. Categorization of Japanese vowels into AusE vowels (percent values).

	AusE response categories												
	/ɪə/	/iː/	/ɪ/	/eː/	/e/	/ɐː/	/ɐ/	/æ/	/oː/	/ɔ/	/ʉː/	/ʊ/	/ɜː/
/ii/		**61.0**	15.3										
/i/		40.8	**43.2**										
/ee/				**25.2**	18.0								16.0
/e/		20.3	**35.0**		19.0								
/aa/						**38.3**	20.2						
/a/							**45.0**			18.3			
/oo/									**33.3**	16.3		25.5	
/o/										**45.8**		22.7	
/uu/											28.2	**39.7**	
/u/											17.7	**39.2**	

When fitted for the response category of AusE /iː/, the logistic regression model found a significant effect of both duration (β_{dur} = 9.63, s.e. = 1.32, z = 7.32, $p < .001$) and F2/F1$^{\mathrm{OFFSET}}$ (β_{OFFSET} = 0.15, p = 0.016). The coefficients indicate that participants were apt to choose AusE /iː/ when vowel duration was long and also when offset spectral quality was peripheral (which is associated with diverging VISC). The model fitted for AusE /ɪ/ also found significant effects of both duration (β_{dur} = -18.1, $p < .001$) and F2/F1$^{\mathrm{OFFSET}}$ (β_{OFFSET} = -0.19, s.e. = 0.07, z = -2.77, p = 0.006), indicating that AusE /ɪ/ was more likely to be chosen when vowel duration was short and also when offset spectral quality was central (which is associated with converging VISC). The model fitted for AusE /ɪə/ found a significant effect of duration (β_{dur} = 9.32, p = 0.0016) and F2/F1$^{\mathrm{ONSET}}$ (β_{dur} = 0.14, p = 0.04), indicating that /ɪə/ was chosen when vowel duration was long and onset spectral quality was peripheral (which is associated with converging VISC). Adding F2/F1$^{\mathrm{CENTER}}$ to the above models did not significantly improve the model fit, according to the likelihood ratio test using R’s *lmtest* package (Zeileis & Hothorn, 2002). The overall result supports the simulations, which predicted that duration and F2/F1$^{\mathrm{OFFSET}}$ would play major roles while the effect of F2/F1$^{\mathrm{ONSET}}$ is smaller and F2/F1$^{\mathrm{CENTER}}$ is virtually uninformative. Crucially, the result confirms that monolingual AusE listeners apply L1-like use of VISC and duration cues to nonnative Japanese perception.

4.3.4 Discussion

4.3.4.1 Interim summary

This study examined monolingual AusE listeners' native perception of AusE /iː, ɪ, ɪə/ and nonnative perception of Japanese /ii, i/. The simulations predicted that AusE listeners would utilize acoustic cues of duration and VISC (as represented by onset and offset formant ratios) to identify their native vowels. More specifically, it was predicted that duration would be important for categorizing short /ɪ/ vs. long /iː, ɪə/, while onset and offset formants would be important for categorizing diverging /iː/ vs. converging /ɪ, ɪə/. Under the assumption that listeners use the same cues in nonnative perception, the simulations further predicted that Japanese long /ii/ would be categorized as AusE /iː/ (due to spectral and temporal similarities) and Japanese short /i/ as either AusE /iː/ or /ɪ/ (due to spectral or temporal similarities), whereas AusE /ɪə/ would rarely be perceived (due to spectral and temporal dissimilarities). The experiment supported these predictions. Participants exhibited L1-like usage of acoustic cues in categorizing nonnative Japanese /ii/ and /i/, where they chose /iː/ when vowel duration was long and offset quality was peripheral (which is associated with diverging VISC) and /ɪ/ when duration was short and offset quality was central (which is associated with converging VISC). AusE /ɪə/ was hardly chosen as a response category.

4.3.4.2 Overcoming the SUBSET scenario

The computational and experimental results are compatible with L2LP's Ingredient 1 (i.e., L1 and L2 optimal perception) and Ingredient 2 (i.e., L2 initial state). The series of simulations of native perception found that optimal perception of AusE /iː, ɪ, ɪə/ can be represented by the use of vowel duration and onset and offset formant ratios, which was congruent with real AusE listeners' perceptual behavior (D. Williams et al., 2018). The AusE optimal perception is somewhat different from the optimal perception of Japanese /ii/ and /i/, in which duration is the single most important cue. The mismatch in cue usage between Japanese and AusE optimal perception is considered to result in a non-optimal perception of the Japanese vowels by naïve AusE listeners. As expected, the simulations predicted non-optimal categorization of Japanese short /i/, which was predicted to be ambiguous between AusE /iː/ or /ɪ/. This was based on the assumption that AusE listeners would attend to VISC cues even though it is irrelevant to the distinction between Japanese /ii/ and /i/, applying a copy of their L1 perception grammar. The experiment confirmed the above prediction. AusE listeners were found to rely on duration as well as onset and offset formant frequencies when categorizing the Japanese vowels, perceiving /ii/ as AusE /iː/ and perceiving /i/ as ambiguous between AusE /iː/ and /ɪ/. Thus, the overall result is compatible with L2LP's optimal perception and *Full Copying* hypotheses.

However, L2LP's predictions regarding the other ingredients turned out to be questionable. Ingredient 3 asserts that learners have both perceptual and representational tasks in the SUBSET scenario, which makes the scenario supposedly more difficult than the SIMILAR scenario with only the perceptual task. However, there does not appear to be any presentational task for the current particular scenario because the 'extraneous' /ɪə/ category was not perceived at all. That is, the number of categories perceived at the initial state of L2 learning coincided with the number of categories required for L2 optimal perception, and therefore there is no need for category reduction. This indicates that the two possible problems specific to the MCA pattern, namely establishing spurious lexical contrasts based on too many categories and unavoidable perception of non-existent categories, do not apply to the current scenario. Therefore, the only remaining task would be the perceptual task of adjusting the non-optimal cue usage, i.e., reinforcing the duration cue and discarding the VISC cues.

A follow-up simulation was conducted to predict how AusE learners of Japanese would develop their L2 perception grammar (Ingredient 4) to achieve L2 optimal perception (Ingredient 5). To this end, the model in Simulation 4 was trained on Japanese input tokens, under the assumption that Japanese /ii/ and AusE /iː/ as well as Japanese /i/ and AusE /ɪ/ are representationally equal. The available acoustic cues to the virtual learner were F2/F1$^{\text{ONSET}}$, F2/F1$^{\text{OFFSET}}$, and duration, of which means and standard deviations were as previously described. The plasticity was inherited from the L1 simulation. The result found that the learner could achieve accurate categorization of Japanese /ii/ and /i/ (above 90%) with only a few thousand training tokens, which is much less than was required for the SIMILAR scenario in Study 1 (i.e., 40,000). During the process, the ranking values of the duration constraints were substantially modified, whereas the ranking values for F2/F1$^{\text{ONSET}}$ and F2/F1$^{\text{OFFSET}}$ were relatively unaffected. This is likely because the error-driven GLA ceased to update the grammar once accurate categorization was achieved based on the duration constraints. Thus, the simulation did not support L2LP's Ingredients 4 and 5 because accurate categorization of L2 sounds was attained by only partially achieving the perceptual task, i.e., modifying the temporal cue usage without adjusting the spectral cue usage. In other words, optimal L2 sound categorization seems to be achievable without perfectly acquiring the optimal perceptual mappings.

In sum, the SUBSET scenario of the current study was found to be not as difficult as L2LP would predict because there is no representational task and also because perfect perceptual mappings are not required for accurate sound categorization. The result thus agrees better with the widespread assumption that the SUBSET scenario poses little perceptual difficulty than with L2LP's claims.

4.3.4.3 Global use of duration

Although the current study focused on AusE listeners' perception of high front vowels, the experimental result revealed that their use of durational cues is not restricted to these vowels. AusE listeners seem to categorize long and short vowels in Japanese as long and short AusE vowels that match in height and backness, as Japanese long /ee, aa, oo/ were most frequently categorized as AusE long /eː, ɐː, oː/ while Japanese short /e, a, o/ were categorized as AusE short /ɪ, ɐ, ɔ/.[10] This duration-based perception is contrastive to AmE listeners' categorization of Japanese vowels (cf. Section 2.4.2), where Japanese long-short pairs were all categorized as tense AmE vowels and thus duration was disregarded.

This dialect-specific differences in cue usage between AusE and AmE are expected to affect L2 phonological acquisition patterns by native listeners of these languages. For example, as far as the acquisition of nonnative vowel quantity contrast is concerned, AusE listeners who utilize durational cues in their L1 may be more privileged than AmE listeners who do not. There is evidence that AusE listeners are capable of discriminating vowel length contrasts in foreign languages fairly accurately (cf. Tsukada (2010)). Tsukada (2012) investigated whether native listeners of Japanese and Arabic,

[10] Duration seems to be ignored in the perception of Japanese /uu/ and /u/, which were both categorized most frequently as AusE short /ʊ/. This is likely because Japanese /uu/ is spectrally distant from AusE long /ʉː/, which is much fronted than /ʊ/ and perhaps is more accurately a high front rounded vowel [ʏ].

both of which are characterized by having phonemic vowel length, could discriminate native (“known”) and nonnative (“unknown”) vowel length contrasts better than AusE control participants. The stimuli were phonologically long and short vowels in Japanese and Arabic. The Japanese stimuli were thus “known” for the Japanese participants and “unknown” for the Arabic and AusE participants. Likewise, the Arabic stimuli were “known” for the Arabic participants and “unknown” for the Japanese and AusE participants. The hypothesis was that the Japanese and Arabic participants would positively transfer their knowledge of phonemic vowel length in their L1 to nonnative speech perception, thus outperforming the control AusE listeners. Contrary to the expectation, the Japanese and Arabic listeners showed no advantage over the AusE listeners, which in turn proves AusE listeners’ sensitivity to nonnative vowel length. Tsukada (2012, p. 511) indeed notes that “the role of duration in the identification of vowels cannot be dismissed in Australian English.”

On the other hand, discrimination of nonnative length contrast is rather challenging for native AmE listeners. Hisagi, Shafer, Strange, and Sussman (2010) found that naïve AmE listeners showed weaker neural responses to and also poorer behavioral categorization of a Japanese vowel length contrast (*tado* vs. *taado*) than native Japanese controls. Tajima, Kato, Rothwell, Akahane-Yamada, and Munhall (2008) also found that native listeners of CE (which shares a very similar vowel system with AmE) who were perceptually trained to perceive Japanese vowel length contrasts improved performance only to a limited extent. Results did not show that trained listeners improved overall performance to a greater extent than untrained controls. Also, the training did not enable listeners to cope with speaking rate variation and did not generalize to untrained contrast types. These studies demonstrate that the acquisition of nonnative vowel length can be quite challenging for AmE and CE listeners.

Given that AusE listeners' use of duration is common across vowel categories in both native and nonnative perception, it can be said that vowel length is (pseudo-)phonemic in AusE. The feature hypothesis (McAllister et al., 2002) provides a straightforward explanation of AusE listeners' sensitivity to nonnative vowel length contrasts found in Tsukada (2012). The length feature is actually used to signal phonological contrasts in AusE (although perhaps to a lesser extent than languages such as Japanese and Arabic), which boosted AusE listeners' discrimination accuracy in nonnative length perception. This relates to Flege's (1995) claim that allophones may be too coarse a unit of analysis and therefore features need to be considered in some instances. The following Study 3 will tackle this issue further by comparing segment- vs. feature-based accounts of new L2 sound category formation.

4.3.4.4 Limitations and future directions

The most significant limitation of the current study is that it examined nonnative rather than L2 perception, which was due to the practical difficulty of recruiting a sufficient number of AusE learners of Japanese for the experiment. The follow-up simulation is useful for making hypotheses concerning AusE listeners' acquisition of the Japanese /ii/-/i/ contrast, but it lacks empirical evidence. The current study thus is a partial test of the L2LP model, whereby Ingredients 1 and 2 (the initial state of L2 perception is copied L1 perception) were supported, and Ingredient 3 (learners have both representational and perceptual tasks) was found to be questionable. Ingredients 4 and 5 (attainment of L2 optimal perception through the learning tasks) need experimental testing using real AusE listeners. Tsukada's (2010) preliminary result suggests that learners can develop sensitivity towards Japanese vowel duration and thus partially supports the simulation result, but it is unclear whether and how their spectral cue usage would change over time. For future research, a longitudinal study on AusE learners' identification of resynthesized Japanese /ii/ and /i/ with different acoustic properties would help reveal changes in their perceptual cue weighting as a result of L2 learning. Another possibility is to test the effect of intensive perceptual training as in Tajima et al. (2008). Such studies are needed to complement the result of the current study.

It also remains unclear whether the ‘extraneous’ AusE category /ɪə/ would eventually disappear from AusE listeners’ L2 system as they become proficient in L2 Japanese. This question is theoretically relevant because SLM and PAM(-L2) make quite a different prediction from L2LP. On the one hand, SLM and PAM(-L2) predict that the /ɪə/ category should remain unaffected in the L1-L2 common phonological space even after extensive L2 learning because the category still needs to be accessible during L1 perception (perhaps except in circumstances where the L1 is less and less used and undergoes attrition). On the other hand, L2LP predicts that the /ɪə/ category should be reduced and ultimately removed from the L2 perception grammar because it is unnecessary for optimally perceiving the L2 sound contrasts, while the category is expected to remain intact in the L1 perception grammar that is independent of the L2 grammar. Therefore, according to L2LP, advanced AusE learners of Japanese should strongly react to /ɪə/-like stimuli presented in the AusE language context but not in the Japanese context, whereas according to SLM and PAM(-L2), they should show comparable reactions regardless of the language context. Brain-imaging techniques such as electroencephalography (EEG) are useful in testing these hypotheses. For example, synthetic AusE /ɪ/- and /ɪə/-like stimuli can be used in an oddball paradigm where the language context is manipulated between sessions as in Study 1. When the stimulus changes from /ɪ/ to /ɪə/, advanced AusE learners of Japanese may exhibit a different degree of mismatch negativity (MMN) as a function of the current language mode (i.e., L1 AusE or L2 Japanese).

Last, it has been assumed throughout the study that the AusE vowels /iː, ɪ, ɪə/ were of equal phonological status in order to focus on the [auditory] to /surface/ mappings. However, the experimental result found that all AusE vowel categories except /ɪə/ were chosen as a response, indicating that the diphthong /ɪə/ may be distinguished from monophthongs at a phonological level. It can be hypothesized that /ɪə/ is, after all, underlyingly a diphthong, which can be phonetically realized with smaller VISC in certain phonetic contexts. Vowels such as /iː/, on the other hand, are underlyingly a monophthong although they can exhibit variable VISC. Such a distinction is comparable to the distinction between ‘true’ and ‘false’ diphthongs in AmE, where the former (i.e., /aɪ, aʊ, ɔɪ/) functions as a sequence of two units while the latter (i.e., /e^{ɪ}, o^{ʊ}/) acts as a phonetically complex single unit (Pike, 1947). This, again, highlights the necessity to consider high-level representations in order to provide the fullest account of L2 perception. Future work could test how AusE listeners perceive Japanese CVV (e.g., /hia/ [çia]) and palatalized CV (e.g., /h^{j}a/ [ça̠])) tokens that show similar acoustic characteristics to AusE /ɪə/ but have different phonological status.

4.3.5 Summary

Study 2 investigated AusE listeners' perception of Japanese /ii/ and /i/, which constitute a SUBSET of their native vowels /iː, ɪ, ɪə/. The experimental result found that naïve AusE listeners applied L1-like cue usage (duration and VISC) to nonnative Japanese perception, as was predicted by L2LP's simulations based on the *Full Copying* hypothesis. However, the model's prediction that a SUBSET scenario is of medium difficulty because of the representational task was not borne out. The experiment found that the 'extraneous' AusE /ɪə/ category was seldom perceived, and thus no representational task was attested for this particular scenario. The simulations also predicted that AusE listeners would easily acquire native-like categorization of Japanese /ii/ and /i/ without adjusting nonnative-like spectral mappings, suggesting that optimal perceptual mappings are not a prerequisite for optimal sound categorization.

4.4 Study 3: NEW scenario

4.4.1 Background

The third study examines Japanese listeners' perception of L1 Japanese /e, a/ and L2 AmE /ɛ, æ, ʌ, ɑ/. This follows a NEW scenario with already-categorized dimensions in L2LP, in which an L2 sound (e.g., /æ/) along the known auditory dimensions (e.g., F1 and F2) does not have an equivalent in L1 and therefore is new to the learner. While this type of scenario has been extensively studied under SLM and PAM(-L2), it has not received ample attention in L2LP because the model has focused on NEW scenarios with non-previously-categorized dimensions only (Escudero & Boersma, 2004). In Study 2, it was suggested that features may need to be considered in addition to segments for an adequate account of L2 perception. The current study extends this notion and explores whether a feature-based implementation of L2LP can provide a formal explanation of new L2 category formation, with a particular focus on AmE /æ/.

As reviewed in Section 2.4.1, Strange et al. (1998, 2001) conducted thorough investigations of cross-linguistic perceptual assimilation of AmE vowels into Japanese vowels. They found that AmE /ɛ/ and /ʌ, ɑ/ were spectrally assimilated to Japanese /e/ and /a/ respectively, with relatively high goodness ratings. The result suggests that the

AmE vowels are fair exemplars of the Japanese vowels.[11] On the other hand, AmE /æ/ was perceived as a very poor exemplar of Japanese /a/, as the vowel was least consistently assimilated and also received the poorest goodness ratings. According to SLM and PAM(-L2), such a deviant exemplar of L1 categories in L2 is subject to new category formation. SLM claims that a new phonetic category can be established for an L2 sound if the learner discerns at least some of the phonetic differences between the L1 and L2 sounds. Based on the above studies, it appears that native Japanese listeners can notice the phonetic differences between AmE /æ/ and Japanese /a/, which is considered to result in new category formation in the L2. Similarly, PAM(-L2) predicts that a new phonetic and phonological category is likely to be formed for a deviant exemplar of an L1 sound. The current case would follow the CG assimilation pattern, in which AmE /æ/ is assimilated to Japanese /a/ but is discrepant from the native "ideal," or possibly the UC pattern, in which the L2 vowel is not assimilated to any of the L1 categories. However, the exact likelihood of new category formation for this L2 sound is unclear because it is difficult to estimate the degree of perceived cross-linguistic phonetic distance objectively (Flege, 1995). Best and Tyler also state that "the exact developmental progression of the deviant phone remains a topic for future research" (2007, p. 29).

[11] Strange et al.'s (2011) result was somewhat different because AmE /ɛ/ was most likely categorized as Japanese /a/ with relatively poor goodness ratings.

The L2LP model provides an alternative account. According to the model, the current case follows a NEW scenario with already-categorized dimensions (F1 and F2). L1 and L2 optimal perception for this scenario is an accurate categorization of two Japanese vowels /e, a/ and four AmE vowels /ɛ, æ, ʌ, ɑ/ respectively, based primarily on the two spectral cues (Ingredient 1). The initial state of L2 perception is non-optimal for the L2 environment because the copied Japanese grammar can perceive fewer categories than required in AmE (Ingredient 2). The learning tasks, then, is to create new sound categories (i.e., representational task) by redistributing or splitting existing L1 perceptual mappings (i.e., perceptual task) along the F1 and F2 dimensions (Ingredient 3). The learner is considered capable of undergoing both tasks as they have access to an L1-like learning device that enables category formation and boundary adjustment (Ingredient 4). The expected end state is L2 optimal perception of the four AmE vowel categories with appropriate perceptual boundaries, with the L1 perception grammar being unaffected (Ingredient 5). However, attainment of L2 optimal perception is considered to be very challenging, as the NEW scenario is considered the most difficult scenario of all. Note that there is another kind of NEW scenario in L2LP, which involves a non-previously-categorized auditory dimension. For example, Spanish learners' acquisition of /iː/ and /ɪ/ in SBE (Escudero & Boersma, 2004) would apply to this case because the auditory dimension of vowel duration is not used phonologically in Spanish. This sub-scenario is considered to be extremely difficult because learners have to create new perceptual mappings on the uncategorized, blank-slate dimension and integrate the newly established cue to create a new sound representation. However, the relative levels of difficulties between already-categorized and non-previously-categorized NEW scenarios have not been identified, which the current study aims to reveal.

While the above explanations of SLM, PAM(-L2), and L2LP concern the relationship between L1 and L2 sound categories, it has been suggested that more fine-grained units such as features may be necessary for adequately explaining L2 phonological acquisition (Flege, 1995). There is emerging evidence that features indeed play an important role in speech perception. Perhaps the most significant is a neurolinguistic study by Mesgarani, Cheung, Johnson, and Chang (2014), who found that human STG, which is associated with the processing of speech sounds (Jacquemot et al., 2003), shows selectivity to distinct phonetic features such as vowel height and backness, not to distinct phonemes. The study used high-density direct cortical surface recordings in six participants while they listened to natural speech samples from the TIMIT Acoustic-Phonetic Continuous Speech Corpus, as part of their clinical evaluation for epilepsy surgery. The study was unprecedented, as previous studies were limited to the use of noninvasive neuroimaging techniques. Other behavioral studies also demonstrate the relevance of features in nonnative and L2 perception. Pajak and Levy (2014) found that the knowledge of a language with phonemic length in vowels leads to enhanced discrimination of nonnative consonant length contrasts. They tested the perception of Polish consonantal length distinctions by four groups of listeners with different L1 backgrounds: Korean, where length is a highly informative cue for both vowels and consonants, Vietnamese, where length is an informative cue only for vowels; Cantonese, where vowel length is a supplementary cue for vowel identity in addition to quality; and Mandarin, where vowel length is uninformative for segmental distinctions. The result found that the informativity of vowel length in the L1 was highly predictable of the sensitivity to nonnative consonantal length, suggesting that the length feature is shared across vowels and consonants. Olson (2019) also found that the acquisition of nonnative voicing features transfers to nontrained phonemes. In the study, native English speakers received phonetic production training on one of the three voiceless stop consonants in L2 Spanish /p, t, k/. Given that English voiceless stops are long-lagged and Spanish voiceless stops are short-lagged, the participants' VOT was expected to shift in the negative direction after training. The result found a significant change in VOT not only for trained (e.g., /p/) but also for untrained phonemes (e.g., /t, k/), suggesting that featural changes generalized to related phonemes sharing the voiceless feature. Thus, the evidence is strong that listeners use features in speech perception, be it L1 and L2. It follows, then, that L2 perception models should be extended to incorporate the role of features.

The current study aims to provide a formal account of new L2 category formation in Japanese listeners' perception of AmE vowels by comparing two versions of L2LP: segment-based and feature-based. Specifically, I attempt to demonstrate that the relationship between perceived cross-linguistic phonetic deviance and the process of new category formation, which is an important aspect of SLM and PAM(-L2) but is currently missing in L2LP, can be formally modeled using the featural version of the model. In what follows, I first present segmental and featural implementations of L2LP to model the current learning scenario (Section 4.4.2). A perception experiment is then presented in Section 4.4.3, which seeks evidence for new category formation in real L1 Japanese learners' L2 AmE perception. Section 4.4.4 then evaluates the two types of modeling in relation to the experimental result, discussing theoretical implications for L2LP and other L2 perception models. Section 4.4.5 provides a summary of the study.

4.4.2 Simulation

This section presents segmental and featural implementations of L2LP to model the process of new L2 category formation in L1 Japanese learners of L2 AmE. The simulations utilize the acoustic data of Japanese and AmE vowels reported in Nishi et al. (2008). Table 4-13 shows the average F1 and F2 of the target vowels in the two languages produced in the sentence condition. The original values in Hz were converted to the mel scale to represent the perceived spectral qualities better. The simulations focus on the spectral cues (F1 and F2) to model how Japanese listeners would establish spectral distinctions among the AmE vowels, without considering temporal cues. This is because spectral cues are generally more important than temporal cues for vowel identity in AmE, even though the target vowels /ɛ, æ, ʌ, ɑ/ are more dependent on temporal cues compared to other vowels (Hillenbrand et al., 2000). Therefore, the simulations do not consider the perception of Japanese long /ee/ and /aa/.

Table 4-13. Mean F1 and F2 of target vowels in AmE and Japanese.

Language	Vowel	F1 (mel)	F2 (mel)
AmE	/ɛ/	721	1368
AmE	/æ/	792	1363
AmE	/ʌ/	724	1144
AmE	/ɑ/	824	1145
Japanese	/e/	573	1421
Japanese	/a/	758	1086

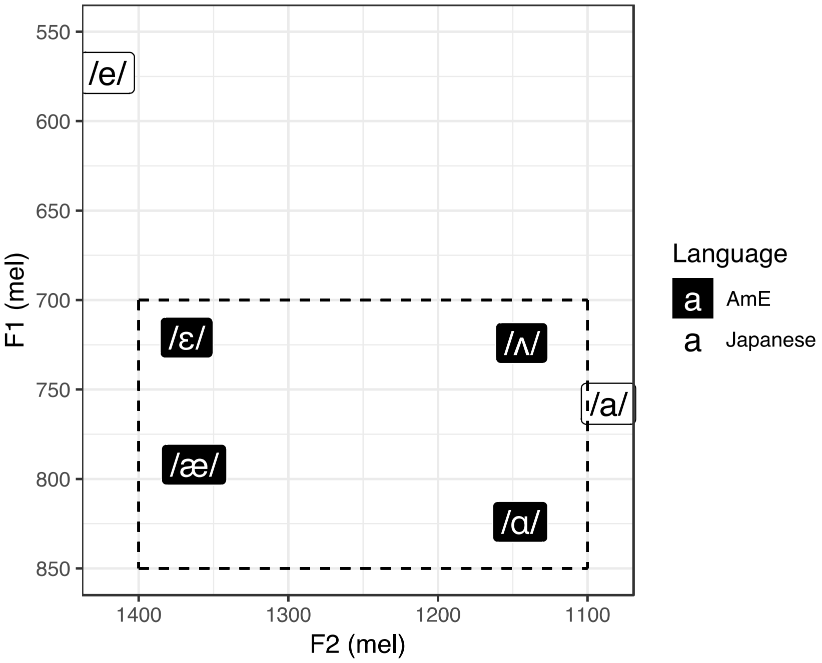


Figure 4-14. Target perceptual space in Japanese learners of AmE.

In order to precisely model the perceptual space, the simulations focus on a range of F1 from 700 mel to 850 mel and a range of F2 from 1100 mel to 1400 mel, as shown in Figure 4-14 (dashed square). The mean F1 and F2 of the target vowels are also shown. The ranges were chosen so that native AmE listeners would perceive the four AmE vowels while native Japanese listeners would perceive the two Japanese vowels within the space. Each range was then divided into 20 equally spaced 'bins' on the mel scale for the simulations. In what follows, I first present the conventional segment-based modeling of new category formation (Section 4.4.2.1), followed by the feature-based modeling (Section 4.4.2.2).

4.4.2.1 Segmental modeling

In this section, L1 Japanese listeners’ acquisition of L2 AmE vowels is modeled using auditory-to-segment constraints as in Studies 1 and 2. The primary focus is on how a new vowel segment is formed for AmE /æ/ out of the two Japanese vowels /e, a/. First, in order to model L1 Japanese perception, each of the F1 and F2 bins was assigned a pair of constraints, one prohibiting the perception of Japanese /a/ and the other prohibiting the perception of Japanese /e/. This resulted in a total of 80 constraints (20 bins × 2 auditory dimensions × 2 vowels), which were all initially ranked at the same height of 100.0. Evaluation noise was fixed at 2.0. The virtual listener then started learning Japanese, receiving randomly generated tokens of Japanese /e/ and /a/ occurring at the same frequency. The F1 and F2 values of each token were randomly drawn from normal distributions, with the means being those in Table 4-13 and the standard deviations being 100 mel. The acoustic values were then rounded to the nearest bins to be evaluated by the relevant constraints. Whenever there was a mismatch between the perceived form and the intended form, the GLA adjusted the ranking values of the relevant constraints by adding or subtracting the plasticity value. The plasticity was initially set at 1.0, which gradually decreased by a factor of 0.7 every 10,000 tokens. The parameter settings are thus mostly the same as Studies 1 and 2.

Figure 4-15 shows the model's perception of Japanese /e/ and /a/ after being trained on 40,000 tokens. The figure was obtained by feeding 400 (20 × 20) F1-F2 pairs to the model 1,000 times each. The labels show the most frequent responses.

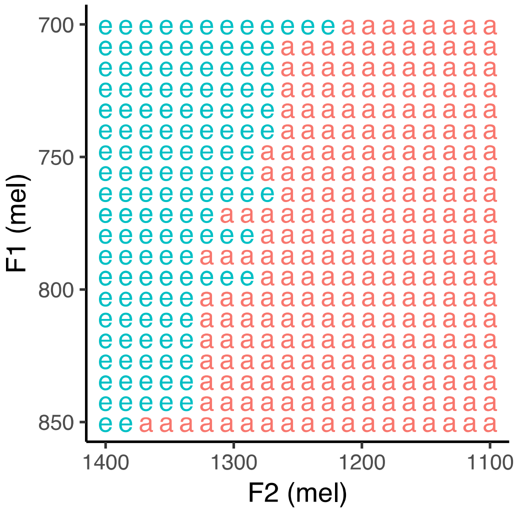


Figure 4-15. Segmental perception of L1 Japanese /e/ and /a/.

The model's perception can be formally represented as in Tableau 4-3, which shows the evaluation of an auditory event [F1 = 800 mel, F2 = 1400 mel]. In this example, the highest-ranked constraint "[F2 = 1400 mel] is not /a/" prohibits the perception of /a/ and thus the other candidate /e/ is perceived, even though the F1 constraints favor the perception of /a/. In other words, while the F1 is more likely that of /a/, the model prioritized the F2 cue and perceived /e/ because the vowel is very fronted.

Tableau 4-3. Segmental perception of [F1 = 800 mel, F2 = 1400 mel] as /e/.

[F1=800 mel, F2=1400 mel]	[F2=1400] not /a/	[F1=800] not /e/	[F1=800] not /a/	[F2=1400] not /e/
☞ /e/		*		*
/a/	*!		*	

The L1 Japanese model was then trained on L2 AmE vowels. At this stage, the model could perceive only Japanese /e/ or /a/ because the initial state of L2 perception is hypothesized to be a copy of the L1 perception grammar. Based on the assumption that AmE /ɛ/ and /ʌ/ are representationally linked to the two Japanese vowels /e/ and /a/ respectively (Strange et al., 1998, 2001), the constraints prohibiting the perception of Japanese /e/ and /a/ were reused as those prohibiting the perception of the two AmE vowels. However, in order to perceive four vowels in AmE, the learner has to create new segments by redistributing or splitting the L1 segments. The most straightforward way of modeling this is to simply add a new candidate and related constraints to the perception grammar, under the assumption that the listener somehow notices that there is more than /e/ (/ɛ/) and /a/ (/ʌ/) in AmE. The process of new category formation is represented in Tableau 4-4, which was created by adding the new candidate /æ/ and related constraints to the grammar of Tableau 4-3. The ranking values of these new constraints are expected to be initially high (e.g., 120) because listeners would prefer to perceive native segments than nonnative segments. However, as the error-driven learning based on the developing lexical-semantic knowledge proceeds (depicted by "←" and "→" in the tableau), the learner would eventually achieve appropriate constraint rankings.

Tableau 4-4. Segmental learning of [F1 = 800 mel, F2 = 1400 mel] as /æ/.

[F1=800, F2=1400]	[F2=1400] not /æ/	[F1=800] not /æ/	[F2=1400] not /a/	[F1=800] not /e/	[F1=800] not /a/	[F2=1400] not /e/
☞ /e/				←*		←*
/a/			*!		*	
✓ /æ/	*!→	*→				

Likewise, a new candidate /ɑ/ and related constraints can be added to the perception grammar so that the grammar can perceive all four possible vowels in AmE. The resultant L2 perception grammar would consist of 80 reused constraints prohibiting two AmE vowels (/ɛ, ʌ/) and 80 newly added constraints prohibiting the other two AmE vowels (/æ, ɑ/). When the ranking values of the 160 constraints were adjusted by the GLA based on 40,000 AmE tokens, the model successfully learned to perceive four AmE segments, as shown in Figure 4-16. The figure was again obtained by feeding 400 F1-F2 pairs to the model 1,000 times. Note that the auditory event of [F1 = 800 mel, F2 = 1400 mel], which used to be perceived as Japanese /e/ or AmE /ɛ/, is now perceived as /æ/. Thus, the virtual learner has successfully completed both the representational task of category creation and the perceptual task of adjusting perceptual boundaries.

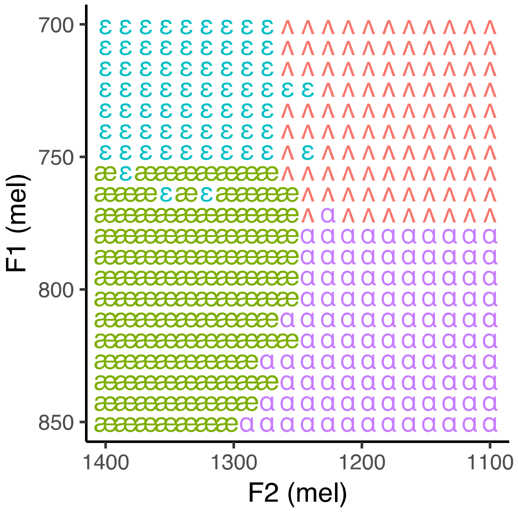


Figure 4-16. Segmental perception of L2 AmE /ɛ, æ, ʌ, a/.

However, the segmental approach is problematic for two reasons. First, new category formation is modeled as a one-time, out-of-the-blue phenomenon. The model does not acquire a new category until the category is explicitly introduced to the model, after which a new representation is firmly and suddenly established. This is unrealistic because learners are considered to gradually develop such representations. Second, the modeling does not directly incorporate the role of perceived L1-L2 deviance in the process of new category formation. The above simulation assumed that a new category should be formed for AmE /æ/, but other possibilities can also be considered, e.g., a new category is formed for /ɛ/. The choice of a 'new' segment is thus arbitrary.

4.4.2.2 Featural modeling

Boersma and Chládková (2011, p. 329) stated that the above segmental model "does not suffice to explain how real human listeners behave." They proposed that, in order to better explain speech perception, sound categories (e.g., /a/) should be represented as a bundle of phonetically based phonological features (e.g., /low, central/) instead of unanalyzed phonemes. In their modeling, an auditory dimension can map only to phonetically related features. For example, the F1 dimension maps only to vowel height features that are phonetically related to the dimension (e.g., /high/, /mid/, /low/) and not to other unrelated features such as backness and length features. The same is true for the F2 dimension, which maps only to vowel backness features (e.g., /front/, /central/, /back/). The cue constraints that handle the perceptual mappings (e.g., "[F1 = 800 mel] is not /high/") are thus comparable to the one-dimensional auditory-to-feature constraints in LP (Figure 3-6). Thus, the relationship between auditory dimensions and phonological representations is not arbitrary, unlike multi-dimensional auditory-to-segment constraints used in the segmental modeling (Figure 3-8). The perceived features are then combined to form a single unit such as /low, central/, which equates with a category such as /a/. Feature combinations are further constrained by feature co-occurrence constraints (e.g., "*/low, front/") whose strictness depends on the organization of features in the language.

Boersma and Chládková's approach is readily applicable to the current learning scenario. Below I present how the featural version of L2LP can model L1 Japanese listeners' acquisition of L2 AmE vowels. First, to simulate L1 Japanese perception, two height (/mid/ and /low/) and two backness features (/front/ and /central/) were identified as relevant to the perception of /e/ (/mid, front/) and /a/ (/low, central/) in Japanese . Each bin on the F1 dimension was assigned a pair of height-related constraints, one prohibiting the perception of the /mid/ feature and the other prohibiting the /low/ feature. Likewise, each bin on the F2 dimension was assigned a pair of backness-related constraints, one prohibiting the /front/ feature and the other prohibiting the /central/ feature. Therefore, a total of 80 auditory-to-feature constraints (20 bins × 2 auditory dimensions × 2 features) were prepared. In addition to cue constraints, four constraints prohibiting certain feature co-occurrence were also prepared: "*/mid, front/," "*/mid, central/," "*/low, front/," and "*/low, central/." The 84 constraints were initially ranked at the same height of 100.0. The virtual listener then started learning Japanese, receiving randomly presented tokens of Japanese /mid, front/ (i.e., /e/) and /low, central/ (i.e., /a/) occurring at the same frequency. The GLA adjusted the ranking values of the relevant constraints whenever there is a mismatch between what was perceived and what was intended. The parameter settings are the same as the segmental modeling.

Figure 4-17 shows the featural model's perception of Japanese /mid, front/ (/e/) and /low, central/ (/a/). The figure was obtained by feeding 400 F1-F2 pairs to the model 1,000 times. Although the perceptual behavior resembles that of the segmental model, a formal representation of the perception grammar highlights a unique strength of the featural model (Tableau 4-5). In the tableau, it can be seen that the model's perception of [F1 = 800 mel, F2 = 1400 mel], which is an AmE /æ/-like sound, is influenced not only by cue constraints but also by feature co-occurrence constraints. Specifically, while the cue constraints would favor the perception of /low, front/, the feature co-occurrence constraints prevent it because such a feature combination does not occur in Japanese. In general, constraints prohibiting a feature combination that does not occur in the language becomes ranked very high, while those prohibiting a combination that occurs becomes ranked low. The combination of cue and feature co-occurrence constraints enables featural modeling to express perceptual deviance, which was not possible in the segmental modeling. For example, the featural listener knows that the sound consists of /low/ and /front/, but has to perceive another phonologically well-formed sound such as /mid, front/ because /low, front/ is phonologically ill-formed in Japanese. In other words, the listener knows the phonetic deviance of the given sound; it is too /low/ for Japanese /e/ but too /front/ for Japanese /a/.

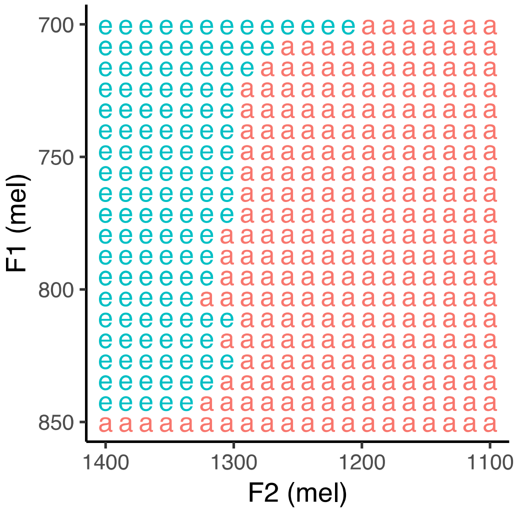


Figure 4-17. Featural perception of L1 Japanese vowels.

Tableau 4-5. L1 featural perception of [F1 = 800 mel, F2 = 1400 mel] as /mid, front/.

[F1=800, F2=1400]	[F2=1400] not /central/	*/low, front/	*/mid, central/	[F1=800] not /mid/	…	*/mid, front/
☞ /mid, front/				*		*
/mid, central/	*!		*	*		
/low, front/		*!				
/low, central/	*!					

Such an L2 sound that is perceived as deviant from L1 exemplars is expected to undergo new category formation according to SLM and PAM(-L2). The featural version of L2LP can model the emergence of new L2 categories as a result of reorganization of existing L1 features to allow illegal feature combinations. Following the *Full Copying* hypothesis, the featural Japanese listener initially categorizes AmE vowels based on the cue and feature co-occurrence constraints inherited from L1 learning. However, as the learner receives more input in L2 AmE, the learner will gradually adjust the ranking values of both types of constraints to achieve L2 optimal perception. For example, given that /low/ and /front/ features do occur simultaneously in AmE (i.e., /æ/), the learner will gradually lower the ranking of the feature co-occurrence constraint "*/low, front/." Likewise, the feature co-occurrence constraint "*/mid, central/" should also be weakened as L2 learning proceeds. Eventually, the learner becomes capable of perceiving all four possible combinations of features, each of which correspond to one of the four AmE vowel categories: /mid, front/ (/ɛ/), /mid, central/ (/ʌ/), /low, front/ (/æ/), /low, central/ (/ɑ/). The boundaries between the features are also adjusted to maximize categorization accuracy.

Shown in Figure 4-18 is the perception of the virtual L1 Japanese learner of L2 AmE after being trained on 40,000 randomly generated AmE tokens. The model is now capable of categorizing all four AmE vowel categories with appropriate perceptual boundaries, which suggests that the learner has successfully completed the representational and perceptual tasks proposed in L2LP. The performance of the featural learner is compatible with that of the segmental learner in Figure 4-16. Tableau 4-6 shows the perception of [F1 = 800 mel, F2 = 1400 mel] by the featural learner, which used to be categorized as /mid, front/ (/e/) but now as /low, front/ (/æ/). When compared to Tableau

4-5, it can be seen that the feature co-occurrence constraint "*/low, front/" that used to be ranked high in L1 Japanese grammar has been weakened through L2 learning. The sound is therefore no longer a deviant, illegal feature combination and is a permissible, well-formed combination in the listeners' L2 perception grammar.

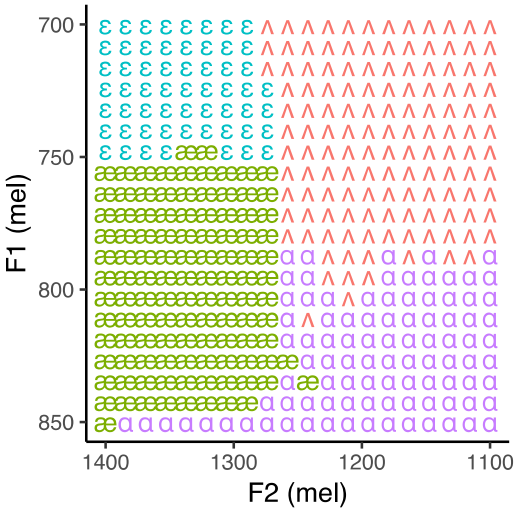


Figure 4-18. Featural perception of L2 AmE vowels.

Tableau 4-6. L2 featural perception of [F1 = 800 mel, F2 = 1400 mel] as /low, front/.

[F1=800, F2=1400]	[F2=1400] not /central/	*/mid, central/	*/mid, front/	[F1=800] not /mid/	…	*/low, front/
/mid, front/			*!	*		
/mid, central/	*!	*		*		
☞ /low, front/						*
/low, central/	*!					

In sum, the above simulations have demonstrated that the featural version of L2LP can model the process of new L2 category formation as a result of feature reorganization. While the predicted perceptual performance was comparable between segmental and featural modeling, the featural account is more advantageous than the segmental account in two points. First, a new category can gradually emerge based on existing L1 features in featural modeling without manipulating the features themselves, whereas in segmental modeling, a new category has to be intentionally added to the grammar at an arbitrary point of time. Second, featural modeling can express perceived L1-L2 phonetic deviance and how it relates to new L2 category formation, whereas such deviance cannot be directly represented in segmental modeling.

Importantly, the segmental and featural implementations of L2LP presented above are based on the fundamental assumption that L1 Japanese learners of L2 AmE would actually develop new perceptual representations in their L2 perception grammar. However, it is an unproven assumption because no previous studies have examined the presence of such representations in Japanese learners of AmE. Therefore, the following perception experiment was conducted to test how the L1 and L2 sounds are represented in the perceptual system of real Japanese learners of AmE.

4.4.3 Experiment

4.4.3.1 Participants

Thirty-six Japanese learners of English who had received formal English language education in Japanese secondary schools participated in the experiment (22 male, 14 female, mean age = 21.3). All of them were graduate or undergraduate students at Waseda University. They received some additional English instruction at the University, of which quality and quantity varied depending on the courses they enrolled in. None of the participants had overseas experience for more than three consecutive months, and none reported any history of hearing impairment. Thus, the participants' linguistic background is similar to that of Study 1, although their L2 proficiency is expected to be lower than those in Study 1 where nearly half of the participants had lived in the United States.

4.4.3.2 Stimuli

As in Study 1, the experiment included Japanese (JP) and English (EN) sessions to test participants' perception in L1 Japanese and L2 AmE modes. Two comparable sets of resynthesized [bVs] stimuli were prepared for the two sessions. The stimuli for the JP session were created from a natural token of Japanese *baasu* /baasu/ 'birth,' produced by a male native Japanese speaker in a carrier sentence *watashi wa* ___ *to iimasu* 'I say ___.'

The high back vowel /u/ is typically devoiced or deleted between two voiceless consonants (Shaw & Kawahara, 2017), and so the token was phonetically realized as [baːsu̥] or [baːs]. Likewise, the stimuli for the EN session was created from a natural token of English *bus* [bʌs], produced by a male native AmE speaker in a carrier sentence *I say* ___ *ten times*. The acoustic properties of the target vowel portion were then manipulated using STRAIGHT (H. Kawahara, 2006) in MATLAB (The MathWorks Inc., 2018). The F1 was set to vary in four equally spaced mel steps (700 mel, 750 mel, 800 mel, and 850 mel), and the F2 varied in another four mel steps (1100 mel, 1200 mel, 1300 mel, 1400 mel). These steps thus match the target perceptual space in the simulations (Figure 4-14). Vowel duration also varied in four logarithmic steps (100 ms, 114 ms, 131 ms, 150 ms) to cover the durational variability of AmE vowels reported in Nishi et al. (2008): /ɛ/ = 98 ms, /æ/ = 147ms, /ʌ/ = 98 ms, and /ɑ/ = 125ms. The F3 was fixed at 1700 mel. The vowels were further modified to have a mean F0 of 120 ms and a peak intensity of 70 dB without changing the original contours. This yielded 64 (4 × 4 × 4) [bVs] tokens for each session, where the F1, F2, and duration of the target vowel differed systematically, but other phonetic properties including consonant realizations and F0 and intensity contours were maintained from the original. Such stimuli are considered to elicit language-specific perception modes (Gonzales & Lotto, 2013).

4.4.3.3 Procedure

The experiment was a forced-choice identification task. The participants categorized the stimuli into four Japanese loanwords *beesu* 'base,' *besu* 'Bess,' *baasu* 'birth,' *basu* 'bus' (written in Japanese *katakana* orthography) in the JP session and into four English words *Bess*, *bass*, *bus*, *boss* in the EN session. Unlike the simulations, the Japanese long vowel response categories /ee/ and /aa/ were included to control for the number of response categories across sessions. Each stimulus was presented four times in random order, giving a total of 256 (64 × 4) trials for each session. Session order was counterbalanced to control for order effects: eighteen participants (11 male, 7 female) attended the JP session first, and the other 18 (11 male, 7 female) attended the EN session first. Oral and written instructions were given in each language, i.e., Japanese in the JP session and English in the EN session, to elicit language-specific perception modes. The experimenter was a Japanese speaker of English. Participants were tested individually in an anechoic chamber at Waseda University, in which the experiment was run on a computer using Praat's ExperimentMFC. Participants heard the stimuli at a comfortable volume through Sennheiser HD 380 Pro headphones and clicked the word choices on the screen using a mouse. The whole experiment took approximately 30 – 40 minutes to complete.

4.4.3.4 Analysis

In order to quantitatively investigate the participants' reliance on the three acoustic cues (F1, F2, and duration), logistic regression analysis was applied to the obtained response data. The model used in the current study is:

$$Ln\left(\frac{P}{1-P}\right) = \alpha + \beta_{F1} \times \text{F1} + \beta_{F2} \times \text{F2} + \beta_{dur} \times \text{duration}$$

where *P* is the probability that a particular response category (e.g., /a/) is chosen, α is the intercept of the model, and βs represent how the acoustic values (i.e., F1, F2, and duration) affect the probability of the response category being chosen (in log odds). The model was applied to each of the response categories, i.e., /ee, e, aa a/ for the JP session and /ɛ, æ, ʌ, ɑ/ for the EN session. The estimated coefficients can be used to graphically represent the participants' response categories in the perceptual space (Morrison, 2007). In the current study, the coefficients for F1, F2, and duration are used to visualize participants' perceptual representations of the Japanese and English vowels in the three-dimensional, F1-F2-duration space. Note that the coefficients are directly comparable between the JP and EN sessions because the three acoustic values were set to vary in the same way in both sessions.

4.4.3.5 Result

Figure 4-19 shows the participants' perceptual representations of the Japanese and AmE vowels in the F1-F2-duration space based on the average logistic regression coefficients. The Japanese (transparent points) and AmE vowels (filled points) are presented within the same space because the coefficients are directly comparable between the JP and EN sessions. Vertical lines are used to help clarify the exact location of the points in the three-dimensional space.

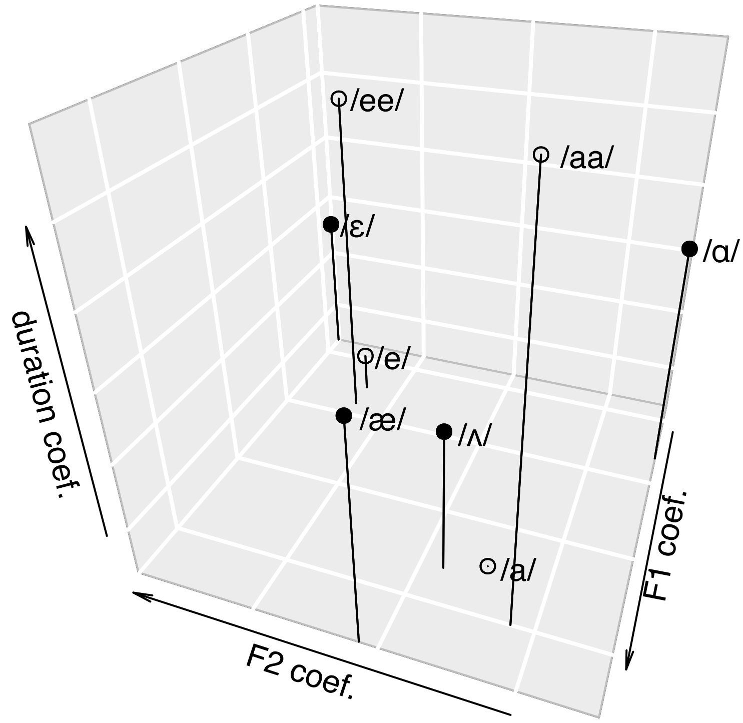


Figure 4-19. Participants' perceptual representations of Japanese and AmE vowels.

Several observations can be made from the figure. Starting with the Japanese vowels, /e/ is likely to be perceived when F1 is low, F2 is high, and duration is short. On the other hand, /a/ is likely to be perceived when F1 is high, F2 is low, and duration is short. The perceptual patterns are consistent with the phonological status of these vowels: /e/ is /mid, front, short/ while /a/ is /low, central, short/. Phonologically long /ee/ and /aa/ seem to be represented as longer versions of their phonologically short counterparts because the long-short vowel pairs occupy almost identical positions in the F1-F2 dimensions but differ along the duration dimension. This is congruent with the phonological description of Japanese long vowels being a sequence of two identical vowels. Moving on to the AmE vowels, it can be seen that /ɛ/ is spectrally close to Japanese /e/ but is temporarily intermediate between Japanese /e/ and /ee/. AmE /æ/ is somewhat distant from Japanese /a/ (/aa/) in terms of spectra and its duration is between Japanese /a/ and /aa/. AmE /ʌ/ is spectrally closer to Japanese /a/ compared to AmE /æ/ and is somewhat longer than Japanese /a/. Finally, English /ɑ/ is spectrally distant from either Japanese /e/ or /a/, and its duration is again intermediate between Japanese long and short vowels. Therefore, the AmE vowels exhibited various spectral distance from the Japanese vowels in the F1-F2 space, while their duration was uniformly intermediate between Japanese phonologically long and short vowels.

In order to quantitatively investigate the response data, LME models were fitted to by-participant results of logistic regression analysis using the *lme4* and *lmerTest* packages in R. The dependent variable was logistic regression coefficient for each of the three acoustic dimensions, and the independent variable was response category where Japanese /a/ was the reference. The models included random intercepts for participant and session order. As for the F1 coefficient, Japanese /e/ and English /ɛ, ɑ/ were significantly lower from Japanese /a/ (p = .013, p < .001, p = .025, respectively),[12] This indicates that the vowels were represented as higher (i.e., more raised) than the reference vowel /a/. As for the F2 coefficient, Japanese /e, ee/ and English /æ/ were significantly higher than Japanese /a/ (*p*s < .001), indicating that the vowels were represented as more fronted than /a/. In contrast, the F2 of English /ɑ/ was significantly lower than that of Japanese /a/ (p < .001), indicating that the vowel was represented as more back than /a/. As for the duration coefficient, all vowels except Japanese /e/ were significantly longer than Japanese /a/ (*p*s < .01). Additional LME models revealed that the AmE categories were not significantly different from each other in terms of perceived duration and that they were significantly shorter than Japanese phonologically long vowels, confirming that all four AmE vowels had an intermediate length between Japanese long and short vowels.

[12] Japanese /ee/ also approached the significance level (p = .056).

The following inferences can be made from the obtained results. First, speculating from the F1-F2 proximity, AmE /ɛ/ is spectrally assimilated to Japanese mid front /e/ while English /ʌ/ is assimilated to Japanese low central /a/. It is worth noting that neither F1 nor F2 coefficients of /ʌ/ were significantly different from those of Japanese /a/, suggesting that the two sounds were not perceptually distinguished. AmE /æ/ is represented as a fronted version of Japanese /a/ because its F2 coefficient was significantly higher than that of /a/. This implies that a new category had been formed for the L2 sound based on the F2 cue or vowel backness. The location of English /ɑ/ in the F1-F2 space is somewhat unexpected because the vowel is reported to be spectrally assimilated to Japanese /a/. However, this can be explained as a result of orthographic influences. Since /ɑ/ is often spelled as "o" in AmE (e.g., *boss*, *not*, *lot*), Japanese learners of English may associate it with the Japanese mid back rounded /o/. The association may be further strengthened by the acoustic similarities of the vowels in F3 (i.e., roundedness), as AmE /ɑ/ is typically produced as low back rounded [ɒ]. Finally, the participants did not rely on durational cues in discriminating the AmE vowels, at least at a statistically significant level. This indicates that the L2 vowels were free from temporal assimilation to Japanese phonological length distinctions.

4.4.4 Discussion

4.4.4.1 Interim summary

This study explored the process of new L2 category formation in Japanese listeners' perception of L2 AmE vowels /ɛ, æ, ʌ, ɑ/. The primary focus was on AmE /æ/, which is reported to be a poor exemplar of Japanese /a/ and therefore is subject to new category formation according to SLM and PAM(-L2). Two versions of L2LP, namely segmental and featural versions, were presented to account for new category formation for the L2 sound. While both types of modeling predicted similar outcomes, the featural version provided an ecologically more valid account because it could represent how an L2 new category would gradually emerge by reorganizing the existing L1 features, especially when the L2 sound comprises a feature combination that is not allowed in the L1 and therefore sounds deviant (e.g., "*/low, front/"). The experiment found that AmE /æ/ was perceptually represented as a fronted version of Japanese low central /a/, indicating that a new category had been formed for the L2 sound based on the vowel backness feature. The result thus aligns with the featural account that Japanese listeners would notice the /low/ and /front/ features in /æ/ and learn to combine them. However, neither segmental nor featural modeling could predict the perception of /ɑ/, which was supposedly assimilated to Japanese /o/ potentially due to the effects of orthography.

4.4.4.2 Segmental and featural accounts

The primary aim of the current study was to compare the segmental and featural accounts of NEW L2 category formation based on L2LP. While both types of modeling yielded similar learning outcomes, the processes of each model reveal important theoretical differences between the two.

Two problems were identified for the segmental model of new category formation. One is that a new segment has to be intentionally added to the grammar by the modeler. The act of adding a new category could be justified because it is possible for real learners to introduce new categories to their grammar by explicit learning about the sound categories (e.g., through textbooks) or by incidental noticing of the lexical-semantic distinctions signaled by the sound contrast (e.g., noticing the difference between *bass* and *bus*). However, it is unclear when and for which L2 sound a new category should be added. This is related to the second problem that a segmental grammar does not represent perceived L1-L2 deviance. While a researcher could refer to the reported perceived deviance of L2 categories to conduct plausible modeling (e.g., a new category should be added for AmE /æ/ that is reportedly deviant to Japanese listeners), the model itself does not incorporate such deviance. This is because the segmental grammar, after all, perceives only segments. That is, no matter poor an exemplar a particular sound is, the grammar has to classify it as one of the available segments in the current grammar. For example, the segmental L1 Japanese grammar perceives both AmE /ɛ/-like sound (e.g., [F1 = 700 mel, F2 = 1400 mel]) and /æ/-like sound (e.g., [F1 = 800 mel, F2 = 1400 mel]) as Japanese /e/ most of the time, despite their difference in F1. This is because the grammar prioritizes the F2 cue over the F1 cue because the former is a more reliable cue for distinguishing Japanese /e/ and /a/ within the relevant perceptual space. Thus, AmE /ɛ/ and /æ/ are perceived as the 'same' L1 segment /e/ equally frequently, even though the former is considered a better exemplar of Japanese /e/ in terms of vowel height.

The featural model does not suffer from either of the above two problems. A strength of the featural approach is that it can model the relationship between perceived L1-L2 deviance and new L2 category formation. The featural L1 Japanese grammar perceives an AmE /æ/-like sound (e.g., [F1 = 800 mel, F2 = 1400 mel]) as consisting of /low/ and /front/ features, which do not occur simultaneously in Japanese. The listener would perceive the sound as /mid, front/ (/e/) because of the feature co-occurrence constraint "*/low, front/" is ranked very high in the L1-appropriate perception grammar. This essentially represents the perceptual deviance of AmE /æ/: it is too low for Japanese /e/ and too back for Japanese /a/. The notion is congruent with Flege's (1995) claim that features used to distinguish L1 sounds can probably not be freely recombined to perceive and produce L2 sounds. Then, new category formation for AmE /æ/ can be modeled as a result of the weakening of the "*/low, front/" constraint. With increased exposure to L2 input, the featural Japanese learner of English learns to combine the existing /low/ and /front/ features, which does occur simultaneously and meaningfully in AmE. In this way, a new category can gradually emerge as a result of feature reorganization without manipulating the features, unlike in segmental modeling.

The experiment in the current study found that the Japanese learners of English could distinguish AmE /æ/ from Japanese /a/ on the basis of F2. This indicates that a new category had been formed for the L2 sound based on the vowel backness feature. The new category can thus be expressed as /low, front/ because /æ/ was perceptually represented as a fronted version of Japanese /a/ (/low, central/). The finding aligns with the featural account that the initially deviant feature combination /low, front/ becomes permissible as a result of L2 learning. While the result does not disprove the segmental account that /æ/ is created by splitting or redistributing Japanese /a/, it would remain unexplained why only the F2 dimension was involved. Therefore, the featural account provides an ecologically more valid account of new L2 category formation. In the next section, I will discuss how the incorporation of features would benefit not only L2LP but also other models of L2 perception.

4.4.4.3 Implications for L2LP and other models

The majority of previous studies under L2LP have adopted the segmental approach. Escudero (2005, p. 48) explicitly commented on the issue of segmental vs. featural approaches: “[t]he reader may wonder why using segmental units, such as /i/, should be preferred over a combination of features, such as /high, front/, for describing the constraints in an adult perception grammar.” She expressed her concern that feature-based modeling with feature co-occurrence constraints such as “*/high, front/” would considerably complicate the model and it might not work anyway. However, the current study has demonstrated that featural modeling does work, possibly with a fewer number of constraints than segmental modeling (featural = 84, segmental = 160). Escudero also proposed that perceptual cue integration could be modeled with constraints such as “[F1 = 300 Hz] is not /high, front/,” which could just as well be abbreviated to “[F1 = 300 Hz] is not /i/”. This is only true if perceptual mapping and cue integration occur simultaneously as in multi-dimensional auditory-to-feature and auditory-to-segment constraints (cf. Figure 3-7 and Figure 3-8). Instead, the featural approach adopted in this study instead assumes one-dimensional mappings where cue integration occurs separately from perceptual mapping (cf. Figure 3-6). The segmental and featural approaches can thus be contrasted as in Figure 4-20.

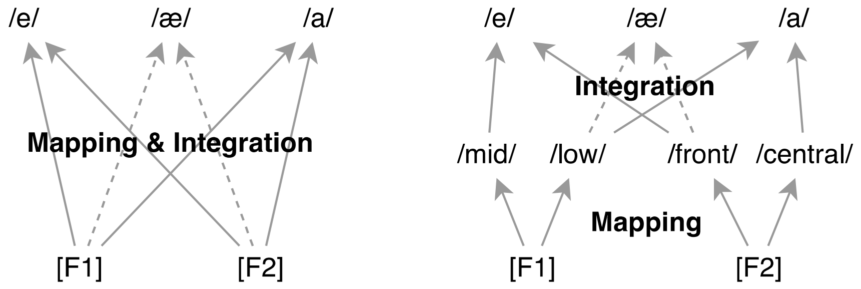


Figure 4-20. Segmental (left) vs. featural (right) approaches.

There are two theoretical advantages in adopting the featural approach. First, the relationship between auditory dimensions and linguistic representations is phonetically faithful in the featural approach, whereas in the segmental approach it is completely arbitrary. The segmental approach would allow perverse-sounding constraints such as "[VOT = 10 ms] is not /i/" because any auditory dimension, in principle, can map to any linguistic representation. In contrast, the featural approach would allow only phonetically relevant constraints such as "[VOT = 10 ms] is not /voiceless/." This explains why the featural model, but not the segmental model, can express perceived phonetic deviance. The segmental grammar can only perceive segments regardless of how the cues are used. Contrarily, the featural grammar maps auditory dimensions to relevant features through cue constraints, while these features are not always faithfully integrated due to structural constraints, thus representing perceived deviance.

Second, the separation of perceptual mapping and cue integration would allow a stricter distinction between the perceptual and representational tasks in L2LP. For example, in segmental modeling, new perceptual mappings suddenly emerge as soon as the new sound representation is added to the grammar. In other words, the perceptual task occurs with the representational task. On the other hand, the perceptual and representational tasks occur independently in the featural modeling, in which the auditory-to-feature constraints are responsible for learning perceptual mappings, and the feature co-occurrence constraints are responsible for acquiring new feature combinations or sound representations. The featural approach thus allows modeling of cases in which the representational task is complete (e.g., the learner knows that there is a vowel category /æ/) but the perceptual task is still in progress (e.g., the perceptual mappings are nonnative-like and being developed), which is realistic (cf. Study 2).

Although L2LP has mostly adopted the segmental approach, it should be noted that the featural approach is nothing new to the model. For example, Escudero and Boersma (2004) explained that Spanish listeners develop a length-based distinction between /iː/ and /ɪ/ in L2 SBE, which is not seen either in the L1 or in the L2, because they split the L1 /i/ category on the blank-slate duration dimension, i.e., /iː/ = /i, long/ and /ɪ/ = /i, short/. Thus, they modeled the non-previously-categorized NEW scenario with the help of the length features, namely /long/ and /short/. Besides, the segmental approach can be seen as a simplified version of the featural approach, in which the two processes of abstraction – perceptual mapping and cue integration – are handled simultaneously. This is why the segmental and featural simulations in the current study yielded comparable results, although their processes are distinctive. Therefore, L2LP is amenable to a featural account, which is expected to help the model to explain various other scenarios in greater detail.

The experimental result of the current study runs contrary to L2LP's prediction that the NEW scenario is the most difficult because the Japanese listeners who had no overseas experience had successfully established a new sound representation and new mappings for /æ/. Given that the listeners were most likely less proficient than those in Study 1 who could not achieve L2 optimal perception, it seems that this particular NEW scenario is not more difficult than the SIMILAR scenario. However, it is also true that a NEW scenario can be extremely difficult, as is the case for Japanese listeners' acquisition of English /r/-/l/. A feature-based account is useful for explaining this discrepancy. I propose that the relative difficulty of a NEW scenario is dependent on whether and how each feature is utilized in the L1. For example, a NEW scenario with non-previously-categorized dimensions is expected to be very difficult because learners have to create completely new features on the blank-slate auditory dimension. Japanese listeners' acquisition of English /r/-/l/ applies to this case. Even though Japanese listeners may be able to achieve a categorical distinction between these sounds (MacKain et al., 1981), they may fail to utilize the appropriate F3 dimension and instead rely on inappropriate cues such as F2 (Iverson et al., 2003) because the F3 is not informative for segmental contrasts in Japanese. In contrast, no additional feature may be necessary for a NEW scenario involving already-acquired dimensions, which was the case for the current study. In such a case, the L1 features can be reused and reorganized to achieve optimal L2 perception, which is expected not to be very difficult. In addition, there can be a case in which a new feature must be added to already-categorized dimensions. For example, native listeners of Arabic are considered to have only two vowel height features /high/ and /low/, which are insufficient for perceiving Japanese vowels that contrast in /high/, /mid/, and /low/ height features. It can be hypothesized that Arabic learners of Japanese would initially rely on their L1 /high/ and /low/ features, but eventually notice that these features are insufficient for L2 Japanese and therefore create a new /mid/ feature along the already-categorized dimension of F1. This scenario is considered to be more difficult than the already-categorized NEW scenario in which the existing features can be simply reused, though not as difficult as the non-previously-categorized NEW scenario.

Finally, SLM and PAM(-L2) would also benefit from the featural approach. As for SLM, the use of features may help better define the degree of perceived phonetic dissimilarities between L1 and L2 sounds, which plays a crucial role in the model's predictions. Regarding the current learning scenario, SLM predicted that Japanese listeners would discern at least some of the phonetic differences between AmE /æ/ from Japanese /a/, resulting in new category formation. Based on the experimental result, the learners are considered to have noticed the distance of AmE /æ/ from Japanese /a/ in the vowel backness feature but not in the height feature. Their representation of /æ/ may thus be different from that of monolingual AmE listeners, in accordance with H6. The perceptual assimilation patterns in PAM(-L2) can also be elaborated by incorporating the notion of features. For example, the model would explain Japanese listeners' perception of AmE /æ/-/ʌ/ as a CG or UC assimilation pattern, in which the former is perceived as a more deviant exemplar of Japanese /a/ or perhaps is uncategorized. By adopting the featural approach, it can be explained that /æ/ is perceived as a deviant exemplar of Japanese /a/ due to its unusually fronted gestural feature. Thus, a new phonetic and phonological category is predicted to be formed for /æ/ based on the F2 dimension only, while no new category formation is expected to occur for the better exemplar /ʌ/. These predictions align well with the experimental result.

4.4.4.4 Limitations and future directions

Some limitations need to be addressed when interpreting the current study's findings. First, the simulations used steady-state F1 and F2 cues only, ignoring other potentially relevant cues such as duration and VISC. In theory, it is possible that Japanese listeners assimilate phonetically long /æ, ɑ/ and short /ɛ, ʌ/ in AmE to phonologically long /ee, aa/ and short /e, a/ in Japanese, respectively. However, the experimental result found no such temporal assimilation. The exclusion of duration from the simulations is thus justified, although it still needs to be explored why the listeners were free from temporal assimilation, unlike the SIMILAR scenario in Study 1. Also, AmE /æ/ can be realized as "tense æ" with centering VISC in some dialects of AmE (Labov et al., 2006), which was not considered in either the simulations and the experiment. Future research could address whether and how dynamic spectral cues affect Japanese listeners' perception. Second, it has been assumed that F1 and F2 cues are equally used. However, one of the cues may be weighted more heavily than the other. For example, young infants were found to utilize the F1 cue over the F2 cue when categorizing CE vowels, possibly because the former has more acoustic energy and thus is acoustically more salient than the latter (Curtin et al., 2009). While adults' perception is expected to be different from infants', it is worth investigating whether the perceptual salience of each cue relates to the process of new category formation. Third, the experimental result revealed that orthographic factors need to be considered because AmE /ɑ/ as in *boss* was presumably assimilated to Japanese /o/ rather than acoustically more proximal /a/. The simulations, whether segmental or featural, could not predict this assimilation pattern because the focus was on [auditory] → /surface/ mappings only. Higher-level representations need to be included to adequately account for the current learning scenario.

Another important avenue for future research is to investigate the effect of L2 proficiency. While the experimental result confirmed that the Japanese learners of English had succeeded in establishing a new category for AmE /æ/, their L2 perception has room for further development because none of the participants had any overseas experience. Featural simulations would predict that, even if the learner has established a new L2 sound representation by reorganizing existing features, their perceptual mappings may still be nonnative-like if the reused features remain L1-like. In contrast, segmental simulations would predict that an almost optimal, native-like mapping emerges as soon as a new sound representation is introduced to the grammar. A follow-up study with Japanese listeners with a higher level of proficiency (e.g., 'returnees' from the US) could help test these predictions.

4.4.5 Summary

Study 3 demonstrated the usefulness of feature-based modeling to account for Japanese listeners' acquisition of a NEW AmE sound /æ/, a deviant exemplar of L1 categories. The experiment found that the vowel was perceptually represented as 'fronted /a/,' indicating that a new category had been formed for the L2 sound. While the segmental model had difficulty in explaining the relevance of L1-L2 perceptual deviance to L2 category formation, the featural model provided a coherent explanation that AmE /æ/ sounds deviant because it consists of /low/ and /front/ features that do not occur simultaneously in Japanese and that these features can be reorganized to form a new category in the L2. However, the result runs contrary to L2LP's prediction that a NEW scenario is the most difficult because the Japanese learners of English, who were not necessarily highly proficient in the L2, could successfully establish a new category for the L2 sound. It is thus proposed that the relative difficulty of a NEW scenario depends on whether and how the relevant features are utilized in the L1.

Chapter 5: General discussion

5.1 Introduction

The three case studies presented in Chapter 4 aimed to test L2LP's predictions per and across scenarios (SIMILAR, SUBSET, and NEW). Computational simulations based on the model yielded very detailed predictions for the scenarios, which were generally supported by the experimental results. The findings of each case study, such as language-specific perception modes and feature-based perception, are unique and have useful implications on their own. Thus, as an overall evaluation, L2LP can be said to be a useful and fruitful model of L2 perception. However, some of the model's predictions were not borne out, especially regarding the relative difficulty of acquisition. This draws attention to a necessity to amend the current model as well as to consult other models of L2 perception.

This chapter presents a discussion of the results of the case studies. I first discuss the implications of the case studies from both theoretical and pedagogical perspectives in Section 5.2. Section 5.3 then discusses the potential limitations of the model, with a particular focus on the difficulty of acquisition across three scenarios, in relation to other models such as SLM and PAM(-L2). The subsequent Section 5.4 discusses how the current limitations should be addressed in future research to extend L2LP. Finally, Section 5.5 provides a summary of the chapter.

5.2 Implications

5.2.1 Implications for research

5.2.1.1 Overall results

The overall result of the case studies confirmed that L2LP is capable of providing very specific and testable predictions, which the other L2 perception models currently lack. This owes mainly to two properties of L2LP: the separation of perceptual mappings and sound representation and the incorporation of computational simulations.

L2LP claims that the separation of perceptual mappings and sound representations is crucial for adequately describing, explaining, and predicting L2 perception. Conversely, the two factors seem to be conflated in SLM and PAM(-L2), in which sound categories are seen to perform the mapping of the speech signal (Escudero, 2005, p. 131). The three case studies suggest that they should indeed be distinguished. First, L2LP best explains the result of Study 1 among the three models because perceptual mappings and sound representations are seen to act independently in the model. In L2LP, the L1 Japanese contrast /ii-/i/ is representationally equated with the L2 AmE contrast /iː/-/ɪ/, while the perceptual mapping patterns can change as a function of language context. In contrast, SLM and PAM(-L2) are compelled to explain the observed shift in cue weighting as a result of new category formation because a change in perceptual mapping is associated with a different category. However, this explanation is at odds with the models' prediction that the L1 and L2 contrasts should be subject to equivalence classification (SLM) or assimilation (PAM). Second, L2LP is the only model that considers the possibility of the SUBSET scenario as in Study 2, in which an L2 representation can map to more than one L1 representation (MCA pattern). The MCA pattern seems to be overlooked in SLM and PAM(-L2) because they focus on the similarities between L1 and L2 categories (i.e., representations) rather than the specific mapping patterns. Thus, the conflation of representations and mappings obscures the potential problems associated with MCA, namely the establishment of spurious lexical contrasts and perception of unnecessary categories. Last, the feature-based implementation of L2LP in Study 3 enabled formal modeling of perceived L1-L2 phonetic deviance and its relation to the process of new L2 category formation. This is again because perceptual mappings (e.g., [F1] → /low/, /mid/, [F2] → /front/, /central/) and representations (e.g., /mid, front/, /low, central/) are expressed separately in L2LP. On the other hand, the perceived phonetic similarity is defined rather vaguely in SLM and PAM(-L2) because the categories themselves are seen as 'similar' or 'dissimilar' to other categories without taking into account the specific mapping patterns. For these reasons, perceptual mappings and sound representations should be strictly distinguished as in L2LP.

Another important property of L2LP is its incorporation of computational simulations. Recall the distinction between a model and a simulation: a model is a simplified representation of a system of interest for understanding, and a simulation is the operation of the model that is useful for making predictions as well as for evaluating the model. Based on the results of the case studies, it can be said that L2LP's simulations based on Stochastic OT and the GLA made the model's predictions more specific and detailed than any other model's. For example, the graphical representation of the simulated perception in Study 1 showed a close resemblance to the real learners' perception, which was directly comparable numerically (e.g., simulated cue weighting = 0.83, cue weighting of Participant 23 = 0.95). The simulations in Study 2 predicted real AusE listeners' nonnative perception patterns accurately, in which Japanese /ii/ was mostly perceived as AusE /iː/ (simulated frequency = 78.1%, observed frequency = 61.0%), while Japanese /i/ was perceived as ambiguous between AusE /iː/ (simulated = 46.4% , observed = 40.8%) and /ɪ/ (simulated = 51.5% , observed = 43.2%), with AusE /ɪə/ being seldom chosen as a response (less than 15% in both the simulation and the experiment). The simulations are not only precise but also amenable, as is illustrated in Study 3 where the feature-based modeling in Boersma and Chládková (2011) was implemented and compared with the conventional segment-based modeling.

The simulations have also served as objective testing of the model. The case studies found that the simulated results were mostly congruent with the model's theoretical predictions, except for the SUBSET scenario in Study 2. For this scenario, the model hypothesized that there would be a representational task of reducing the number of categories. However, the simulations predicted that AusE listeners would not perceive the 'extraneous' AusE /ɪə/ category out of the Japanese vowels from the initial state, and therefore there would be no representational task. In addition, the simulated L2 developmental pattern was somewhat different from the model's theoretical prediction because the virtual learner arrived at the end state by incompletely achieving the perceptual task. While one may consider the above discrepancy as a failure of the model, I consider that this is rather a strength because it suggests that the model and the simulations do not conspire to predict the same outcome, unlike SLM and PAM(-L2) that are subject to HARKing (cf. Section 4.2.4.3). Thus, it has been shown that the computational simulations alone can be used to test the theoretical components of the model. To summarize, the computational simulations can not only provide a good estimate of L2 perception patterns for various learning scenarios and but also serve as an objective self-test of the model, which the previous L2 perception models may need in order to overcome their alleged lack of concreteness.

5.2.1.2 Results per case study

The three case studies also provide distinct implications for L2 perception research, which are discussed per scenario below. First, Study 1 found that language mode affects L2 listeners' online speech perception, to which L2 researchers should pay dedicated attention. Grosjean (2001) notes that language mode can be an independent, control or confounding variable and hence needs to be heeded at all times, even if it is not the main variable being investigated. Language mode can be influenced by a number of factors, including the person(s) being spoken or listened to (e.g., language proficiency, language mixing habits and attitudes, usual mode of interaction, kinship relation, and socioeconomic status), the situation (e.g., physical location, presence of monolinguals, and degree of formality and of intimacy), the form and content of the message being uttered or listened to (e.g., language used, topic, type of vocabulary needed, and amount of mixed language), the function of the language act (e.g., to communicate information, to request something, to create a social distance between the speakers, to exclude someone, and to take part in an experiment, etc.), and specific research factors (e.g., the aims of the study taking place, the type and organization of the stimuli, and the task used). While it would be impossible to control for all those factors, L2 researchers should make careful attempts to put bilinguals on the preferred position along the language mode continuum.

For instance, when testing a bilingual's perception in one of their languages, the language setting (including the instructions, stimuli, and the experimenter) should be completely monolingual in the language of investigation. If the experimental goal is to elicit L1-L2 intermediate perception, then the language settings should involve mixed language. Unfortunately, many past studies have not taken thorough measures to sufficiently control for language contexts, making language mode as a potential confounding variable. For example, a highly proficient L2 learner's performance may become more L1-like than the actual competence when the L2 mode is not fully activated. Without controlling for language mode, it would be uncertain whether the learner's performance is affected by the L1 or by the experimental condition, or both.

Second, the follow-up simulation in Study 2 found that native-like perceptual mappings are not a prerequisite for accurate sound categorization. Other attested cases support this proposition, such as Japanese listeners who manage to establish native-like sound representations of English /r/ and /l/ based on inappropriate cues such as F2 (Hattori & Iverson, 2009; Iverson et al., 2003). This reminds the previously discussed importance of separating perceptual mappings and sound representations, which is important not only theoretically but also methodologically. For example, to test whether Japanese listeners can acquire a native-like perception of English /r/ and /l/, it would be insufficient to examine correct classification rates of naturally produced /r/ and /l/ tokens because then it would not be known how the classification is being done. Instead, an intricate manipulation of synthetic or resynthesized stimuli (as in Studies 1 and 3) or detailed acoustic analysis of natural stimuli (as in Study 2) is advisable. While L2LP claims that the separation of perceptual mappings and sound representations is crucial, it must be noted that the model does not seem to distinguish the two components in its optimal perception hypothesis. This may have caused the model to yield inaccurate predictions concerning the relative levels of difficulty across scenarios, which is discussed further in Section 5.3.1.

Lastly and perhaps most importantly, Study 3 revealed the usefulness of feature-based modeling, which is readily applicable to other types of learning scenarios. For example, the language-specific perception modes in Study 1 can be explained as the act of length and tenseness features in Japanese and AmE. Using the feature-based approach, the Japanese /ii/-/i/ contrast can be represented as /high, front, short/ and /high, front, long/, whereas the AmE /iː/-/ɪ/ contrast can be represented as /high, front, tense/ and /high, front, lax/. It follows that Japanese /ii/ and AmE /iː/ as well as Japanese /i/ and AmE /ɪ/ are associated with each other by the shared /high/ and /front/ features, which might explain why AmE/ɪ/ is not perceived as Japanese /e/ despite the acoustic proximity.

Japanese learners of English are expected to initially utilize the inappropriate length feature to distinguish the two L2 vowels, though they may learn to disuse it and acquire the new tenseness feature. The observed shift in cue weighting between language contexts can be explained as different levels of activation of the length (L1 mode) and tenseness features (L2 mode), although the L1 feature is considered to be more robust than the newly acquired tenseness feature. As for Study 2, it was suggested that the length feature is (pseudo-)phonemic in AusE, which explains AusE listeners' sensitivity to nonnative phonemic length (cf. the feature hypothesis). The learning task for AusE learners of Japanese, then, is to adjust the boundary between the L1 /long/ and /short/ features to accommodate to the L2 production environment. While the perception of VISC cues might be represented by such features as /closing/ and /opening/, /wide/ and /narrow/, and /falling/ and /rising/ (McArthur & McArthur, 2005), adjustment of these features is of secondary importance for the learning scenario because they are most likely not used in the perception of Japanese vowels with little or no VISC. In this way, featural accounts can provide new insights on various types of learning scenarios that have previously been investigated with the conventional segment-based approach.

There are several caveats one should keep in mind before applying the featural approach to other L2 learning scenarios. First of all, features do not necessarily have to be specified in a binary manner (i.e., [±]) as in traditional phonology. For example, Chládková, Escudero, and Lipski (2015) investigated Dutch listeners' neural sensitivity to vowel duration in native /aː/ and /ɑ/ as in the word *maan* 'moon' and *man* 'man,' which constitute a phonological length contrast. They found that durational changes in the stimuli evoked larger MMN for /aː/ than for /ɑ/, indicating that duration is phonemically relevant for the *maan* vowel that is represented as "long," while it is not phonemically specified for the *man* vowel. The result suggests that listeners do not necessarily encode the durational distinction as a binary [±long] value. Kawahara and Braver (2013) investigated Japanese speakers' emphatic vowel lengthening and found that the speakers could manifest a maximum of six levels of distinction in duration as a function of the level of emphasis, e.g., /itai/ 'aching' (no emphasis), /itaai/ (level 1 emphasis), /itaaai/ (level 2 emphasis), /itaaaai/ (level 3 emphasis), /itaaaaai/ (level 4 emphasis), and /itaaaaaai/ (level 5 emphasis). Traditional binary features would not suffice to describe such phenomena. Instead, univalent features in more recent phonology (Gussenhoven & Jacobs, 2017) should be used. Second of all, the phonetic property of a feature is considered language-specific. For example, Tsukada's (2012) finding that Japanese and Arabic listeners were less accurate in discriminating vowel length in the "unknown" language than in the "known" language indicates that Japanese and Arabic have the same type of length features with different featural boundaries. That is, what is perceived as “long” in Japanese is not necessarily also “long” in Arabic, and vice versa. This is natural because the phonetic property of a phoneme, which consists of features, are also language-specific. However, the idea challenges the widespread assumption that having the same feature in the L1 and the L2 is always equally advantageous for L2 learning. The language-specific featural boundaries may also change as a result of L2 learning, which is expected to generalize to multiple phonemes sharing the feature. This is exactly what Olson (2019) found, in which phonetic training on one of three voiceless stops in L2 Spanish /p, t, k/ affected L1 English speakers’ production of untrained phonemes, suggesting a global shift of the boundary between “voiced” and “voiceless” features. Last of all, a few other caveats mentioned in Flege (1995) are worth readdressing here. First, L1 features are not freely recombined to form a new L2 category, which is compatible with the notion of feature-cooccurrence constraints in Study 3. Second, certain features may inherently be weighted more heavily than others, such as the length feature being weighted more heavily than the height and backness features (Bohn, 1995). Third, features may be evaluated differently depending on the position in the syllable or the frequency of occurrence, which inherits from SLM’s H1. All these points, among others, would need to be carefully considered when adopting the featural approach.

5.2.2 Implications for education

5.2.2.1 Overall results

While the present thesis has focused mainly on the theoretical aspects of L2LP, the model has the potential to contribute to the field of L2 education as well. Of particular relevance are the model's emphasis on the role of input and its view of perceptual learning as distributional and meaning-driven.

L2LP emphasizes that the role of input is of prime importance for L2 learning. The model claims that L2 learners' perceptual behavior is predictable from the acoustic properties of the input they receive. This suggests that a large amount of native-like input is required for the acquisition of L2 optimal perception. The simulated learners in the present study were trained in a very rich learning environment where they had access to such input. However, this is not always the case for real learners, particularly for those in classroom settings. For example, phonetic input to learners of English as a foreign language (EFL) in Japan tends to be impoverished because the majority of the teachers are native speakers of Japanese who themselves are not necessarily fluent in English[13] and thus the input is most likely Japanese-accented. Flege and Eefting (1987, p. 81)

[13] According to the 2018 survey of the Japanese Ministry of Education, Culture, Sports, Science, and Technology (MEXT), less than 40% of junior high school teachers had passed the B2 level of the Common European Framework of Reference for Languages (CEFR).

suggest that such “non-authentic L2 input” may result in “incomplete approximation to L2 phonetic norms.” Using L2LP-based simulations, it can be shown that virtual learners do not acquire L2 optimal perception if the input is heavily foreign-accented even when the plasticity is high (i.e., in childhood). This raises a serious question on the effectiveness of early EFL education that is currently under implementation in Japanese education system[14]. The effect of the ongoing curriculum reform will likely be limited at best unless more proper attention is paid to the quality and quantity of input the learners actually receive. There is no guarantee that a simple lowering of AOL would benefit the learners (Harada, 2011). Also, a growing number of secondary and tertiary institutions in Japan have started adopting English immersion programs, where students learn English as a medium of instruction. However, careful consideration is needed for assessing such programs, especially in the context of Japan where Japanese is used as the sole de facto official language; English, after all, is a foreign language. L2LP-based simulations are useful for estimating the effectiveness of various L2 learning settings including immersion. A great strength of a simulation is that the model can be reconfigured and experimented with, which is usually impossible, too expensive, or impractical in reality (Maria, 1997). For example, a comparison can be made between the traditional EFL

[14] In 2011, English instruction became compulsory starting in the fifth grade. It is also planned to make English activity classes mandatory for third- and fourth-grade students by 2020.

setting and the immersion setting by manipulating the quality and quantity of input (e.g., “traditional setting” with 50% native Japanese, 40% Japanese-accented English, and 10% native English input vs. ”immersion setting’” with 60% native English and 40% Japanese-accented English input). The effect of learners’ age can also be investigated by manipulating the plasticity values (e.g., “immersion setting” from the age of 13 vs. from the age of 19). Although such simulations were beyond the scope of the present study, L2 education researchers may find L2LP quite useful and insightful.

L2LP models perceptual learning as a distributional and meaning-driven phenomenon, which is relevant to input enhancement in L2 education. According to Smith (1993), instructional input can be manipulated to create either positive or negative input enhancement. Positive input enhancement is intended to make certain forms more salient in the input, while negative input enhancement is intended to mark certain forms as incorrect. As for speech perception, it has been empirically shown that enhanced acoustic cues improve adult learners’ categorization of L2 sound categories. For example, Escudero, Benders, and Wanrooij (2011) trained adult Spanish learners of Dutch on a natural bimodal or an enhanced bimodal distribution of Dutch /ɑ/ and /aː/, with the average productions of the vowels or more extreme values as the endpoints respectively. Categorization improved for learners who listened to the enhanced distribution but not for those who listened to the average productions, suggesting that adults benefit from enhanced distributional learning. L2LP's simulations can demonstrate the effect of such positive input enhancement by manipulating the distributions of the acoustic input. The effect of negative input enhancement on perceptual learning has also been attested. Goudbeek, Swingley, and Smits (2009) tested adult listeners' learning of a category distinction within a psychophysical space with or without supervision (i.e., corrective feedback). Supervision proved beneficial, especially for maintaining category learning beyond the learning phase. The importance of corrective feedback is implicitly assumed in L2LP's error-driven learning mechanism, the GLA, in which the learner has direct access to the correct form based on the semantic context. However, real learners do not always have such access and therefore need to notice their perceptual error in some way in order to make the input intake (Coder, 1967; Gass, 1988). This explains why form-focused instructions (Ellis, 2001), in which attention is paid to language form while maintaining an instructional emphasis on meaning, is very effective (Lyster & Saito, 2010) because it is considered to facilitate the meaning-driven learning as modeled in the GLA. L2LP is the only model that can account for the effects of positive and negative input enhancement in L2 perceptual learning, which may provide new insights for researchers and practitioners engaging in L2 education.

5.2.2.2 Results per case study

The three case studies have distinct pedagogical implications on their own, which I present per scenario below. First, the effect of perception modes found in Study 1 calls for a necessity to control for language modes in L2 teaching settings. For example, the majority of EFL instructions in Japanese secondary and tertiary institutions are conducted in Japanese[15], which would place the learners close to the L1 monolingual mode. For better learning, they should instead be placed in the L1-L2 bilingual mode or the L2 monolingual mode, depending on their levels of competence and confidence. While it may be impractical for all EFL teachers to provide native-like input, they can certainly attempt to use as much English as possible to elicit the learners' L2 mode. Also, it should be acknowledged that immersion does not necessarily ensure that learners are placed in the monolingual L2 mode, as proficient bilinguals often mixed their languages. While such language use is not 'inappropriate,' if the purpose of immersion is to improve L2 proficiency, and if there are other speakers who do not share the same L1, then the use of mixed language should be avoided at least in the classroom. L2 teachers should be aware of the importance of controlling for the language context in order to provide an effective and fair learning environment to the learners.

[15] According to the 2018 survey of MEXT, less than 20% of English teachers in Japanese junior high schools and high schools "speak in English most of the time during class."

Next, the difference between accurate categorization and optimal perceptual mappings found in Study 2 is closely related to the distinction between intelligibility, comprehensibility, and accentedness in SLA. Derwing and Munro (2005) define intelligibility as the extent to which the speaker's intended utterance is actually understood by a listener, comprehensibility as a listener's perception of how difficult it is to understand an utterance, and accentedness as a listener's perception of how different a speaker's accent is from that of the L1 community. While it is often assumed that greater accentedness always entails reduced intelligibility and comprehensibility, they are in fact partially independent. Specifically, while listeners who find a specific L2 utterance to be unintelligible and incomprehensible always perceive it as heavily accented, they often assign good intelligibility and comprehensibility ratings to utterances that they have also rated as heavily accented (Munro & Derwing, 1995). Given that the primary goal of linguistic communication is to understand and to be understood, L2 teachers are advised to prioritize intelligibility and comprehensibility over accent-free speech. What this means for speech perception is that, as long as the listener can identify what is intended by the speaker, native-like perceptual behavior is not always necessary. In other words, accurate sound categorization may be prioritized over native-like perceptual mappings. This point is revisited in Section 5.3.1.

Finally, the relevance of features in L2 perception found in Study 3 sheds new light on the effect of input enhancement. Escudero et al.'s (2011) finding that enhanced acoustic distributions improve speech categorization suggests that enhanced features such as exaggerated vowel height and backness facilitate perceptual learning. It follows that, given that featural changes generalize to related categories (Olson, 2019), enhancement of a certain feature may facilitate the learning of multiple categories sharing the feature. For example, enhancement of the vowel length feature in Japanese /ii/ vs. /i/ may help L2 learners of Japanese to acquire other length contrasts such as /aa/ vs. /a/ and /ee/ vs. /e/. It may even help the acquisition of the singleton-geminate contrasts in Japanese (e.g., *kata* 'shoulder' – *katta* 'won'), as the length feature seems to be related across vowels and consonants (Pajak & Levy, 2014). Conversely, if a certain feature is found to be problematic for L2 learners, the feature may need to be enhanced across multiple categories. Input feature enhancement is commonly seen in infant-directed speech (Kalashnikova et al., 2017) and foreign-accented speech (Scarborough et al., 2007; Uther et al., 2007) as a way of facilitating speaker-hearer interactions and language acquisition, which have useful implications for L2 education.

5.3 Limitations

5.3.1 Difficulty of acquisition

The most significant limitation of L2LP turned out to be its inaccurate prediction of the relative levels of difficulty of the learning scenarios. L2LP claims that the difficulty of acquiring L2 optimal perception depends on the number and nature of learning tasks involved. Specifically, the SIMILAR scenario is expected to be the least difficult because there is only one perceptual task (boundary adjustment); the SUBSET scenario is of intermediate difficulty because there are one perceptual task (boundary adjustment) and one representational task (category reduction); the NEW scenario is the most difficult because there are two perceptual tasks (creating new mappings and cue integration) and one representational task (category creation). Thus, the predicted order of difficulty is SIMILAR < SUBSET < NEW. However, the three case studies found that obtaining native-like cue weighting for SIMILAR L2 sounds can be quite challenging for Japanese listeners with overseas experience (Study 1), that accurate categorization of SUBSET sounds can be fairly easy for naïve AusE listeners (Study 2), and that NEW category formation on already-categorized dimensions is achievable for Japanese learners without overseas experience (Study 3), though a non-previously categorized NEW scenario can be extremely difficult. Thus, the attested levels of difficulty are SUBSET < SIMILAR ≦ NEW.

This discrepancy is considered to result from L2LP's assumption that optimal perception entails the completion of both perceptual and representational tasks. As discussed earlier, L2LP strictly distinguishes between perceptual mappings and sound representations, which benefits the model in various ways. However, the model's optimal perception hypothesis assumes that both optimal perceptual mappings and optimal sound representations must be established for L2 optimal perception, thus conflating the two. These two components should instead be distinguished because sound representations can sometimes be accurately categorized without native-like perceptual mappings, as Study 2 has illustrated. Given that the primary goal of speech perception is to extract meaningful linguistic representations for communication, L2 learners are expected to prioritize the extraction of sound representations over native-like perceptual mappings. Furthermore, nonnative-like perceptual mappings could be sufficient as long as the learners find them useful (e.g., Japanese listeners who manage to categorize English /r/ and /l/ with non-essential cues such as the F2). Therefore, learners may cease to adjust their perceptual mappings once they find their sound categorization sufficiently accurate, which possibly leads to fossilization (Selinker, 1972). Fossilization is not predicted by L2LP (Escudero, 2005, p. 121) because learners are hypothesized to continue learning until both perceptual mappings and sound representations become optimal, which seems unrealistic.

The separation of optimal mappings and optimal representations in L2LP would help explain the observed difficulty of the three learning scenarios in Studies 1 – 3. Regarding Study 1, Japanese listeners' persistent use of duration in distinguishing AmE /iː/ and /ɪ/ can result from their already accurate categorization of the two vowel representations based on the duration-based mapping. Although vowel duration is not the most important cue for the AmE contrast, the vowels do exhibit systematic differences in duration as a side effect of realizing vowel tenseness. Japanese learners of English may thus be capable of accurately distinguishing the vowels based on the duration cue without learning to utilize the spectral cues. Consequently, for this particular SIMILAR scenario, optimal perceptual mappings are difficult to attain because the sound categorization is already quasi-optimal. Study 2 is a similar case, in which naïve AusE listeners are already fairly accurate at detecting the durational difference between Japanese /ii/ and /i/ representations, even though their perceptual mappings are L1-like. The only task for AusE learners of Japanese would be to adjust the durational mappings to increase discrimination accuracy. Adjustment of spectral mappings is not a must because it is irrelevant to the target L2 length contrast. Thus, this particular SUBSET scenario is not very difficult in terms of sound categorization, though the attainment of native-like perceptual mappings may be challenging, particularly in the spectral domain. Last, Study

3 found that Japanese learners of English can establish a new L2 category for AmE /æ/ with relative ease, while the other three vowels /ɛ, ʌ, ɑ/ were perceptually assimilated to L1 categories /e, a, o/. The establishment of a new sound representation is fairly achievable because learners can reuse and reorganize L1 features along already-categorized dimensions. However, native-like perceptual mappings may be difficult to acquire because learners could find their categorization of the four AmE vowels based on L1-like features (e.g., distinguishing *bed* from *bad* based on Japanese /mid/ and /low/ features) sufficiently useful for communicative purposes. Therefore, for this particular NEW scenario, the representational task of category creation is not difficult per se, whereas the perceptual task of boundary adjustment could be challenging.

In this way, the separation of perceptual mappings and sound representations in the L2LP's optimal perception hypothesis would allow the model to make coherent predictions regarding the difficulty of acquisition across scenarios. It would also redefine the definition of 'difficulty' in the model, as the perceptual and representational tasks should be associated with different types of difficulty. Given that learners can cease to adjust perceptual mappings once quasi-optimal sound categorization is achieved, the perceptual task is most likely more difficult than has previously been assumed, while the representational task may not be very difficult per se.

5.3.2 Necessity of other L2 models

SLM and PAM(-L2) help further explain the relative difficulty of the learning scenarios in the present thesis. For example, PAM(-L2) explains the SIMILAR scenario of Study 1 as a TC assimilation pattern, in which each L2 phonological category is perceived as equivalent to a different L1 phonological category. The model predicts that the learner would have little difficulty in discriminating minimally contrasting words for these distinctions, which consequently makes further perceptual learning unlikely or at least small in magnitude (unless the L2 sound is phonologically, but not phonetically, assimilated to the L1 sound). This prediction is compatible with the above discussion that optimal perceptual mappings become difficult to achieve when sound categorization is already accurate. On the other hand, SLM can explain the relative ease of the representational task and the relative difficulty of the perceptual task of the NEW scenario in Study 3. According to the model, new category formation is likely to occur because the L2 sound is phonetically dissimilar to the L1 sounds (H3). However, the phonetic category established for the L2 sound could differ from a monolingual's, as the learners' representation may be based on L1-like features or feature weights (H6). Therefore, SLM and PAM(-L2) can be used to complement L2LP's predictions, as they describe and explain the difficulty of various learning scenarios fairly well.

There are also a few phenomena in L2 phonology that SLM explains better than L2LP. One example is L1 attrition. L2LP predicts that "no attrition of L1 sound perception will be attested" (Escudero, 2005, p. 121) because L1 and L2 perception grammars are considered to exist independently of each other. However, it has been shown that the disuse of L1 can result in a complete loss of the perceptual ability to discriminate L1 sound contrasts (Ventureyra et al., 2004). L1 attrition is compatible with SLM, which considers that the phonetic systems remain flexible over the lifespan and that L1 categories can change under the influence of L2 acquisition. L1 attrition is closely related to L1 phonetic drift, namely the shift of L1 categories as influenced by the acquired L2 sound categories. L2LP would attribute L1 phonetic drift to language modes, in which L1 perception can be shifted towards the L2 side because the L2 grammar can be activated along with the L1 grammar. However, language modes do not explain why L1 categories can also drift away from L2 categories, i.e., dissimilation (Flege & Eefting, 1987). In contrast, SLM would explain that L1 and L2 categories are deflected away from each other to maintain sufficient phonetic contrast between categories in the common space. Thus, as far as the effect of L2 on the L1 is concerned, SLM's notion of common L1-L2 phonological space provides a more plausible account of L1-L2 category interactions than L2LP's separate grammars hypothesis.

PAM(-L2) provides a very detailed description of L1-L2 assimilation patterns, which also complements L2LP. For example, what would be treated as a NEW scenario in L2LP corresponds to either of CG, SC, and UC assimilation patterns in PAM(-L2), each of which is associated with a different level of discrimination difficulty and a different likelihood of category formation. Such detailed description helps determine which L2 sound is subject to category formation and which is subject to assimilation. Such information is useful for L2LP because the number of L1 and L2 sounds, which is claimed to determine the types of scenarios, is not very informative on its own, especially for SUBSET and NEW scenarios where extraneous and new categories need to be identified.

Finally, it must be noted that L2LP can explain certain phenomena better than the other two models, such as the distributional and meaning-driven nature of perception, language-specific perception modes, the SUBSET scenario and the MCA pattern, to name a few. The point here is this: since each model focuses on a different aspect of L2 speech perception, all models help to understand the phenomenon from a different perspective. By the nature of models being a simplified representation of a system of interest, no single model would suffice to explain every aspect of the whole system. Borrowing Cutler's words, "(e)very theory, after all, is ultimately wrong in some way" (2012, p. xv). Thus, it is advised that different types of models are consulted instead of a single one.

5.3.3 Remaining matters

There are a few remaining matters that L2LP wound need to address further. Discussed first is vowel normalization. In the present thesis and many other vowel perception studies based on L2LP, the input acoustic data for simulations were obtained from one gender, typically from male speakers. This is because the simulated grammar would perform poorly if the input comes from both genders since the spectral characteristics of male and female speech differ substantially due to their physiological differences in vocal tract length. Real listeners, on the other hand, are capable of handling such variability without trouble (McQueen & Cutler, 2010). The simulations in the present thesis dealt with this problem by using the average formant values by male speakers and by manipulating the standard deviations. However, this approach is not perfect because the results of the simulations are very sensitive to the exact values of the standard deviations (Escudero & Boersma, 2003). The model often fails to converge when the standard deviations are 'inappropriate,' but 'appropriate' values are difficult to identify because real values do not necessarily work. [16] Thus, it appears that some sort of normalization algorithm should be incorporated into the model.

[16] Reducing standard deviations increases the likelihood of convergence to some extent (Weiand, 2007). However, too small a standard deviation results in an 'untrained' region along an auditory dimension, where the model's response becomes random because no token has occurred in the region during training.

Escudero and Bion (2007) tackled this issue by modeling normalization and LP as sequential processes. In their study, the input acoustic data was manipulated by various formant normalization methods available in the literature (Adank et al., 2004), which were then fed to a Stochastic OT-based perception grammar. The result found that virtual listeners endowed with a normalization algorithm outperformed the control listener without it, suggesting that this sequential modeling is promising. However, the normalization methods used in the study are all vowel-extrinsic (i.e., they use information across multiple vowels), which is not realistic from a psycholinguistic point of view because real listeners can accurately identify vowels even when different speakers' voices are randomly mixed (Verbrugge et al., 1976). Thus, a more listener-oriented, psychoacoustic method of normalization needs to be identified. There are many proposals for vowel-intrinsic normalization methods, including those using F0 and higher formants, those using auditorily based scales such as mel, ERB and Bark, and those using formant ratios. While these methods aim to remove idiosyncratic information from the speech signal to obtain invariant abstract categories, an alternative proposal is that linguistic categories should be modeled as collections of experienced instances (Johnson, 2008). How listeners cope with the variable acoustic signal is an important question for vowel perception modeling under L2LP.

Another issue to be addressed is age. L2LP attributes the negative effect of age on L2 acquisition to a gradual loss of cognitive or neural plasticity. This is also reflected in the GLA’s plasticity scheme, which is initially set high and then decreases at a certain rate. L2LP also claims that the role of input overrules decreased plasticity, suggesting that the acquisition of L2 optimal perception is possible in adulthood. However, most studies under L2LP, including those in the present thesis, have investigated adult L2 learners with controlled AOL, and thus have not directly investigated the age factor. While there is ample evidence that children acquire L2 production faster than adults (Oh et al., 2011) and that late L2 learners can obtain native-like pronunciation (Birdsong, 2007), the effect of age on L2 perception still requires further investigation. A practical obstacle to studying it is the difficulty of testing infant and child L2 learners. This is particularly the case in Japan, where most L2 research takes place in universities and therefore the samples tend to be restricted to adult learners only. Nonetheless, Japan is expected to provide an interesting environment for testing the effect of age in the upcoming future because, as mentioned earlier, the age of onset of EFL instructions is being lowered and can possibly be lowered further. Thus, it will be possible to test L2 learners with roughly the same L1 backgrounds but different AOL. The result may then be compared with L2LP-based simulations to evaluate the validity of the plasticity scheme.

Research has also suggested that early and late bilinguals may acquire L2 phonology in fundamentally different ways, which questions whether the current plasticity scheme of L2LP adequately explains the effect of age. For example, Yang, Fox, and Jacewicz (2015) documented a bilingual child (AOL 3;7) who established his L2 vowel space by first reorganizing the existing L1 vowel space to make the two vowel spaces maximally distinct from each other, then adjusting it, suggesting that the starting point of L2 production is not necessarily L1 categories as is assumed in L2LP. Kartushina et al. (2016) also suggest that early but not simultaneous bilinguals (AOL < 8) are subject to reorganization of the L1 phonetic space, whereas later bilinguals are subject to L2-L1 assimilation. Therefore, it is an open question whether the current version of L2LP is adequate for explaining L2 perceptual acquisition by child learners.

Last, individual differences have not received adequate treatment in L2LP or any other model of L2 perception. Skehan (1991) classified individual differences into four main areas: aptitude, motivation, learning strategies, and learning styles. Of particular interest to L2 speech acquisition is aptitude, which involves the “phonemic coding ability” or the capacity to make sound discriminations and code nonnative sounds so that they can be recalled later (Carroll, 1965). Although little is known as to how such an ability should be modeled, one study has suggested that the ‘compactness’ of L1 categories predicts L2 production accuracy (Kartushina & Frauenfelder, 2014), which could be applicable to perception as well. In the study, adult Spanish learners of French were tested on their production of French /e/-/ɛ/, which are similar to Spanish /e/, and French /ø/-/œ/, which are distinct from any of the five Spanish vowels /i, e, a, o, u/. The participants' native productions were analyzed to assess the variability in the production of L1 vowels (i.e., the compactness of vowel categories in the F1-F2 space) as well as the position of the vowels in the acoustic space. The result found that speakers with more compact distributions for Spanish /e/ were better at producing the similar French /e/ and /ɛ/ vowels. Likewise, those with more compact overall distributions for the five Spanish vowels were better at producing the uncategorized French /ø/ and /œ/ vowels. Using SLM, the authors explained that speakers whose L1 productions are more compact would be more likely to discern the phonetic differences between L1 and L2 sounds as there is less spectral overlap, compared to those whose productions are more dispersed. This leads to a more likelihood of establishing new L2 categories and therefore more native-like productions. The explanation is in line with the notion of Carroll's phonemic coding ability. Provided that the same kinds of sound categories are used in both perception and production, language learning aptitude as represented by L1 category compactness should be able to be modeled in L2LP.

5.4 Future directions

5.4.1 Addressing the identified limitations

Despite the limitations, L2LP is one of the most comprehensive and compelling models of L2 speech perception available to date, which should be extended to overcome its current insufficiencies. Below I summarize suggestions for extending the model based on the above discussion:

(1) Adopting the featural approach. L2LP is amenable to featural accounts, which would enable modeling of various perceptual phenomena, including new feature acquisition (e.g., tenseness features in Study 1), transfer of features (e.g., length features in Study 2, cf. feature hypothesis), and feature reorganization (e.g., "*/low, front/" in Study 3). Featural modeling would also result in more phonetically faithful mappings and stricter separation of perceptual and representational tasks.

(2) Modification of the optimal perception hypothesis. Optimal perceptual mappings and optimal sound representations should be distinguished to explain the relative levels of difficulty of the scenarios better. The perceptual task is considered to be more difficult than the representational task, as the learner may cease to adjust mappings once they achieve a satisfactory level of accuracy in categorizing sound representations. The definition of 'difficulty' in the model may also need to be redefined.

(3) Incorporating vowel normalization. Escudero and Bion's (2007) sequential modeling of vowel normalization and LP, in which the acoustic input is normalized and then fed to the OT perception grammar, seems promising. However, since vowel-extrinsic normalization methods are psycholinguistically implausible, a more listener-oriented method of vowel normalization should be identified and incorporated into the model.

(4) Investigation of the plasticity scheme. Given that the effect of age on L2 speech acquisition has received less empirical testing in perception than in production, and given the possibility that early and later bilinguals acquire the L2 in fundamentally different manners, it should be investigated whether the current plasticity scheme in the GLA is adequate. Japan may provide a good environment for testing the age factor in the future because of the ongoing EFL curriculum reform to lower the AOL.

(5) Modeling individual differences. Kartushina et al.'s (2014) finding that individually variant compactness of L1 categories predict success in L2 production is expected to be applicable to L2 perception, assuming that these categories are shared between perception and production. Other types of individual differences such as motivation are also worth investigating.

5.4.2 Beyond auditory-to-surface mapping

Another important avenue for future research is to expand the scope of inquiry to higher-level representations. The present thesis has focused on [auditory] → /surface/ mappings only, following the original L2LP model. However, the three case studies revealed that this is insufficient for adequately modeling L2 perception and acquisition. Specifically, Study 1 suggested that an L2 category could be phonologically, but not phonetically, assimilated to an L1 category. Study 2 suggested that monophthongs and diphthongs may be treated in fundamentally different ways at the phonological level. Study 3 suggested that orthographic influences need to be considered to explain the assimilation patterns. In addition, there are many perceptual phenomena that would not be explained by simple [auditory] → /surface/ mappings. Examples of such phenomena are perceptual vowel epenthesis (Dupoux et al., 1999; Mazuka et al., 2011), in which acoustically non-present "illusory" vowels are perceived (e.g., Japanese listeners hear [ebzo] as /e.bu.zo/), and word recognition, by which pre-lexical perception may be overridden (e.g., [klɪn] is recognized as *clean* because *clin* is not a word). Below, two recent models that are capable of handling higher-level representations are presented: the revised L2LP model (van Leussen & Escudero, 2015) and the Bidirectional Phonology and Phonetics (BiPhon) model (Boersma, 2011).

5.4.2.1 Revised L2LP model

The revised L2LP model (van Leussen & Escudero, 2015) is a multi-layered, connectionist-inspired extension of L2LP, which consists of four levels of representations: [auditory], /surface/, |underlying|, and <lexical>. Units on adjacent levels are connected, and the process of perceiving and recognizing a word is represented as a four-step path through this network (i.e., [auditory] → /surface/ → |underlying| → <lexical>). Figure 5-1 shows how the revised L2LP represents L1 Dutch listeners' perception of L2 Spanish front vowels (SUBSET scenario). In this example, the [auditory] input [tʃVka, F1(V) = 4.0 Bark] corresponds to a realization of either *chica* 'girl' or *checa* 'Czech female' in Spanish, with an F1 value of 4 Bark for the front vowel (V). Under the assumption that the L2 grammar is initially a full copy of the L1 grammar, the [V] connects to one of three front vowels in Dutch on the /surface/ level, i.e., /tʃika/, /tʃɪka/, and /tʃɛka/. These in turn connect to three |underlying| representations, i.e., |tʃika|, |tʃɪka|, and |tʃɛka|, eventually leading to either of two <lexical> items, namely <girl> or <Czech.F>. This yields a total of 18 paths (3 × 3 × 2) for each acoustic input. The relative strengths of connections along these paths decide the optimal route and thus what is perceived and recognized. The strengths of the connections are altered over the course of learning. Knowledge of a language is thus stored in the connection strengths.

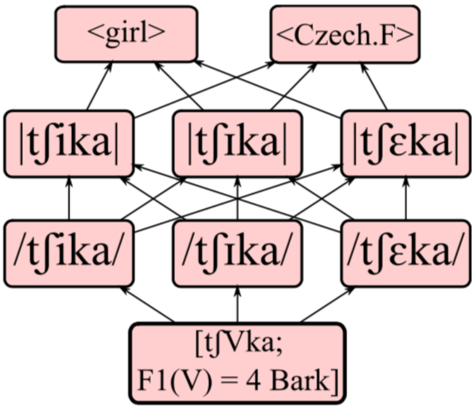


Figure 5-1. Structure of revised L2LP (van Leussen & Escudero, 2015).

Following Stochastic OT, the strength of each connection is represented by a *ranking value*, which is distorted slightly by *evaluation noise* to obtain a temporary *selection point* at each time of evaluation. The optimal path is not defined by the sum of the selection points, but rather is the one containing the least weak connections, in accordance with the central tenet of OT. Furthermore, learning is modeled as error-driven updating of the connections in the same way as the GLA. When there is a mismatch between the recognized <lexical> form and the intended form, the current grammar is updated by weakening all connections along the path that led to the incorrect form and strengthening all connections along the path to the intended form[17], by subtracting and adding the *plasticity* value that is set to decrease during learning. In essence, the revised L2LP is a functional extension of Stochastic OT and the GLA.

[17] The learner parses a single path to the correct form to decide which connections to strengthen. Evaluation noise is reapplied prior to parsing, which enhances the likelihood of convergence (Jarosz, 2013)

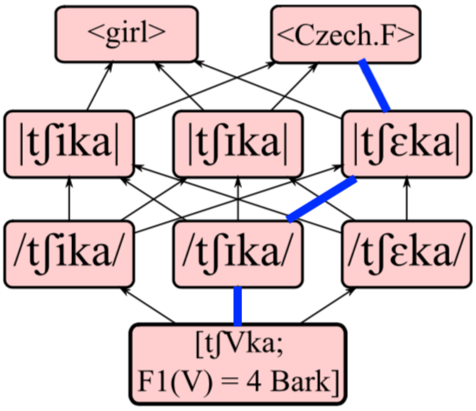


Figure 5-2. Unfaithful /surface/ → |underlying| mapping in revised L2LP.

The revised L2LP model is capable of modeling various phenomena that could not be modeled by [auditory] → /surface/ mappings alone. For example, the aforementioned example of the SUBSET scenario illustrates a case in which perception does not always equate with recognition. It can be seen in Figure 5-1 that the listener can perceive three surface forms (/tʃika/, /tʃɪka/, and /tʃɛka/) and recognize three underlying forms (|tʃika|, |tʃɪka|, and |tʃɛka|), even though Spanish has only two front vowels /i/ and /e/ and two corresponding lexical forms *chica* and *checa*. A possible result of this discrepancy is depicted in Figure 5-2, in which the learner perceives [tʃVka; F1 (V)= 4 Bark] as the surface form /tʃɪka/, which is then mapped unfaithfully to the underlying form |tʃɛka|, leading to the recognition of <Czech.F>. The unfaithful /surface/ → |underlying| mapping is an important conceptual shift from the original L2LP model, in which the surface forms are assumed to always faithfully map to underlying forms.

Another advantage of the revised L2LP is that it can model the two processes of perception and recognition as either sequential (i.e., bottom-up) or interactive (i.e., bottom-up and top-down). The original model held a sequential view, in which the listener always evaluates the [auditory] → /surface/ connections of perception prior to the /surface/ → |underlying| and |underlying| → <lexical> connections of recognition. However, it is still a matter of debate in psycholinguistics whether the outcome of pre-lexical perception forms the input to recognition or whether the two processes interact with each other. In revised L2LP, both sequential and interactive versions of modeling can be implemented by assigning the connections *stratum indices* besides the ranking values. At each time of evaluation, connections are ordered first by stratum, then by selection point (i.e., ranking value plus evaluation noise). For example, if the [auditory] → /surface/ connections of perception are assigned a higher stratum than the connections of recognition, perception precedes recognition and thus the two processes are modeled as sequential. On the other hand, by assigning all connections the same stratum, the connections of recognition may influence the outcome of perception and thus the two processes interact. Testing the two versions of L2LP is expected to contribute to the long-standing debate in psycholinguistics, which makes the revised L2LP an appealing successor to the original L2LP.

Given the advantages of the revised L2LP, the reader may wonder why I have not utilized it at all in the present thesis. There are three reasons for this. First, the structure of the revised model is peculiar and difficult to apply to other scenarios. While one might envision the revised L2LP as a kind of artificial neural network that can process any acoustic input to yield an optimal output, the reality is somewhat different. In the model, a single network as in Figure 5-1 can process only one representable acoustic value, e.g., [F1 = 4 Bark]. For example, van Leussen and Escudero (2015) divided the F1 dimension from 2 to 8 Bark into steps of 0.1 Bark, resulting in a total of 61 representable acoustic values. Each of the representable values was assigned to one of 61 independently prepared networks, rather than a single coherent network. This unusual and somewhat implausible implementation comes from the need to conform to the constraint-based principle of OT. For example, the constraint "[F1 = 4 Bark] is not /i/" would not assess any acoustic input but [F1 = 4 Bark], and so is "[F1 = 4.1 Bark] is not /i/". This makes it difficult to extend the current implementation to more than one auditory dimension, as another auditory dimension would require another set of independent networks. Such multi-dimensional modeling based on revised L2LP would be too complex and beyond the ability and inclination of most L2 researchers. Thus, while revised L2LP is faithful to the original L2LP, it is theoretically less plausible and practically less accessible.

Second, the ‘phonological’ representations in revised L2LP do not in fact suffice to represent phonological representations. Perceptual vowel epenthesis illustrates this point. Dupoux et al. (1999) found that Japanese listeners perceive ‘illusory’ vowels inside consonant clusters in VCCV stimuli, using a continuum of nonsense stimuli ranging from no vowel (e.g., [ebzo]) to full vowel between the consonants (e.g., [ebuzo]). The Japanese participants reported the presence of a vowel [u] between consonants, even in stimuli with no vowel. They also had difficulty in discriminating between VCCV and VCuCV stimuli. This perceptual ‘illusion’ is considered to result from the Japanese syllable structure prohibiting consonant clusters and syllable coda. While the revised L2LP may be able to express this phenomenon as [eb(V)zo] → /ebzo/, /ebuzo/ → |ebzo|, |ebuzo| (there is no <lexical> form because *ebuzo* is a nonword), it fails to account for why epenthesis would occur in the first place and why the vowel /u/ should be perceived out of the five Japanese vowels. This is because the connections do not really represent phonological constraints that prohibit consonant clusters and that prefer certain vowels under certain phonological conditions. The mere addition of /surface/ and |underlying| levels in revised L2LP is thus insufficient for modeling phonological processes in perception. In order to explain perceptual epenthesis, one would need a different kind of grammar such as regular OT or more traditional rule-based phonology.

Third, since the revised L2LP is faithful to the principles of the original, the two models yield very similar or even identical predictions. This makes the choice of which model to use a methodological rather than theoretical preference. For example, van Leussen and Escudero (2015) used the SUBSET scenario of Boersma and Escudero (2008), i.e., L1 Dutch listeners' perception of L2 Spanish front vowels, to illustrate how the computational implementation of the original L2LP could be extended. Aside from methodological differences, the original and revised models predict identical acquisition paths for this scenario: the learner starts with a copy of the L1 Dutch grammar with three vowels /i, ɪ, ɛ/ (*Full Copying* hypothesis), which they adjust through the GLA (*Full Access* hypothesis) by shifting categorical boundaries (perceptual task) and getting rid of the extraneous /ɪ/ category (representational task) to optimally perceive L2 Spanish /i, e/ (optimal perception hypothesis). The only theoretical difference would be the absence or presence of the |underlying| level[18], although the |underlying| level in revised L2LP does not strictly represent phonological representations as discussed above. Thus, there is nothing against adhering to the original model. For these reasons, I prefer Escudero's (2005) original implementation as a more straightforward and accessible version of L2LP compared to the revised implementation.

[18] Note that the <lexical> level is implicitly assumed in the original L2LP because it is indispensable to meaning-driven learning by the GLA.

5.4.2.2 Bidirectional Phonology and Phonetics (BiPhon) model

A promising alternative to the revised L2LP is the BiPhon model, which is a predecessor of L2LP as stated in Chapter 3. BiPhon is an OT-based grammar model that aims to handle 'all' of phonetics and phonology, including both the listening process (comprehension) and the speaking process (production). Figure 5-3 shows the structure of the model, which is intended to be capable of *whole-language simulations* of a language. The task of the listener is to travel up the figure from the [[auditory]] form eventually to the "context" or semantic representations. In contrast, the task of the speaker is to travel down the figure from the intended "context" to an [articulatory] form. When traveling up or down the figure, the speaker or listener visits a number of intermediate representations (/surface/ form, |underlying| form, and <morphemes>). These processes are modeled by Optimality-Theoretic constraints that evaluate either a single level of representation (structural and articulatory constraints) or a relation between two levels of representation (semantic, lexical, faithfulness, cue, and sensorimotor constraints). The constraints are used bidirectionally in the sense that language users use the same constraints when they speak as when they listen, with the same rankings (Smolensky, 1996). As in L2LP, the constraint rankings are expressed as ranking values in Stochastic OT and are learned through the error-driven algorithm of the GLA.

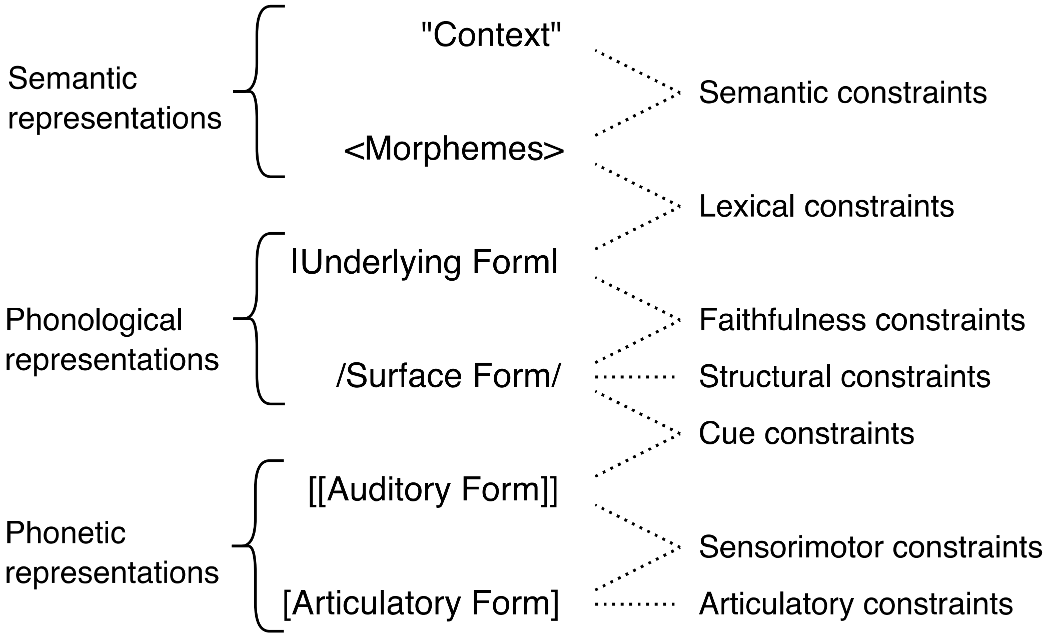


Figure 5-3. Representations and constraints in BiPhon (Boersma, 2011).

Below I illustrate how the BiPhon model is capable of modeling various perceptual phenomena, including perceptual vowel epenthesis and word recognition. To begin with, pre-lexical perception, which was the primary focus of the present thesis, is modeled as a [[auditory]] → /surface/ mapping through cue constraints as in the original L2LP model (e.g., "[[F1 = 300 Hz]] is not /i/"). An example is shown in Tableau 5-1, in which the auditory form [[F1 = 300 Hz]] is perceived as the surface form /i/ by a Japanese listener. Note that the finger points backward ("☜") to mark that the candidate wins in the comprehension direction; candidates that win in the production direction are marked by the regular OT finger ("☞").

Tableau 5-1. Vowel perception in Japanese.

[[F1=300 Hz]]	*/e/ [[300 Hz]]	*/i/ [[300 Hz]]
☜ /i/ [[300 Hz]]		*
/e/ [[300 Hz]]	*!	

Unlike L2LP, BiPhon considers another kind of constraint that poses restrictions on the outcome of prelexical perception. These are called structural constraints, which are synonymous with markedness constraints in traditional OT.[19] Structural constraints are required for modeling perceptual vowel epenthesis. A notable strength of BiPhon is that any OT-based phonological analysis can be incorporated into the model. For example, Otaki (2012) analyzed vowel epenthesis in Japanese loanword adaptation using OT. According to him, /u/ is epenthesized within consonant clusters and after syllable coda as a result of markedness constraints on syllable structure, namely *COMPLEX and NOCODA. The vowel /u/ is chosen as the winner due to markedness constraints on epenthetic vowel structure (EPENTHVOWEL): */-high/, */-back/, and */+low/. The ranking is based on the principle that epenthetic vowels should be perceptually least salient; high vowels are less sonorant than non-high vowels, and high back vowels are shorter than high front vowels. The above analysis can be reinterpreted as a perceptual process in BiPhon, where [[abzo]] is perceived as /ebuzo/ (Tableau 5-2).

[19] Feature co-occurrence constraints can also be seen as a structural constraint.

Tableau 5-2. Perceptual vowel epenthesis in Japanese.

[[ebzo]]	*/COMPLEX/	/EPENTHVOWEL/		
		*/-high/	*/-back/	*/+low/
/ebzo/ [[ebzo]]	*!			
/ebizo/ [[ebzo]]			*!	
/ebezo/ [[ebzo]]		*!	*	
/ebazo/ [[ebzo]]		*!	*	*
/ebozo/ [[ebzo]]		*!		
☜ /ebuzo/ [[ebzo]]				

After prelexical perception comes phonemic and word recognition, i.e., the /surface/ → |underlying| mapping and the |underlying| → <lexical> mapping. Phonemic recognition is mediated by faithful constraints, which require identity between /surface/ and |underlying| forms. Word recognition is mediated by lexical constraints, which require the |underlying| form to correspond to an existent <morpheme>. Presented below is an example of word recognition, adapted from Boermsa (2011). The target [[auditory]] form is a vowel in /klVn/ in SE. In Tableau 5-3, it can be seen that the auditory form [[F1 = 450 Hz]] is mapped to the lax vowel /ɪ/ through cue constraints, leading to the perception of the surface form /klɪn/.

Tableau 5-3. Vowel perception in SE.

[[F1=450 Hz]]	*/i/ [[450 Hz]]	*/ɪ/ [[450 Hz]]
/klin/ [[450 Hz]]	*!	
☜ /klɪn/ [[450 Hz]]		*

While the perceived from /klɪn/ does not violate any structural constraint (i.e., it is phonologically well-formed), there is a problem that no |underlying| form exists for /klɪn/ because *clin* is a nonword. The closest substitute would be |klin|, i.e., <clean> or possibly |klæn|, i.e., <clan>. Therefore, native SE listeners are expected to unfaithfully recognize the perceived surface form /klɪn/ as *clean* or *clan*. This can be modeled with the help of faithfulness constraints such as "*|æ|/ɪ/" ("an underlying form |æ| does not connect to a surface form /ɪ/") and the lexical constraint "*<>|*x*|" ("the underlying form |*x*| must correspond to any morpheme"), as shown in Tableau 5-4. In the tableau, the lexical constraint "*<>|*x*|" prohibits the recognition of |klin| because the listener would prefer to perceive any meaningful word rather than a nonword. The question, then, is whether |klin| <clean> or |klæn| <clean> should be perceived, which is evaluated by the faithfulness constraints. In this example, "*|æ|/ɪ/" outranks "*|i|/ɪ/," possibly because /i/ is phonetically and phonologically closer to /ɪ/ than to /æ/. Consequently, |klin| <clean> is recognized.

Tableau 5-4. Word recognition in SE.

/klɪn/	*<> |*x*|	*|æ| /ɪ/	*|i| /ɪ/
☜ <clean> |klin| /klɪn/			*
< > |klɪn| /klɪn/	*!		
<clan> |klæn| /klɪn/		*!	

Nonword recognition can also be modeled by simply lowering the ranking of the lexical constraint. In Tableau 5-6, /klɪn/ is faithfully perceived as |klɪn| (*clin*).

Tableau 5-5. Nonword recognition in SE.

/klɪn/	*|æ| /ɪ/	*|i| /ɪ/	*<> |*x*|
<clean> |klin| /klɪn/		*!	
☜ <> |klɪn| /klɪn/			*
<clan> |klæn| /klɪn/	*!		

The above examples modeled perception and recognition as sequential processes (left side of Figure 5-4). However, BiPhon can implement a parallel evaluation of the two processes (right side of Figure 5-4) as well, by including all the constraints for perception and recognition in the same tableau. Parallel modeling of perception and recognition in SE is shown in Tableau 5-6. An interesting outcome is that the presence of a lexical item can affect the auditory boundary, which is known as the Ganong effect (Ganong, 1980). It can be seen in the tableau that the cue constraints prefer /ɪ/, while the faithfulness constraint prefers /i/. Consequently, the listener perceives [[F1 = 450 Hz]] as /klin/, while the same auditory input would have been perceived as /klɪn/ if it were not for lexical information as in Tableau 5-3. That is, the auditory boundary between /i/ and /ɪ/ was shifted towards the /ɪ/ side because of the lexical representation <clean> |klin|.

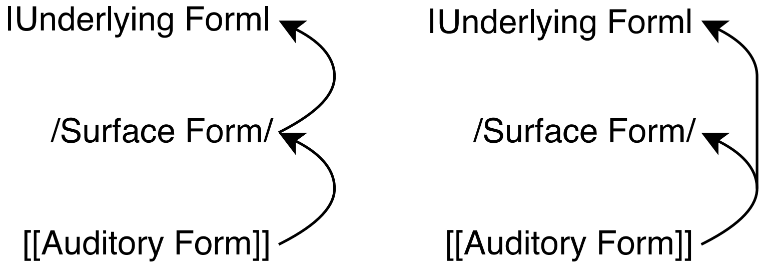


Figure 5-4. Serial (left) and parallel (right) models of speech comprehension.

Tableau 5-6. Lexical effect on auditory boundary in SE.

[[F1=450 Hz]]	*<> /x/	*|i| /ɪ/	*/i/ [[450 Hz]	*/ɪ/ [[450 Hz]]
< > |klɪn| /klɪn/ [[450 Hz]]	*!			*
< > |klɪn| /klin/ [[450 Hz]]	*!		*	
<clean> |klin| /klɪn/ [[450 Hz]]		*!		*
☜ <clean> |klin| /klin/ [[450 Hz]]			*	

To summarize, BiPhon is capable of modeling various perceptual phenomena involving higher-level representations, which the current L2LP (original or revised) may not be capable of. As L2LP grew out of BiPhon, the two models share fundamental similarities, indicating that applying the BiPhon approach to L2LP would not be so challenging. The most significant advantage of incorporating BiPhon into L2LP would be that L2 speech production can be modeled, which is illustrated in the following section.

5.4.3 Toward speech production

As mentioned earlier in the thesis, the cause of L2 production difficulties has been ascribed to “perceived foreign accentedness” (Strange, 1995), which is why most current theories and models of L2 speech acquisition focus on perception. SLM explicitly states that the cause of foreign accents resides in learners’ inaccurate perception of L2 sounds. L2LP agrees with SLM; according to Escudero (2005, p. 3), “it can be concluded that perception develops first and needs to be in place before production development can occur, and also that the difficulties with L2 sounds have a perceptual basis such that incorrect perception leads to incorrect production.” While the relationship between perception and production is complex and controversial (Hattori, 2009; Sheldon & Strange, 1982), the consensus is that there are at least partial associations between the two. If that is the case, then L2 perception models such as L2LP should be extendable to L2 production, which opens a great avenue for future research.

This final section discusses the possibility of extending L2LP to speech production, with the help of the BiPhon framework that models both the listening (comprehension) and speaking (production) processes using the same Optimality-Theoretic constraints. Since BiPhon and L2LP can be seen as variants of the same model, the bidirectional approach in BiPhon should be readily applicable to L2LP.

BiPhon models both comprehension and production of speech by using the same sets of constraints, with the same rankings. This bidirectionality is a great strength of the model, which enables whole-language simulations according to Boersma (2011). For example, the process of prelexical perception ([[auditory]] → /surface/) can be reversed, i.e., /surface/ → [[auditory]]. This can be termed as prototype selection, in which speakers judge what would be the best [[auditory]] form that is associated with a given phonological /surface/ form, via cue constraints acquired through perceptual learning. Tableau 5-7 shows an example of prototype selection, in which the auditory form [[F1 = 290 Hz, duration = 200 ms]] is chosen as the most prototypical exemplar of /i/ in SE. Note that the regular OT hand ("☞") is used here because this is part of the production process. However, this process cannot really be called 'production' yet because articulatory considerations are not involved. Not restricted by articulatory effort, the winning [[auditory]] form may be much more peripheral or extreme (lower F1 and higher duration) than the average auditory realization, e.g., [[330 Hz, 100 ms]]. The winner is a better token (i.e., a prototype) than the average token because it has less chance of being perceived as anything but /i/. For example, given that a typical /ɪ/ has an auditory form [[500 Hz, 80 ms]], the best exemplar of /i/ would have an auditory value that is farthest away from it.

Tableau 5-7. Prototype selection in SE.

/i/	*/i/ [[74 ms]]	*/i/ [[349 Hz]]	*/i/ [[200 ms]]	*/i/ [[290 Hz]]
/i/ [[349 Hz, 74 ms]]	*!	*		
/i/ [[290 Hz, 74 ms]]	*!			*
/i/ [[349 Hz, 200 ms]]		*!	*	
☞ /i/ [[290 Hz, 200 ms]]			*	*

Modeling the process of speech production requires an additional [articulatory] level of representation and two constraints associated with it, namely sensorimotor and articulatory constraints. Suppose that a SE speaker attempts to produce the vowel /i/. The speaker first computes an auditory prototype of the surface form /i/, e.g., [[F1 = 290 Hz]], which is then put into an articulation. Whereas the sensorimotor constraints (e.g., "*[[290 Hz]][330 Hz]," i.e., "an auditory form [[290 Hz]] does not correspond to an articulatory form [330 Hz]") would prefer an articulation that produces [290 Hz], the articulatory constraints (e.g., "*Energy[290 Hz]," i.e., "the articulators do not spend articulatory energy required for achieving an articulatory form [290 Hz]") would prevent this. Tableau 5-8 shows how vowel production can be modeled in BiPhon. Here, [[290 Hz]] has already been chosen as the most prototypical auditory form of /i/, and thus the cue constraints are omitted from the tableau. The sensorimotor and articulatory constraints compete with each other to decide which [articulatory] form should be produced. The sensorimotor constraints prefer an articulatory form that is closer to the prototype, i.e., [330 Hz] > [310

Hz] > [290 Hz] in the order of constraint strictness. However, the articulatory constraints prefer to spend as little articulatory energy as possible. Given that an articulation of more peripheral vowel requires more articulatory energy, the ranking would be [290 Hz] > [310 Hz] > [330 Hz] in the order of strictness. In this example, an articulatory form of [330 Hz] is ultimately produced because the articulatory constraints are ranked higher than the sensorimotor constraints. In other words, the speaker chose to spend less articulatory energy at the expense of a more perfect correspondence between the [[auditory]] and [articulatory] forms. In some occasions such as careful speech, sensorimotor constraints are considered to outrank articulatory constraints to produce more prototypical forms.

Tableau 5-8. Vowel production in SE.

/i/	*ENGY [290]	*ENGY [310]	[[290]] [330]	[[290]] [310]	*ENGY [330]	[[290]] [290]
/i/ [[290]] [290]	*!					
/i/ [[290]] [310]		*!		*		
☞ /i/ [[290]] [330]			*		*	*

The above example modeled production in a serial manner, in which prototype selection proceeds phonetic articulation (left side of Figure 5-5). However, these processes can also be evaluated in parallel (right side of Figure 5-5) by evaluating the cue, sensorimotor, and articulatory constraints all at the same time within a single tableau.

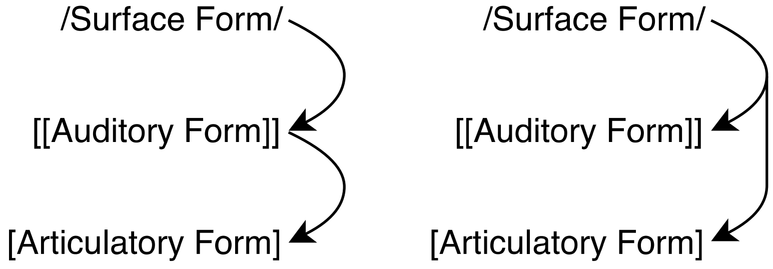


Figure 5-5. Serial (left) and parallel (right) models of speech production.

Boersma (2011) explains that the parallel model of speech production may be more preferable because the GLA cannot acquire the appropriate rankings of articulatory constraints in the serial model. When [[auditory]] → [articulatory] mappings are learned alone, the sensorimotor constraints always outrank the articulatory constraints, which become ranked so low that they stop determining articulatory output. Since learning occurs only when there is a mismatch between [[auditory]] and [articulatory] forms, the model learns to always achieve a perfect correspondence between these forms, regardless of articulatory effort. This problem can be solved in the parallel model, in which /surface/ → [articulatory] mappings are also considered. For example, if the correct articulatory form of /an+pa/ is [ampa] due to place assimilation, the model learns to produce [ampa] by ranking articulatory constraints for place assimilation high, overriding cue and sensorimotor constraints that prefer /an+pa/ → [[anpa]] and [[anpa]] → [anpa] mappings.

However, Boersma (2011) also notes that the serial version may suffice to model speech production because articulatory constraints might be unlearnable unlike other types of constraints. That is, given that articulatory constraints are connected to the articulatory periphery, their ranking is perhaps directly determined by articulatory effort and therefore fixed. If constraint reranking is a process that occurs in the brain, then it is questionable whether articulatory constraints should be treated in the same way as other constraints. If this is the case, then the problem of articulatory constraints being ranked too low in the serial model is no longer problematic, making the serial model eligible.

Whether serial or parallel, BiPhon proposes that the knowledge between /surface/ and [[auditory]] forms, which is learned through perception, is directly used in the process of speech production. It follows that the perceptual acquisition of [[auditory]] → /surface/ mappings in L2LP, which was the primary focus of the present thesis, is applicable to the modeling of L2 speech production. Also, the articulatory and sensorimotor constraints in BiPhon may help reveal partial dissociations between perception and production. The extension of L2LP to production with the help of BiPhon seems a promising avenue for future research in SLA, as it should deepen our understanding of the nature of L2 phonological acquisition.

5.5 Chapter summary

L2LP was found to be a comprehensive and useful model of L2 perception with high explanatory and predictive power. The model's precise and testable predictions benefit from the separation of perceptual mappings and sound representations as well as the incorporation of computational simulations, both of which currently lack in other models such as SLM and PAM(-L2). L2LP also has the potential for bridging the gap between L2 perception research and L2 education due to its focus on the role of input and distributional and meaning-driven learning. Also, the three case studies have demonstrated that research under L2LP sheds new light on previously under-researched areas such as language-specific perception modes and feature-based perception.

However, it was found that L2LP's predictions on the relative levels of difficulty of the learning scenarios were inadequate. The model also had difficulty in explaining certain aspects of L2 phonological acquisition such as L1 phonetic drift and different types of L1-L2 assimilation, indicating a necessity for other models to complement the model. While L2LP can be extended to many different directions to overcome its current insufficiencies, a promising avenue for future research seems a collaboration with the BiPhon model, which would enable comprehensive modeling and simulation of the whole process of L2 phonological acquisition in both perception and production.

Chapter 6: Conclusions

This thesis has provided a thorough empirical test of a recent model of L2 speech perception, the L2LP model, by conducting three case studies that correspond to each of the proposed three types of learning scenarios: SIMILAR (Study 1), SUBSET (Study 2), and NEW (Study 3). The specific learning scenarios of interest were Japanese listeners' perception of L2 AmE /iː/-/ɪ/ vs. L1 Japanese /ii/-/i/ (Study 1), naïve AusE listeners' perception of nonnative Japanese /ii, i/ vs. native AusE /iː, ɪ, ɪə/ (Study 2), and Japanese listeners' perception of L2 AmE /ɛ, æ, ʌ, ɑ/ vs. L1 Japanese /e, a/ (Study 3). These studies were presented in the order of the predicted levels of difficulty according to the model: SIMILAR < SUBSET < NEW.

A great advantage of L2LP over previous models of L2 perception, such as SLM and PAM(-L2), is its incorporation of computational simulations. Simulations are useful not only for making predictions but also for evaluating the model on which the simulations are based. In each of the three case studies, computational implementations of L2LP were first presented to provide specific and testable predictions of the model regarding the particular learning scenario. The predictions were then compared with the result of a perception experiment on real listeners, which served as an empirical self-test of the model. The overall findings of each scenario are summarised below.

Study 1 found that Japanese learners of English exhibit language-dependent cue weighting in perceiving SIMILAR L1 and L2 sounds depending on which language they think they hear. The simulations predicted that learners would show primarily duration-based perception for Japanese /ii/-/i/, while their perception for AmE /iː/-/ɪ/ would be more dependent on spectra and less on duration. The experiment, which manipulated the language context in two sessions, supported the predictions. Listeners showed the predicted shift in cue weighting between sessions, despite the stimuli being identical. The results were thus compatible with L2LP's separate grammars hypothesis and the language mode hypothesis. However, duration remained as the dominant cue for most learners, which is at odds with L2LP's prediction that the SIMILAR scenario is of least difficulty.

Study 2 found that naïve AusE listeners adopt L1-like cue usage in perceiving nonnative vowels that constitute a SUBSET of their native vowels. The simulations predicted that listeners would fully transfer their use of duration and VISC cues, which are necessary for perceiving AusE /iː, ɪ, ɪə/, to the perception of Japanese /ii, i/. The prediction was borne out in the experiment, which found that listeners relied on duration and onset and offset formants in categorizing nonnative Japanese vowels. The result thus supports L2LP's *Full Copying* hypothesis. However, the SUBSET scenario did not seem as difficult as L2LP would predict because no representational task of category reduction was attested, as 'extraneous' AusE /ɪə/ was not perceived at all. Also, the simulations predicted that AusE listeners would be able to learn to distinguish Japanese /ii/ and /i/ fairly quickly without changing their nonnative-like spectral perception, suggesting that optimal perceptual mappings are not a prerequisite for accurate sound categorization.

Study 3 found that feature-based account is useful in modeling the acquisition of NEW AmE sounds by Japanese listeners. The case follows a sub-scenario of the NEW scenario involving already-categorized auditory dimensions. While the segmental modeling had difficulty in explaining why AmE /æ/ is perceived as a deviant exemplar of L1 segments and how this relates to new L2 category formation, the featural modeling provided a coherent explanation that AmE /æ/ sounds deviant because it consists of /low/ and /front/ features that do not co-occur in Japanese and that these features can be rearranged to form a new, phonologically well-formed category in L2 AmE. The experiment found that AmE /æ/ was represented as a 'fronted version of /a/' in Japanese listeners' perceptual space, which is compatible with the featural account. The overall result indicates that the process of new category formation is not necessarily difficult per se, as opposed to L2LP's claim. Instead, whether and how the features constituting the new L2 sound are utilized in the L1 (already-categorized vs. non-previously-categorized) seem to predict the relative difficulty of acquisition better.

In general, L2LP-based simulations for each particular learning scenario were found to be accurate, showing a close resemblance to real listeners' perception. Therefore, as an overall evaluation, the present thesis has verified the validity and usefulness of the model. However, the model's prediction concerning the difficulty of acquisition across scenarios was found to be inaccurate. As opposed to the prediction (SIMILAR < SUBSET < NEW), the attested levels of difficulty in the present thesis were SUBSET < SIMILAR ≦ NEW. This discrepancy, at least in part, comes from the model's assumption that L2 learners would continue to refine their perception until both optimal mapping and optimal representations are established. In reality, learners may cease to adjust their perceptual mappings once they achieve a satisfactory level of accuracy in sound categorization (i.e., fossilization), a possibility that L2LP has not considered.

While L2LP may outperform SLM and PAM(-L2) in concreteness, the two models are necessary to account for certain phenomena in L2 phonology that L2LP has difficulty explaining. For example, L1 phonetic drift and attrition are more straightforwardly explained by the notion of common phonological space in SLM than by L2LP's separate perception and language mode hypotheses. PAM(-L2)'s detailed description of perceptual assimilation patterns complements L2LP that focuses more on the number of sound categories in the L1 and the L2, which itself is not sufficiently informative. A model, by nature, simplifies the system of interest for understanding, and thus no single model would suffice to explain everything. Therefore, L2 researchers are advised to utilize or at least pay attention to different types of available models in the field.

There are many directions in which L2LP can be extended. Based on the present study's findings, the following suggestions are proposed to improve the current model: (1) adopting the featural approach, (2) separating perceptual mappings and sound representations in the optimal perception hypothesis, (3) incorporating a listener-oriented method of speaker normalization, (4) testing of the effects of age, and (5) modeling individual differences. While a recent revision to the model by van Leussen and Escudero (2015) addresses some of the limitations of the original model, the practical inaccessibility and theoretical implausibility of the revised version are seen as problematic. Instead, I propose that a collaboration between L2LP and BiPhon is a more promising avenue for future research. The bidirectionality of BiPhon, which aims to model both the listening process and the speaking process under the single framework of OT, would endow L2LP with the ability to comprehensively model the acquisition of L2 speech in both perception and production. Such modeling and empirical testing thereof will contribute to our better understanding of the nature of L2 speech acquisition.

References

Adank, P., Smits, R., & van Hout, R. (2004). A comparison of vowel normalization procedures for language variation research. *The Journal of the Acoustical Society of America*, *116*(5), 3099–3107.

Antoniou, M. (2010). *One head, two languages: Speech production and perception in Greek-English bilinguals*. Western Sydney University.

Antoniou, M., Best, C. T., Tyler, M., & Kroos, C. (2010). Language context elicits native-like stop voicing in early bilinguals' productions in both L1 and L2. *Journal of Phonetics*, *38*(4), 640–653.

Antoniou, M., Best, C. T., Tyler, M., & Kroos, C. (2011). Inter-language interference in VOT production by L2-dominant bilinguals: Asymmetries in phonetic code-switching. *Journal of Phonetics*, *39*(4), 558–570.

Arai, T., Behne, D. M., Czigler, P., & Sullivan, K. P. H. (1999). Perceptual cues to vowel quantity: Evidence from Swedish and Japanese. *Proceedings of the Swedish Phonetics Conference (Fonetik)*, *81*, 8–11.

Bak, T. H., Nissan, J. J., Allerhand, M. M., & Deary, I. J. (2014). Does bilingualism influence cognitive aging? *Annals of Neurology*, *75*(6), 959–963.

Bardovi-Harlig, K., & Sprouse, R. A. (2017). Negative versus positive transfer. In J. I. Liontas (Ed.), *The TESOL Encyclopedia of English Language Teaching* (pp. 1–6). John Wiley & Sons.

Bates, D., Mächler, M., Bolker, B. M., & Walker, S. C. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48.

Behne, D. M., Czigler, P. E., & Sullivan, K. P. H. (1997). Swedish quantity and quality: A traditional issue revisited. *Reports from the Department of Phonetics*, *4*, 81–83.

Benkí, J. R. (2001). Place of articulation and first formant transition pattern both affect perception of voicing in English. *Journal of Phonetics*, *29*(1), 1–22.

Best, C. T. (1994). The emergence of native-language phonological influences in infants: A perceptual assimilation model. In *The development of speech perception: The* *transition from speech sounds to spoken words* (pp. 167–224). MIT Press.

Best, C. T. (1995). A direct realist view of cross-language speech perception. In W. Strange (Ed.), *Speech perception and linguistic experience: Issues in cross-language research* (pp. 171–204). York Press.

Best, C. T., McRoberts, G. W., & Sithole, N. M. (1988). Examination of perceptual reorganization for nonnative speech contrasts: Zulu click discrimination by English-speaking adults and infants. *Journal of Experimental Psychology: Human Perception and Performance*, *14*(3), 345–360.

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

Bialystok, E. (2001). *Bilingualism in development: Language, literacy, and cognition*. Cambridge University Press.

Bion, R., Miyazawa, K., Kikuchi, H., & Mazuka, R. (2013). Learning phonemic vowel length from naturalistic recordings of Japanese infant-directed speech. *PLOS ONE*, *8*(2), e51594.

Birdsong, D. (2007). Nativelike pronunciation among late learners of French as a second language. In O.-S. Bohn & M. J. Munro (Eds.), *Language experience in second language speech learning: In honor of James Emil Flege* (pp. 99–116). John Benjamins.

Boersma, P. (1998). *Functional Phonology: Formalizing the interactions between articulatory and perceptual drives*. University of Amsterdam.

Boersma, P. (2011). A programme for bidirectional phonology and phonetics and their acquisition and evolution. In A. Benz & J. Mattausch (Eds.), *Bidirectional Optimality Theory* (pp. 33–72). John Benjamins.

Boersma, P. (1997). How we learn variation, optionality, and probability. *Proceedings of the Institute of Phonetic Sciences of the University of Amsterdam*, *21*, 43–58.

Boersma, P., & Chládková, K. (2011). Asymmetries between speech perception and production reveal phonological structure. In W.-S. Lee & E. Zee (Eds.), *Proceedings of the 17th International Congress of Phonetic Sciences* (pp. 328–331). The University of Hong Kong.

Boersma, P., & Escudero, P. (2008). Learning to perceive a smaller L2 vowel inventory: An Optimality Theory account. In P. Avery, E. Dresher, & K. Rice (Eds.), *Contrast in phonology: Theory, perception, acquisition* (pp. 271–302). Mouton de Gruyter.

Boersma, P., & Hayes, B. (2001). Empirical tests of the Gradual Learning Algorithm. *Linguistic Inquiry*, *32*(1), 45–86.

Boersma, P., & Weenink, D. (2019). *Praat: doing phonetics by computer (Version 6.1.08) [Computer software]*. http://www.praat.org/

Bohn, O.-S. (1995). Cross-language speech perception in adults: First language transfer doesn't tell it all. In W. Strange (Ed.), *Speech perception and linguistic experience: Issues in cross-language speech research* (pp. 275–300). York Press.

Bongaerts, T. (1999). Ultimate attainment in L2 pronunciation: The case of very advanced late L2 learners. In D. Birdsong (Ed.), *Second language acquisition and the critical period hypothesis* (pp. 133–159). Lawrence Erlbaum.

Browman, C. P., & Goldstein, L. (1992). Articulatory phonology: An overview. *Phonetica*, *49*(3–4), 155–180.

Caramazza, A., Yeni-Komshian, G., & Zurif, E. B. (1974). Bilingual switching: The phonological level. *Canadian Journal of Psychology/Revue Canadienne de Psychologie*, *28*(3), 310–318.

Carroll, J. B. (1965). The prediction of success in foreign language training. In R. Glaser (Ed.), *Training, research, and education* (pp. 87–136). Wiley.

Casillas, J. V. (2015). Production and perception of the /i/-/ɪ/ vowel contrast: The case of L2-dominant early learners of English. *Phonetica*, *72*(2–3), 182–205.

Casillas, J. V., & 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, C. B. (2012). Rapid and multifaceted effects of second-language learning on first-language speech production. *Journal of Phonetics*, *40*(2), 249–268.

Chang, C. B. (2013). A novelty effect in phonetic drift of the native language. *Journal of Phonetics*, *41*(6), 520–533.

Chládková, K., Escudero, P., & Lipski, S. C. (2015). When "AA" is long but "A" is not short: Speakers who distinguish short and long vowels in production do not necessarily encode a short long contrast in their phonological lexicon. *Frontiers in Psychology*, *6*, 438.

Chomsky, N. (1965). *Aspects of the theory of syntax*. MIT Press.

Clopper, C. G., Pisoni, D. B., & de Jong, K. (2005). Acoustic characteristics of the vowel systems of six regional varieies of American English. *The Journal of the Acoustical Society of America*, *118*(3), 1661–1676.

Coder, S. P. (1967). The significance of learners' errors. *International Review of Applied Linguistics in Language Teaching*, *5*(4), 161–170.

Cook, V. (2003). Introduction: The changing L1 in the L2 user's mind. In V. Cook (Ed.), *Effects of the Second Language on the First* (pp. 1–18). Channel View Publications.

Cox, F. (2006). The Acoustic characteristics of /hVd/ vowels in the speech of some Australian teenagers. *Australian Journal of Linguistics*, *26*(2), 147–179.

Cox, F., & Palethorpe, S. (2007). Australian English. *Journal of the International Phonetic Association*, *37*(03), 341–350.

Curtin, S., Fennell, C., & Escudero, P. (2009). Weighting of vowel cues explains patterns of word-object associative learning. *Developmental Science*, *12*(5), 725–731.

Cutler, A. (2012). *Native listening: Language experience and the recognition of spoken words*. MIT Press.

Derwing, T. M., & Munro, M. J. (2005). Second language accent and pronunciation teaching: A research-based approach. *TESOL Quarterly*, *39*(3), 379–397.

Detey, S., & Nespoulous, J.-L. (2008). Can orthography influence second language syllabic segmentation? Japanese epenthetic vowels and French consonantal clusters. *Lingua*, *118*(1), 66–81.

Dupoux, E., Kakehi, K., Hirose, Y., Pallier, C., & Mehler, J. (1999). Epenthetic vowels in Japanese: A perceptual illusion? *Journal of Experimental Psychology: Human Perception and Performance*, *25*(6), 1568–1578.

Ellis, R. (2001). Introduction: Investigating form-focused instruction. *Language Learning*, *51*(s1), 1–46.

Elman, J. L., Diehl, R. L., & Buchwald, S. E. (1977). Perceptual switching in bilinguals. *The Journal of the Acoustical Society of America*, *62*(4), 971–974.

Elvin, J., Williams, D., & Escudero, P. (2016). Dynamic acoustic properties of monophthongs and diphthongs in Western Sydney Australian English. *The Journal of the Acoustical Society of America*, *140*(1), 576–581.

Escudero, P. (2005). *Linguistic perception and second language acquisition: Explaining the attainment of optimal phonological categorization*. LOT Dissertation Series 113.

Escudero, P. (2007). Second-language phonology: The role of perception. In *Phonology in Context* (pp. 109–134). Palgrave Macmillan.

Escudero, P. (2009). The linguistic perception of SIMILAR L2 sounds. In P. Boersma & S. Hamann (Eds.), *Phonology in Perception* (pp. 151–190). Mouton de Gruyter.

Escudero, P., Benders, T., & Lipski, S. C. (2009). Native, non-native and L2 perceptual cue weighting for Dutch vowels: The case of Dutch, German, and Spanish listeners. *Journal of Phonetics*, *37*(4), 452–465.

Escudero, P., Benders, T., & Wanrooij, K. (2011). Enhanced bimodal distributions facilitate the learning of second language vowels. *The Journal of the Acoustical Society of America*, *130*(4), EL206–EL212.

Escudero, P., & Bion, R. (2007). Modeling vowel normalization and sound perception as sequential processes. In J. Trouvain & W. J. Barry (Eds.), *Proceedings of the 16th International Congress of Phonetic Sciences* (pp. 2–5). Saarland University.

Escudero, P., & Boersma, P. (2003). Modelling the perceptual development of phonological contrasts with Optimality Theory and the Gradual Learning Algorithm. *University of Pennsylvania Working Papers in Linguistics*, *8*(1), 71–85.

Escudero, P., & Boersma, P. (2004). Bridging the gap between L2 speech perception research and phonological theory. *Studies in Second Language Acquisition*, *26*(4), 551–585.

Escudero, P., & Boersma, P. (2002). The subset problem in L2 perceptual development: Multiple-category assimilation by Dutch learners of Spanish. *Proceedings of the 26th Annual Boston University Conference on Language Development*, 208–219.

Escudero, P., & Wanrooij, K. (2010). The effect of L1 orthography on non-native vowel perception. *Language and Speech*, *53*(3), 343–365.

Flege, J. E. (1981). The phonological basis of foreign accent: A hypothesis. *TESOL Quarterly*, *15*(4), 443–455.

Flege, J. E. (1995). Second language speech learning: Theory, findings, and problems. In W. Strange (Ed.), *Speech perception and linguistic experience: Issues in cross-language research* (pp. 233–277). York Press.

Flege, J. E., & Eefting, W. (1987). Production and perception of English stops by native Spanish speakers. *Journal of Phonetics*, *15*, 67–83.

Flege, J. E., & Hillenbrand, J. M. (1987). Differential use of closure voicing and release burst as cue to stop voicing by native speakers of French and English. *Journal of Phonetics*, *15*, 203–208.

Flege, J. E., Munro, M. J., & MacKay, I. R. (1995). Factors affecting strength of perceived foreign accent in a second language. *The Journal of the Acoustical Society of America*, *97*(5), 3125–3134.

Flege, J. E., & Port, R. (1981). Cross-language phonetic interference: Arabic to English. *Language and Speech*, *24*(2), 125–146.

Fox, M. M., & Maeda, K. (1999). Categorization of American English vowels by Japanese speakers. In J. J. Ohala, Y. Hasegawa, M. Ohala, D. Granville, & A. C. Bailey (Eds.), *Proceedings of the 14th International Congress of Phonetic Sciences* (pp. 1437–1440). University of California.

Fox, R. A., & Jacewicz, E. (2009). Cross-dialectal variation in formant dynamics of American English vowels. *The Journal of the Acoustical Society of America*, *126*(5), 2603–2618.

Fridland, V. (2008). Patterns of /uw/, /ʊ/, and /ow/ fronting in Reno, Nevada. *American Speech*, *83*(4), 432–454.

Ganong, W. F. (1980). Phonetic categorization in auditory word perception. *Journal of Experimental Psychology: Human Perception and Performance*, *6*(1), 110–125.

García-Sierra, A., Diehl, R. L., & Champlin, C. A. (2009). Testing the double phonemic boundary in bilinguals. *Speech Communication*, *51*(4), 369–378.

García-Sierra, A., Ramírez-Esparza, N., Silva-Pereyra, J., Siard, J., & Champlin, C. A. (2012). Assessing the double phonemic representation in bilingual speakers of Spanish and English: An electrophysiological study. *Brain and Language*, *121*(3), 194–205.

Gass, S. M. (1988). Integrating research areas: A framework for second language studies. *Applied Linguistics*, *9*(2), 198–217.

Gluszek, A., & Dovidio, J. F. (2010). Speaking with a nonnative accent: Perceptions of bias, communication difficulties, and belonging in the United States. *Journal of Language and Social Psychology*, *29*(2), 224–234.

Gonzales, K., & Lotto, A. J. (2013). A bafri, un pafri: Bilinguals' pseudoword identifications support language-specific phonetic systems. *Psychological Science*, *24*(11), 2135–2142.

Goudbeek, M., Swingley, D., & Smits, R. (2009). Supervised and unsupervised learning of multidimensional acoustic categories. *Journal of Experimental Psychology: Human Perception and Performance*, *35*(6), 1913–1933.

Grosjean, F. (1989). Neurolinguists, beware! The bilingual is not two monolinguals in one person. *Brain and Language*, *36*(1), 3–15.

Grosjean, F. (2001). The bilingual's language modes. In J. Nicol (Ed.), *One Mind, Two Languages: Bilingual Language Processing* (pp. 1–22). Blackwell.

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.

Gussenhoven, C., & Jacobs, H. (2017). *Understanding phonology* (4th ed.). Routledge.

Hagiwara, R. (1997). Dialect variation and formant frequency: The American English vowels revisited. *The Journal of the Acoustical Society of America*, *102*(1), 655–658.

Harada, T. (2011). Does early foreign language learning benefit L2 pronunciation? Implications from immersion education. *Waseda Review of Education*, *25*(1), 1–14.

Harrington, J., Cox, F., & Evans, Z. (1997). An acoustic phonetic study of broad, general, and cultivated Australian English vowels. *Australian Journal of Linguistics*, *17*(2), 155–184.

Hattori, K. (2009). *Perception and production of English /r/-/l/ by adult Japanese speakers*. University College London.

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.

Hillenbrand, J. M., Clark, M., & Houde, R. (2000). Some effects of duration on vowel recognition. *The Journal of the Acoustical Society of America*, *108*(6), 3013–3022.

Hillenbrand, J. M., Getty, L., Clark, M., & Wheeler, K. (1995). Acoustic characteristics of American English vowels. *The Journal of the Acoustical Society of America*, *97*(5), 3099–3111.

Hirata, Y. (2004). Effects of speaking rate on the vowel length distinction in Japanese. *Journal of Phonetics*, *32*(4), 565–589.

Hirata, Y., & Tsukada, K. (2009). Effects of speaking rate and vowel length on formant frequency displacement in Japanese. *Phonetica*, *66*(3), 129–149.

Hisagi, M., Shafer, V. L., Strange, W., & Sussman, E. S. (2010). Perception of a Japanese vowel length contrast by Japanese and American English listeners: Behavioral and electrophysiological measures. *Brain Research*, *1360*, 89–105.

Holt, L. L., & Lotto, A. J. (2008). Speech perception within an auditory cognitive science framework. *Current Directions in Psychological Science*, *17*(1), 42–46.

Hosoda, M., & Stone-Romero, E. (2010). The effects of foreign accents on employmentrelated decisions. *Journal of Managerial Psychology*, *25*(2), 113–132.

Hume, E., & Johnson, K. (2001). A model of the interplay of speech perception and phonology. In E. Hume & K. Johnson (Eds.), *The role of speech perception in phonology* (pp. 3–26). Academic Press.

Hyman, L. M. (2001). The limits of phonetic determinism in phonology: *NC revisited. In E. Hume & K. Johnson (Eds.), *The role of spech perception in phonology* (pp. 141–185). Academic Press.

Ioup, G., Boustagui, E., El Tigi, M., & Moselle, M. (1994). Reexamining the critical period hypothesis: A case study of successful adult SLA in a naturalistic environment. *Studies in Second Language Acquisition*, *16*, 73–98.

Iverson, P., Kuhl, P., Akahane-Yamada, R., Diesch, E., Tohkura, Y., Kettermann, A., & Siebert, C. (2003). A perceptual interference account of acquisition difficulties for non-native phonemes. *Cognition*, *87*(1), B47–B57.

Jacquemot, C., Pallier, C., LeBihan, D., Dehaene, S., & Dupoux, E. (2003). Phonological grammar shapes the auditory cortex: A functional magnetic resonance imaging study. *The Journal of Neuroscience*, *23*(29), 9541–9546.

Jarosz, G. (2013). Learning with hidden structure in Optimality Theory and Harmonic Grammar: Beyond Robust Interpretive Parsing. *Phonology*, *30*, 27–71.

Johnson, K. (2008). Speaker normalization in speech perception. In D. B. Pisoni & R. E. Remez (Eds.), *The handbook of speech perception* (pp. 363–389). Blackwell.

Kalashnikova, M., Carignan, C., & Burnham, D. (2017). The origins of babytalk: Smiling, teaching or social convergence? *Royal Society Open Science*, *4*(8), 170306.

Kartushina, N., & Frauenfelder, U. H. (2014). On the effects of L2 perception and of individual differences in L1 production on L2 pronunciation. *Frontiers in Psychology*, *5*, 1246.

Kartushina, N., Frauenfelder, U. H., & Golestani, N. (2016). How and when does the second language influence the production of native speech sounds: A literature review. *Language Learning*, *66*(S2), 155–186.

Kawahara, H. (2006). STRAIGHT, exploitation of the other aspect of VOCODER:

Perceptually isomorphic decomposition of speech sounds. *Acoustical Science and Technology*, *27*(6), 2006.

Kawahara, S., & Braver, A. (2013). The phonetics of multiple vowel lengthening in Japanese. *Open Journal of Modern Linguistics*, *3*(2), 141–148.

Kawahara, S., Erickson, D., & Suemitsu, A. (2017). The phonetics of jaw displacement in Japanese vowels. *Acoustical Science and Technology*, *38*(2), 99–107.

Keating, P. A., & Huffman, M. K. (1984). Vowel variation in Japanese. *Phonetica*, *41*(4), 191–207.

Kerr, N. L. (1998). HARKing: Hypothesizing after the results are known. *Personality and Social Psychology Review*, *2*(3), 196–217.

Klatt, D. H., & Klatt, L. C. (1990). Analysis, synthesis, and perception of voice quality variations among female and male talkers. *The Journal of the Acoustical Society of America*, *87*(2), 820–857.

Kriengwatana, B., Terry, J., Chládková, K., & Escudero, P. (2016). Speaker and accent variation are handled differently: Evidence in native and non-native listeners. *PLOS ONE*, *11*(6), e0156870.

Kuhl, P. (2000). A new view of language acquisition. *Proceedings of the National Academy of Sciences of the United States of America*, *97*(22), 11850–11857.

Kuhl, P. (2004). Early language acquisition: Cracking the speech code. *Nature Reviews Neuroscience*, *5*(11), 831–843.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, *82*(13), 1–26.

Labov, W., Ash, S., & Boberg, C. (2006). *The atlas of North American English: Phonetics, phonology and sound change: A multimedia reference tool*. Walter De Gruyter.

Ladefoged, P., & Maddieson, I. (1996). *The sounds of the world's languages*. Wiley-Blackwell.

Lenneberg, E. H. (1967). *The biological foundations of language*. Wiley.

Lev-Ari, S., & Peperkamp, S. (2013). Low inhibitory skill leads to non-native perception and production in bilinguals' native language. *Journal of Phonetics*, *41*(5), 320–331.

Liberman, A. M., Cooper, F. S., Shankweiler, D. P., & Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological Review*, *74*(6), 431–461.

Liberman, A. M., & Mattingly, I. G. (1985). The motor theory of speech perception revised. *Cognition*, *21*(1), 1–36.

Lindau, M. (1980). The story of /r/. *The Journal of the Acoustical Society of America*, *67*, S27.

Lisker, L. (1999). Perceiving final voiceless stops without release: Effects of preceding monophthongs versus nonmonophthongs. *Phonetica*, *56*(1–2), 44–55.

Lyster, R., & Saito, K. (2010). Interactional feedback as instructional input: A synthesis of classroom SLA research. *Language, Interaction and Acquisition*, *1*(2), 276–297.

MacKain, K., Best, C. T., & Strange, W. (1981). Categorical perception of English /r/ and /l/ by Japanese bilinguals. *Applied Psycholinguistics*, *2*(4), 369–390.

Maekawa, K. (2003). Corpus of Spontaneous Japanese: Its design and evaluation. *Proceedings of the ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition*, 7–12.

Maria, A. (1997). Introduction to modeling and simulation. In S. Andradóttir, K. J. Healy, D. H. Withers, & B. L. Nelson (Eds.), *Proceedings of the 1997 Winter Simulation Conference* (pp. 7–13).

Mazuka, R., Cao, Y., Dupoux, E., & Christophe, A. (2011). The development of a phonological illusion: A cross-linguistic study with Japanese and French infants. *Developmental Science*, *14*(4), 693–699.

Mazuka, R., Hasegawa, M., & Tsuji, S. (2014). Development of non-native vowel discrimination: Improvement without exposure. *Developmental Psychobiology*, *56*(2), 192–209.

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.

McArthur, T., & McArthur, R. (Eds.). (2005). *Concise Oxford companion to the English language*. Oxford University Press.

McCarthy, J. J. (2007). What is Optimality Theory? *Language and Linguistics Compass*, *1*(4), 260–291.

McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology*, *18*(1), 1–86.

McQueen, J. M., & Cutler, A. (2010). Cognitive processes in speech perception. In W. J. Hardcastle, J. Laver, & F. E. Gibbon (Eds.), *The Handbook of Phonetic Sciences* (pp. 489–520). Blackwell.

Mesgarani, N., Cheung, C., Johnson, K., & Chang, E. F. (2014). Phonetic feature encoding in human superior temporal gyrus. *Science*, *343*(6174), 1006–1010.

Miyawaki, K., Verbrugge, R., Strange, W., Liberman, A. M., Jenkins, J. J., & 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.

Mora, J. C., & Nadeu, M. (2012). L2 effects on the perception and production of a native vowel contrast in early bilinguals. *International Journal of Bilingualism*, *16*(4), 484–500.

Morrison, G. S. (2007). Logistic regression modelling for first-and second-language perception data. In M. J. Solé, P. Prieto, & J. Mascaró (Eds.), *Segmental and prosodic issues in Romance phonology* (pp. 219–236). John Benjamins.

Morrison, G. S. (2002). Perception of English /i/ and /ɪ/ by Japanese and Spanish listeners: Longitudinal results. In G. S. Morrison & L. Zsoldos (Eds.), *Proceedings of the Northwest Linguistic Conference 2002* (pp. 29–48). Simon Fraser University Linguistics Graduate Student Association.

Morrison, G. S., & Assmann, P. F. (Eds.). (2013). *Vowel inherent spectral change*. Springer.

Morrison, G. S., & Nearey, T. M. (2007). Testing theories of vowel inherent spectral change. *The Journal of the Acoustical Society of America*, *122*(1), EL15–EL22.

Munro, M. J., & Derwing, T. M. (1995). Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. *Language Learning*, *45*(1), 73–97.

Nearey, T. M., & Assmann, P. F. (1986). Modeling the role of inherent spectral change in vowel identification. *The Journal of the Acoustical Society of America*, *80*(5), 1297–1308.

Nishi, K., Strange, W., Akahane-Yamada, R., Kubo, R., & Trent-Brown, S. A. (2008). Acoustic and perceptual similarity of Japanese and American English vowels. *The Journal of the Acoustical Society of America*, *124*(1), 576–588.

Nogita, A., Yamane, N., & Bird, S. (2013). The Japanese unrounded back vowel /ɯ/ is in fact unrounded central/front [ʉ - ʏ]. *Ultrafest VI Program and Abstract Booklet*, 39–42.

Norris, D., McQueen, J. M., & Cutler, A. (2000). Merging information in speech recognition: Feedback is never necessary. In *Behavioral and Brain Sciences* (Vol. 23, Issue 3, pp. 299–325).

Norris, D., McQueen, J. M., & Cutler, A. (2003). Perceptual learning in speech. *Cognitive Psychology*, *47*(2), 204–238.

Oh, G. E., Guion-Anderson, S., Aoyama, K., Flege, J. E., Akahane-Yamada, R., & Yamada, T. (2011). A one-year longitudinal study of English and Japanese vowel production by Japanese adults and children in an English-speaking setting. *Journal of Phonetics*, *39*(2), 156–167.

Olson, D. J. (2019). Feature acquisition in secondlLanguage phonetic development: Evidence from phonetic training. *Language Learning*, *69*(2), 366–404.

Otaki, Y. (2012). A phonological account of vowel epenthesis in Japanese loanwords: Synchronic and diachronic perspectives. *Phonological Studies*, *15*, 35–42.

Pajak, B., & Levy, R. (2014). The role of abstraction in non-native speech perception. *Journal of Phonetics*, *46*, 147–160.

Patkowski, M. S. (1990). Age and accent in a second language: A reply to James Emil Flege. *Applied Linguistics*, *11*(1), 73–89.

Peirce, J. W. (2007). PsychoPy-Psychophysics software in Python. *Journal of Neuroscience Methods*, *162*(1–2), 8–13.

Penfield, W., & Roberts, L. (1959). *Speech and brain mechanisms*. Princeton University Press.

Peterson, G. E., & Barney, H. L. (1952). Control methods used in a study of the vowels. *The Journal of the Acoustical Society of America*, *24*(2), 175–184.

Peterson, G. E., & Lehiste, I. (1960). Duration of syllable nuclei in English. *The Journal of the Acoustical Society of America*, *32*(6), 693–703.

Pike, K. L. (1947). On the phonemic status of English diphthongs. *Language*, *23*(2), 151–159.

Prince, A., & Smolensky, P. (1993). Optimality Theory: Constraint interaction in generative grammar. *Rutgers University Center for Cognitive Science Technical Report*, *2*.

Prince, A., & Smolensky, P. (2004). *Optimality Theory: Constraint interaction in generative grammar*. Blackwell.

R Core Team. (2019). *R: A language and environment for statistical computing*.

Salameh, M. Y. B., & Abu-Melhim, A.-R. (2014). The phonetic nature of vowels in Modern Standard Arabic. *Advances in Language and Literary Studies*, *5*(4), 60–67.

Scarborough, R., Dmitrieva, O., Hall-Lew, L., Zhao, Y., & Brenier, J. (2007). An acoustic study of real and imagined foreigner-directed speech. In J. Trouvain & W. J. Barry (Eds.), *Proceedings of the 16th International Congress of Phonetic Sciences* (Vol. 121, pp. 2165–2168). Saarland University.

Schwartz, B. D., & Sprouse, R. A. (1996). L2 cognitive states and the Full Transfer/Full Access model. *Second Language Research*, *12*(1), 40–72.

Selinker, L. (1972). Interlanguage. *International Review of Applied Linguistics in Language Teaching*, *10*(3), 209–231.

Shaw, J. A., & Kawahara, S. (2017). The lingual articulation of devoiced /u/ in Tokyo Japanese. *Journal of Phonetics*, *66*, 100–119.

Sheldon, A., & Strange, W. (1982). The acquisition of /r/ and /l/ by Japanese learners of English: Evidence that speech production can precede speech perception. *Applied Psycholinguistics*, *3*(3), 243–261.

Shinohara, Y., & Iverson, P. (2018). High variability identification and discrimination training for Japanese speakers learning English /r/-/l/. *Journal of Phonetics*, *66*, 242–251.

Simonet, M. (2016). The phonetics and phonology of bilingualism. In *Oxford Handbooks Online* (pp. 1–25). Oxford University Press.

Skehan, P. (1991). Individual differences in second language learning. *Studies in Second Language Acquisition*, *13*, 275–298.

Smith, M. S. (1993). Input enhancement in instructed SLA: Theoretical bases. *Studies in Second Language Acquisition*, *15*(2), 165–179.

Smolensky, P. (1996). On the comprehension/production dilemma in child language. *Linguistic Inquiry*, *27*, 720–731.

Strange, W. (1992). Learning non-native phoneme contrasts: Interactions among subject, stimulus, and task variables. In Y. Tohkura, E. Vatikiotis-Bateson, & Yoshinori Sagisaka (Eds.), *Speech perception, production and linguistic structure* (pp. 197–219).

Strange, W. (1995). Cross-language studies of speech perception: A historical review. In W. Strange (Ed.), *Speech perception and linguistic experience: Issues in cross-language research* (pp. 3–45). York Press.

Strange, W., Akahane-Yamada, R., Kubo, R., Trent-Brown, S. A., Nishi, K., & Jenkins, J. J. (1998). Perceptual assimilation of American English vowels by Japanese listeners. *Journal of Phonetics*, *26*(4), 311–344.

Strange, W., Akahane-Yamada, R., Kubo, R., Trent, S. A., & Nishi, K. (2001). Effects of consonantal context on perceptual assimilation of American English vowels by Japanese listeners. *The Journal of the Acoustical Society of America*, *109*(4), 1691–1704.

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

Tajima, K., Kato, H., Rothwell, A., Akahane-Yamada, R., & Munhall, K. G. (2008). Training English listeners to perceive phonemic length contrasts in Japanese. *The Journal of the Acoustical Society of America*, *123*(1), 397–413.

Takagi, N. (1993). *Perception of American English /r/ and /l/ by adult Japanese learners of English: A unified view*. University of California-Irvine.

ter Schure, S. M. M., Junge, C. M. M., & Boersma, P. (2016). Semantics guide infants' vowel learning: Computational and experimental evidence. *Infant Behavior and Development*, *43*, 44–57.

Tesar, B., & Smolensky, P. (1998). Learnability in Optimality Theory. *Linguistic Inquiry*, *29*(2), 229–268.

The MathWorks Inc. (2018). *MATLAB (Version R2018b) [Computer software]*.

Trubetzkoy, N. (1939). *Grundzüge der phonologie*. Travaux du Cercle Linguistique de Prague.

Trubetzkoy, N. (1969). *Principles of phonology*. University of California Press.

Tsukada, K. (2012). Comparison of native versus nonnative perception of vowel length contrasts in Arabic and Japanese. *Applied Psycholinguistics*, *33*(3), 501–516.

Tsukada, K. (2010). Pattern of perceptual assimilation of Japanese vowels to Australian English vowels: Comparison of learners and non-learners of Japanese. In M. Tabain, J. Fletcher, D. Grayden, J. Hajek, & A. Butcher (Eds.), *Proceedings of the 13th Australasian International Conference on Speech Science and Technology* (p. 29). Causal Productions.

Tsukada, K., Cox, F., Hajek, J., & Hirata, Y. (2018). Non-native Japanese learners' perception of consonant length in Japanese and Italian. *Second Language Research*, *34*(2), 179–200.

Umeda, N. (1975). Vowel duration in American English. *The Journal of the Acoustical Society of America*, *58*(2), 434–445.

Uther, M., Knoll, M. A., & Burnham, D. (2007). Do you speak E-NG-L-I-SH? A comparison of foreigner- and infant-directed speech. *Speech Communication*, *49*(1), 2–7.

van Leussen, J.-W., & Escudero, P. (2015). Learning to perceive and recognize a second language: The L2LP model revised. *Frontiers in Psychology*, *6*, 1000.

VanPatten, B., & Benati, A. G. (2015). *Key terms in second language acquisition* (2nd ed.). Bloomsbury Academic.

Ventureyra, V. A. G., Pallier, C., & Yoo, H. Y. (2004). The loss of first language phonetic perception in adopted Koreans. *Journal of Neurolinguistics*, *17*(1), 79–91.

Verbrugge, R. R., Strange, W., Shankweiler, D. P., & Edman, T. R. (1976). What information enables a listener to map a talker's vowel space? *The Journal of the Acoustical Society of America*, *60*(1), 198–212.

Weiand, K. (2007). Implementing Escudero's model for the SUBSET problem. *Rutgers Optimality Archive*, *913*, 1–24.

Weinberger, S. (1987). The influence of linguistic context of syllable structure simplification. In G. Ioup & S. Weinberger (Eds.), *Interlanguage phonology* (pp. 401–418). Newbury House.

Weinreich, U. (1957). On the description of phonic interference. *Word*, *13*(1), 1–11.

Werker, J. F., & Logan, J. S. (1985). Cross-language evidence for three factors in speech perception. *Perception & Psychophysics*, *37*(1), 35–44.

Werker, J. F., & Yeung, H. H. (2005). Infant speech perception bootstraps word learning. *Trends in Cognitive Sciences*, *9*(11), 519–527.

Whang, J., Yazawa, K., & Escudero, P. (2019). Perception of Japanese vowel length by Australian English listeners. In S. Calhoun, P. Escudero, M. Tabain, & P. Warren (Eds.), *Proceedings of the 19th International Congress of Phonetic Sciences* (p. 265). Australasian Speech Science and Technology Association Inc.

Williams, D., Escudero, P., & Gafos, A. (2018). Spectral change and duration as cues in Australian English listeners' front vowel categorization. *The Journal of the Acoustical Society of America*, *144*(3), EL215-221.

Williams, L. (1977). The perception of stop consonant voicing by Spanish-English bilinguals. *Perception & Psychophysics*, *21*(4), 289–297.

Yang, J., Fox, R. A., & Jacewicz, E. (2015). Vowel development in an emergent Mandarin-English bilingual child: A longitudinal study. *Journal of Child Language*, *42*(5), 1125–1145.

Yazawa, K., & Kondo, M. (2019). Acoustic characteristics of Japanese short and long vowels: Formant displacement effect revisited. In S. Calhoun, P. Escudero, M. Tabain, & P. Warren (Eds.), *Proceedings of the 19th International Congress of Phonetic Sciences* (p. 100). Australasian Speech Science and Technology Association Inc.

Yazawa, K., Whang, J., Kondo, M., & Escudero, P. (2019). Language-dependent cue weighting: An investigation of perception modes in L2 learning. *Second Langauge Research*, 1–25. https://doi.org/10.1177/0267658319832645

Zeileis, A., & Hothorn, T. (2002). Diagnostic checking in regression relationship. *R News*, *2*(3), 7–10. https://cran.r-project.org/doc/Rnews/