¿Estoy yo obligado, a dicha, siendo, como soy, caballero,
a conocer y destinguir los sones . . . ?

El Quijote, Tercera parte, Capítulo XX.

University of Alberta

L1 & L2 Production and Perception of English and Spanish Vowels: A Statistical Modelling Approach

by

Geoffrey Stewart Morrison ©

A thesis submitted to the Faculty of Graduate Studies and Research
in partial fulfilment of the requirements for the degree of

Doctor of Philosophy

Department of Linguistics

Edmonton, Alberta
Fall 2006

Library and
Archives Canada

Published Heritage
Branch

395 Wellington Street
Ottawa ON K1A 0N4
Canada

Bibliothèque et
Archives Canada

Direction du
Patrimoine de l'édition

395, rue Wellington
Ottawa ON K1A 0N4
Canada

Your file Votre référence
ISBN: 978-0-494-23085-5
Our file Notre référence
ISBN: 978-0-494-23085-5

Abstract

The present study explores L1-Spanish speakers' learning of the English /i/–/ɪ/ contrast via acoustic analysis of vowel productions and perception of synthetic stimuli. L1-English, L1-Spanish, and L2-Spanish perception and production are also explored. The vowels examined are English /i/, /ɪ/, adjacent English /e/, /ɛ/, and Spanish /i/, /ei/, /e/. The acoustic properties examined are vowel duration, and initial and final first- and second-formant values. Diphthongisation / vowel inherent spectral change (VISC) is an important factor in the perception of /ɪ/ in the Canadian English dialect examined. Consistent with current theories that L1 and L2 learners build speech sound categories on the basis of the statistical distribution of acoustic properties, discriminant analysis and logistic regression are used to build models of production and perception data. Models trained on monolingual Spanish data predict that Spanish listeners just beginning to learn English will perceive most instances of English /i/ as Spanish /i/, and most instances of English /ɪ/ as Spanish /e/; hence English /i/ and /ɪ/ will be easily distinguished. However, cross-sectional and longitudinal data from L1- Spanish learners of English suggest that they confuse English /i/ and /ɪ/, and begin to distinguish them via a multidimensional category-goodness-difference assimilation to Spanish /i/. A minority of L2-English learners are hypothesised to label more-Spanish-/i/-like vowels (short duration, low F1, zero VISC) as English /i/, and less-Spanish-/i/-like vowels (longer duration, higher F1, converging VISC) as English /ɪ/. Since spectral cues are used in the same direction by L1-English listeners and are most important for L1-English listeners, this immediately results in relatively L1-English-like perception. However, the results for

the majority of L2-English participants were consistent with them beginning with the reverse labelling, and, since only duration cues are positively correlated with L1-English speakers' productions, increased exposure to English leads to a greater weighting for duration cues. Eventually L1-English-like use of spectral cues may be bootstrapped off duration cues. The initial association of English /ɪ/ with good examples of Spanish /i/ is hypothesised to be due to (mis)education/orthography, rather than phonetic/perceptual factors.

Acknowledgments

I would like to thank the members of my committee (Keith Johnson, John Hogan, Megan Hodge, Robert Kirchner, and Terry Nearey) for all their hard work, and especially Terry Nearey and Robert Kirchner with whom I have worked most closely at the University of Alberta. I came to the University of Alberta in order to work with Terry, and after three years here am very happy with that decision. I have learnt a great deal from him about statistical modelling of speech data, and about being an academic researcher and teacher. I am also very grateful for all the opportunities that have been offered to me by the department and the university. I must thank all the anonymous participants who took part in the research, I only wish I could do more to express my appreciation for their essential contribution. Several people assisted with practical tasks such as reading instructions, screening stimuli, recruiting participants, and commenting on drafts of parts of the dissertation; thanks to: Blanca Casado, Bronwen Evans, Carlos Bengoa, Chunling Zhang, Isabel Klint, Juan Pablo Villamor, Natalía Gómez, Paul Iverson, Ron Thomson, Vianey Varela, and Wolf Wikeley. Thanks also to Gorka Elordieta and Lourdes Oñederra for the use of the soundbooth at the University of the Basque Country. Several other people have encouraged me along the way, and I would like to mention in particular Catherine Louise Boucher, Christine Shea, Eta Schneiderman, Gill Morrison, Ian MacKay, Ilana Mezhevich, Joanne Paradis, Maureen Morrison, Michael Kiefte, Murray Munro, Ocke-Schwen Bohn, Paola Escudero, Peter Louwerse, Ryan Klint, and Stewart Morrison. My research was supported financially by a Social Sciences and Humanities Research Council of Canada Doctoral Fellowship, a University of Alberta Walter H Johns Graduate Fellowship, and a Social Sciences and Humanities Research Council of Canada Standard Research Grant held by Terrance M. Nearey.

Table of Contents

List of Tables

List of Figures

1. Introduction

The present study investigates the perception and production of English and Spanish vowels by English and Spanish speakers. It seeks to advance our understanding of second-language (L2) speech learning by collecting a substantial body of data which will be used to better describe first-language (L1) Spanish speakers' L2-English learning and L1-English speakers' L2-Spanish learning. The findings may ultimately have practical implications for language teaching and learning.

1.1 Spanish Vowels and Canadian-English Vowels

Most dialects of Spanish have five monophthongs /i, e, a, o, u/. Examples of mean first and second formant (F1 and F2) values for Spanish monophthongs are given in Figure 1.1 (blue symbols within circles, data from Martínez Celdrán, 1995).[1] Spanish also has a number of diphthongs including /ai, ia, au, oi, ei, ie, eu, ue/ characterised by initial and final targets which approximate pairs of monophthongs, a stable rate of transition between initial and final targets, and longer duration associated with the target with higher F1 relative to the target with lower F1 (Borzone de Manrique, 1979; see also Hualde & Prieto, 2002).[2]

[1] Other acoustic studies of Spanish monophthong production include Álvarez González (1980), Cervera, Miralles, & González-Álvarez (2001), Godínez (1978), Guirao & Borzone de Manrique (1975), Madrid Servín & Marín Rodríguez (2001), Quilis & Esgueva (1983), and Skelton (1969). Perception studies using synthetic vowels include Álvarez González (1980), Fernández Planas (1993), García Bayonas (2004), and León Valdés (1998). Martínez Megar (1990) presented an acoustic study of monophthong production in an Andalucian dialect with ostensively ten monophthongs.

[2] Borzone de Manrique (1979) reported durations for steady-state portions of the diphthong associated with the initial and final targets, à la Lehiste & Peterson (1961). Spanish also has sequences of vowels in hiatus, which differ from diphthongs in that longer duration is associated with the target with lower F1 relative to the target with higher F1.

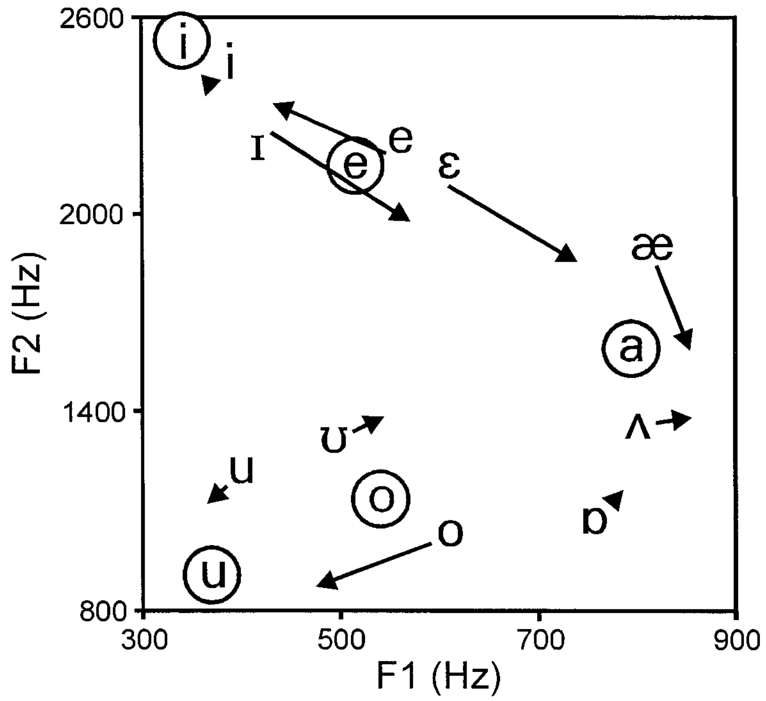


Figure 1.1 Symbols within circles: Mean F1–F2 values for Peninsular Spanish monophthongs measured in the middle of vowels in isolated words. Data from Martínez Celdrán (1995), averaged over five male and five female speakers. Symbols with arrows: Mean F1–F2 values for Canadian English nominal monophthongs and phonetic diphthongs measured at the beginning and end of isolated vowels. Data from Nearey & Assmann (1986), averaged over five male and five female speakers.

The number of vowels in English varies from dialect to dialect; General Canadian English[3] has eight stressed vowels that are traditionally called monophthongs /i, ɪ, ɛ, æ, ɒ, ʌ, ʊ, u/, and two stressed vowels which are traditionally called phonetic diphthongs /e, o/, the latter often transcribed using digraphs such as [eɪ, oʊ] (the start points for the phonetic diphthongs do not correspond perceptually to any of the traditional monophthongs in this dialect). Canadian English also has three vowels traditionally called true diphthongs /aɪ, aʊ, ɔɪ/, and a number of glide plus vowel sequences including /jɛ, ju, wi, wʌ/. In addition to differences in F1 and F2, Canadian-English vowels also differ in terms of duration; for example, all else being equal, Canadian-English /i/ is longer than Canadian-English /ɪ/. Also, most Canadian-English nominal monophthongs are actually produced with vowel inherent

[3] *General Canadian English* is the dialect of English spoken in Western and Central Canada. The dialect is reported to be relatively homogeneous across its geographical expanse, especially in urban settings (Avis, 1973, 1975; Chambers, 1991; Wells, 1982), although there is little instrumental evidence to support these reports.

spectral change (VISC). The green arrows in Figure 1.1 indicate spectral change from the beginning to the end of Canadian-English nominal monophthongs and phonetic diphthongs. VISC has been found to be relevant for L1-Canadian-English listeners' perception; listeners' correct-identification rates deteriorate when stimuli do not include VISC information, and higher correct classification rates and higher correlation with listeners' responses is obtained by statistical models which make use of dynamic spectral information compared to models which only use static spectral values (Andruski & Nearey, 1992; Assmann, Nearey, & Hogan, 1982; Nearey, 1995; Nearey & Assmann, 1986). A review of research on VISC is presented in Appendix 1.

Spanish speakers learning Canadian English are therefore faced with learning to differentiate a larger number of nominal monophthongs. They are also faced with learning to make appropriate use of duration and VISC, acoustic cues which have not been reported to be relevant for perceptually distinguishing Spanish monophthongs.[4]

1.2 Speech-Learning Theories

The discussion below is a synthesis and interpretation of current theories of L1 and L2 speech learning theories. The interpretation is heavily influenced by Nearey & Hogan's (1986) Normal A Posteriori Probability (NAPP) Model, which relates categorisation of speech sounds in perception to the distribution of acoustic properties from speech production.

1.2.1 Theories of L1 learning

Recent theories of L1-speech learning such as the Native Language Magnet model (NLM, Kuhl & Iverson, 1995; Iverson & Kuhl, 2000), the Native Language Neural Commitment hypothesis (NLNC, Kuhl, 2004, see also Kuhl, 2000), the Processing Rich Information from Multidimensional Interactive Representations framework (PRIMIR, Werker & Curtin, 2005), and the Linguistic Perception model (LP, Escudero, 2005, chap. 2) posit that infants establish speech-sound categories on the basis of the distribution of

[4] Differences in the duration of Spanish monophthong productions are small, but high vowels are consistently reported to be shorter than non-high vowels (Bradlow, 1993; Cervera, Miralles, & González-Álvarez, 2001; Marín Gálvez, 1995; Vaquero & Guerra, 1992).

acoustic properties in the language(s) to which they are exposed (see also Maye & Weiss, 2003; Maye, Werker, & Gerken, 2002).

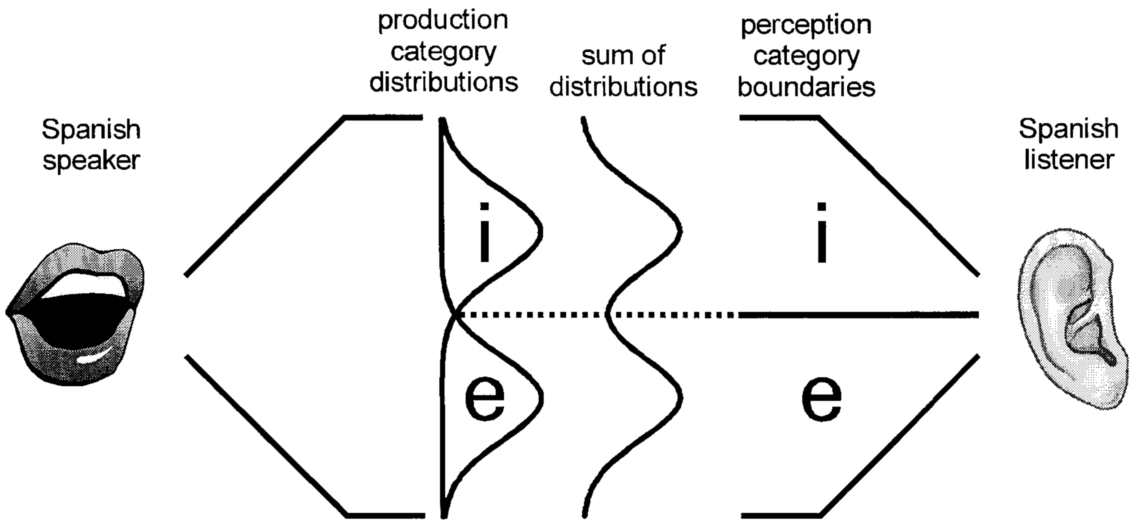

production category distributions
sum of distributions
perception category boundaries
Spanish speaker
Spanish listener
i
e
i
e

Figure 1.2 A stylised one-dimensional illustration of L1 speech perception learning, see text. The dotted line represents the location of a boundary which optimally categorises the Spanish vowel productions.

A stylised one-dimensional illustration of L1 speech perception learning is given in Figure 1.2. The single dimension runs vertically on the page and may be thought of as a representation of either traditional vowel height, or the first formant (F1) for vowels, with lower F1 values towards the top. An L1-Spanish speaker produces Spanish vowels. Each Spanish vowel category produced by the speaker has a distribution of F1 values, most instances of a vowel category will fall near the most prototypical F1 value for that category, but some instances will have more peripheral F1 values, and the tails of the distributions will have some overlap. The L1-Spanish learner is exposed to the sum of the distributions of the Spanish vowels (produced in different contexts by many speakers), infers the multimodal distribution of F1 values, and establishes a category corresponding to each mode. The boundaries between categories are assumed to be optimal so that most Spanish vowel productions are correctly classified by the mature L1-Spanish speech-perception system. The perceptual boundary between Spanish /i/ and /e/ is located at the F1 value at which, in the sum of the distributions of the Spanish vowel productions, there is a valley between the peaks corresponding to the Spanish /i/ and /e/ categories. Note that categories are established prior to lexical learning so that category formation is based on unlabelled input, and infants must construct categories purely on the basis of the multimodal distribution of acoustic properties in the input.[5] Once categories have been established and the L1-learner establishes a lexicon, if they initially misclassify a speech sound on the basis of incoming acoustic information, the misclassification can at least potentially be flagged on the basis of semantic and contextual information, and subsequent learning can therefore be based on labelled input.

1.2.2 Theories of L2 learning

Theories of L2-speech learning such as the Speech Learning Model (SLM, Flege, 1988, 1995, 2003), and the Second Language Linguistic Perception model (L2LP, Escudero, 2005, chap. 3) posit that L2 learners have access to the same processes and mechanisms which were active when they learnt their L1. This leads to the prediction that L2 learners will be able to develop L2-speech-sound categories on the basis of the distribution of acoustic properties in the L2 input; however, L2 learning differs from L1 learning in that the starting state for L2 learning is the existing L1-speech-perception system, whereas the starting state for L1 learning consists only of natural biases in the auditory processing system. At the initial state of L2 learning, learners perceive the L2 speech sounds through the filter of their existing L1-speech-perception system, and the existing speech-perception system interferes with their ability to infer the distribution of the acoustic properties of the L2 speech sounds.

A stylised one-dimensional illustration of the initial state of L2 speech perception is given in Figure 1.3. An L1-English speaker produces English vowels. Each English vowel category produced by the speaker has a distribution of F1 values. The L1-Spanish listener categorises each vowel they hear using their L1-Spanish speech-perception system. Most or all instances of the English /i/ category are perceptually categorised as instances of Spanish /i/ – instances of English /i/ are assimilated to Spanish /i/. Most instances of English /ɛ/ are assimilated to Spanish /e/. Most instances of English /ɪ/ are assimilated to Spanish /e/, but some are assimilated to Spanish /i/. The L1-Spanish listener will easily distinguish instances

[5] Some theories such as the NLM posit that biases in the auditory processing system provide natural boundaries and view the learning task is to learn which natural boundaries to ignore on the basis of the distribution of the acoustic input.

of English /i/ and /ɛ/ by mapping the English categories onto the Spanish /i/ and /e/ categories; however, an instance of English /ɪ/ will be difficult to distinguish from an instance of one of the other two English vowel categories when they are both assimilated to the same Spanish category.

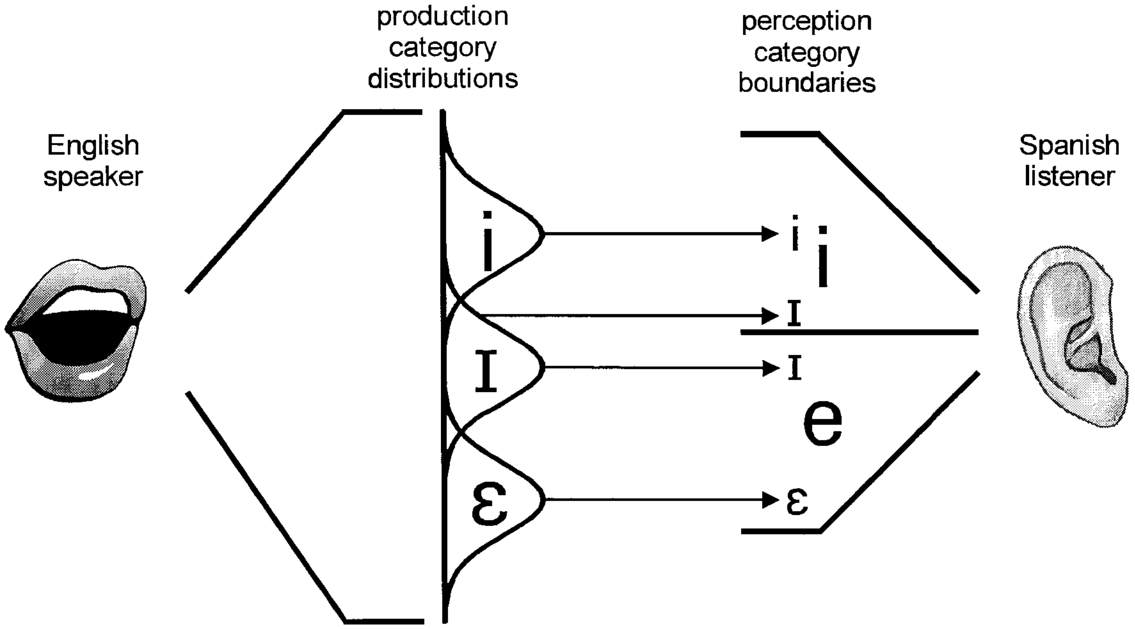

production category distributions
perception category boundaries
English speaker
Spanish listener
i
ɪ
ɛ
i
ɪ
ɪ
ɛ
i
e

Figure 1.3 A stylised one-dimensional illustration of the initial state of L2 speech perception. Arrows represent examples of individual instances of English vowels.

The Perceptual Assimilation Model (PAM, Best, 1994, 1995a, 1995b) describes and gives names to various patterns of assimilation of pairs of L2 speech sounds to L1-speech-sound categories.

– In the stylised illustration, almost all instances of English /i/ are assimilated to Spanish /i/, and almost all instances of English /ɛ/ are assimilated to Spanish /e/. In PAM terminology, instances of English /i/ and /ɛ/ undergo *two-category assimilation* to Spanish /i/ and /e/ respectively, and are predicted to be easily distinguished by L2 listeners.

– Instances of English /ɪ/ and /ɛ/ which are assimilated to Spanish /e/ are on average equally far from the prototypical F1 values for Spanish /e/ and will therefore be perceived as equally good (or equally poor) examples of Spanish /e/. In PAM terminology, instances of English /ɪ/ and /ɛ/ undergo *single-category assimilation* to Spanish /e/, and are predicted to be difficult to distinguish for L2 listeners.

– Compared to instances of English /i/ which are assimilated to Spanish /i/, instances of English /ɪ/ which are assimilated to Spanish /i/ are further from the prototypical F1 values for Spanish /i/, and the latter are therefore more likely to be noticed as deviant examples of the Spanish /i/ category. In PAM terminology, instances of English /i/ and /ɪ/ undergo *category-goodness-difference assimilation* to Spanish /i/, and are predicted to be relatively easily distinguished by L2 listeners.

The PAM terminology is useful for qualitative description (and can be expanded, e.g., Escudero & Boersma, 2002); however, numerical or graphical representations are needed to portray subtleties such as the proportion of instances of English /ɪ/ assimilated to Spanish /i/ and the proportion assimilated to Spanish /e/, and the distance from the prototypical F1 values for Spanish /i/ of an instance of English /ɪ/. Although Best initially described the PAM in a direct realist framework, I have interpreted it here in a psychoacoustic framework. I have also been careful to discuss assimilation in terms of an instance of a production of an L2 category being perceptually classified as an L1 category, that is a physical entity generated on the basis of an L2 mental construct is decoded as representing an L1 mental construct. Since some instances of the English /ɪ/ category are assimilated to the Spanish /i/ category and some to Spanish /e/ category, one cannot talk about assimilation of L2 categories to L1 categories except in relative terms.

Having established a picture of what the initial state for L2 learning might look like, we turn our attention to how the L2 learner might go about learning to perceive the speech sounds of the L2. In the stylised illustration, in order to perceive English in a more L1-English-like manner, the L1-Spanish L2-English listener must learn at least one new category and position category boundaries in appropriate locations for distinguishing the three English vowel categories.

Flege's SLM (1988, 1995, 2003) posits that the probability that a *new* L2-category will be formed depends on the perceived phonetic dissimilarity between the L2 sound and the closest L1 sound. If there are insufficient perceived differences between an L2 sound and the closest L1 sound (the sounds are *identical* or *similar*), then the L2 sound will be treated as equivalent to the L1 sound. The L1 and the L2 sound will be linked as a *diaphone*, a single perceptual category used for both sounds. Diaphones are predicted to take on a mixture of the properties of the L1 and L2 sounds, eventually resulting in perception which is intermediate between that of a monolingual speaker of the L1 and a monolingual speaker of the L2. In contrast the properties of a new L2 category are based only on the L2 sound and are predicted to match the L2 norms for that sound; however, the SLM posits a single phonological space for both L1 and L2 sounds (including diaphones and new L2 sounds), and in order to maintain a contrast between all the L1 and L2 sounds, new sounds may be *deflected* away from L2 norms or L2 listeners may weight perceptual cues differently compared to L2 norms (e.g., the L2 learner may make more use of duration and less use of spectral cues). A problem for the SLM is determining a measure of phonetic similarity between L1 and L2 sounds. One approach which could be adopted is to examine the degree of overlap in the distributions of the L1 and L2 categories: In the stylised illustration, more instances of English /ɪ/ are assimilated to Spanish /e/ than to Spanish /i/, and English /ɪ/ is therefore relatively more similar to Spanish /e/ than to Spanish /i/. The relative similarity could be quantified according to the proportion of instances of English /ɪ/ assimilated to each of the Spanish vowel categories. However, instances of an L2 sound may be assimilated to an L1 sound but be out on the tail of the distribution of the L1 sound and be noticeably deviant members of that category. Phonetic similarity could be measured in terms of the probability density function (PDF) for the L1 category. In the stylised illustration, instances of English /ɛ/ are on the tail of the distribution of the Spanish /i/ category, and may therefore have sufficient perceived phonetic dissimilarity to seed a new L2-English /ɛ/ category. What would also be needed to convert the SLM into a quantitative model is a formal mechanism for category formation.

Escudero’s L2LP (2005) incorporates distribution-based learning, and develops a formal model based on Stochastic Optimality Theory (Stochastic OT, Boersma, 1998), which focusses on the boundaries between speech sound categories. Like the SLM, Escudero’s L2LP model deals with new and similar L2 speech sound scenarios; however, rather than thinking in terms of whether an L2 category is new or similar to an L1 category, the L2LP focusses more on the relative proximity of the boundary between two L2 categories and two L1 categories. The L2LP also deals with the *subset* scenario, which occurs when the L1 has more categories than the L2 so that two or more L1 sounds occupy the same part of the acoustic space as one L2 sound. The following is a brief sketch of the three scenarios:

– *New scenario*: Returning to the stylised illustration, most instances of English /ɛ/ and English /ɪ/ are assimilated to Spanish /e/, the L2 learner will therefore initially perceive instances of the two English vowels as members of a single category. L2 learners are posited to have access to the same learning mechanisms as L1 learners, the L2-English learner will therefore be able to establish a new category boundary on the basis of the statistical distribution of the acoustic properties of the input. The learning of the new English /ɪ/–/ɛ/ boundary is therefore analogous to the learning of an L1 boundary, and the L2 learner must first posit that there are two L2 categories, and establish initial criteria for distinguishing the two, before they can begin to use lexical information. The L2LP posits that new categories will be easier to form on dimensions which were previously unused by the L1. For example, rather than splitting the Spanish /e/ category in the F1 dimension, L1-Spanish L2-English learners may develop a new English /ɛ/–/ɪ/ boundary on the duration dimension which is not used in Spanish (the example in Escudero, 2005, is a duration-based English /i/–/ɪ/ boundary splitting of the Spanish /i/ category).

– *Similar scenario*: The English /i/–/ɪ/ boundary is a little lower in terms of F1 than the Spanish /i/–/e/ boundary; hence, almost all instances of English /i/ are assimilated to Spanish /i/, and most instances of English /ɪ/ are assimilated to Spanish /e/. An L2 English learner will therefore be able to reuse this boundary and the Spanish /i/ and /e/ category labels. Almost all instances of the English vowel /i/ in the English word *sheep* /ʃip/$_{\text{Eng}}$ will be assimilated to Spanish /i/, the L2-English learner will therefore assign the label /ʃip/$_{\text{Sp}}$ to the ovine (the subscripts indicate whether the vowel symbols in the phonemic transcriptions correspond to English or Spanish category labels). Some instances of the English vowel /ɪ/ in the English word *ship* /ʃɪp/$_{\text{Eng}}$ will be assimilated to Spanish /i/, but since most instances will be assimilated to Spanish /e/ the L2 learner will assign the label /ʃep/$_{\text{Sp}}$ to the marine vessel. When the L2-English learner perceives the word /ʃip/$_{\text{Sp}}$, but context makes it clear that the object of discussion is a marine vessel rather than an ovine, then context provides a label indicating that they have misheard the word. Over time, increased exposure to English will provide more labelled input which will allow the L2-English learner to shift their boundary towards the optimal location for distinguishing English /i/ and /ɪ/.

– *Subset scenario*: If an L1-English listener is learning Spanish, then some instances of Spanish /e/ will be assimilated to English /ɛ/ and some to English /ɪ/. Because categories already exist, the L2-Spanish learner will be able to make use of lexical information, and once they realise that words to which they initially assign the labels /tʃɛka/$_{\text{Eng}}$ and /tʃɪka/$_{\text{Eng}}$ have the same referent, ‘checa’ *Czech* /tʃeka/$_{\text{Sp}}$, they can conflate the two L1-English vowels into a single L2-Spanish category. The scenario is complicated by the relative location of the English /i/–/i/ and Spanish /i/–/e/ boundaries resulting in some instances of Spanish ‘chica’ *girl* /tʃika/$_{\text{Sp}}$ also being perceived as /tʃɪka/$_{\text{Eng}}$, but since most instances will be perceived as /tʃika/$_{\text{Eng}}$ this aspect of the subset scenario is analogous to the similar scenario.

Both the SLM and the L2LP posit the L1 perception system as the initial state for the L2 system, but in contrast to the SLM’s which posits a single phonological space for L1 and L2 sounds, the L2LP posits separate perception grammars for L1 and L2 sounds. The L2LP therefore predicts that L2 learners can potentially become optimal perceivers of the L2, and remain optimal perceivers of their L1. The SLM, in contrast, predicts that the ultimate state achievable by a bilingual is an optimal perceiver of the entire set of L1 and L2 speech sounds (including L1–L2 diaphones and new L2 sounds), which will not be optimal for either the L2 or the L1. Neither assumption is essential to developing a distribution-based model of L2 learning, Figure 1.4 shows what the ultimate state achievable might be for a bilingual listener using distribution-based learning and the SLM assumption of a single phonological space, and the following discussion is a sketch of how the SLM might be interpreted in terms of distributions. An L1-Spanish L2-English learner will initially assimilate instances of English /ɪ/ and /ɛ/ to Spanish /e/, but with sufficient exposure to English, the L2-English learner will eventually detect the bimodal distribution of these two vowels and establish a boundary between them. Note, however, that to establish new boundaries, the L2 learner is assumed to be working only with the sum of the distributions of the input as was the case for L1 learners. In the single phonological space the sum of the distributions includes the Spanish /e/ as well as the English /ɪ/ and /ɛ/ categories. As predicted by the SLM, the resulting Spanish /e/ + English /ɪ/ diaphone – new English /ɛ/ boundary (Sp/e/+Eng/ɪ/ diaphone – new Eng/ɛ/ boundary) is therefore deflected away from where the English /ɪ/–/ɛ/ boundary would be for a monolingual English listener. Once the boundary has been established, the L2-English learner will be able to make use of lexical information and fine tune the location of the boundary on the basis of labelled input. If the establishment of the new boundary is not to be disastrous for L1 perception, then it may be supposed that vowels which are initially categorised as either the Sp/e/+Eng/ɪ/ diaphone or the new Eng/ɛ/ are subsequently conflated and labelled as Spanish /e/ when listening is Spanish mode. Because of the large overlap between the English /i/ and Spanish /i/ distributions, these two vowels are highly similar and will form a diaphone. The resulting distributions for the Sp/i/+Eng/i/ diaphone and the Sp/e/+Eng/ɪ/ diaphone, and lexical information providing labelled input, will result in a Sp/i/+Eng/i/ diaphone – Sp/e/+Eng/ɪ/ diaphone boundary which, as predicted by the SLM, is intermediate between the monolingual Spanish /i/–/e/ boundary and the monolingual English /i/–/ɪ/ boundary.

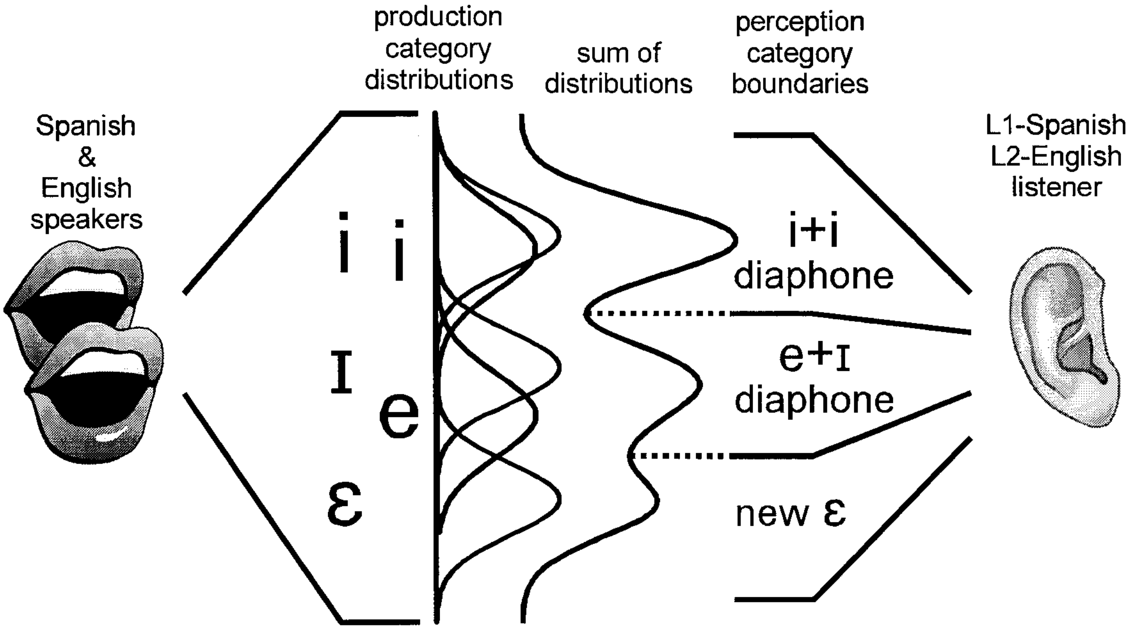

production category distributions
sum of distributions
perception category boundaries
Spanish & English speakers
L1-Spanish L2-English listener
i i
ɪ e
ɛ
i+i diaphone
e+ɪ diaphone
new ɛ

Figure 1.4 A stylised one-dimensional illustration of the ultimate achievable state of L2 speech perception learning in a common phonological space.

Vallabha & McClelland (2005) present a model which is intermediate between the possible extremes of a single phonological space and completely separate perception grammars. The model is a neural network with a single topographical map layer between the input and output layers. In contrast to the SLM and L2LP, the topographical map model is designed to be a neurologically plausible model of auditory cortex behaviour. When the network is trained on the L1, it develops Hebbian attractors which warp the topographical map simulating perceptual magnet effects (Grieser & Kuhl, 1989; Kuhl & Iverson, 1995). The network is able to learn L1 categories in an unsupervised manner solely on the basis of the distribution of the acoustic properties presented to the input layer, but labelled input can also be provided by simultaneously activating a node on the output layer which acts as a category label to identify the input pattern. Once the network has been trained on the L1, it has achieved the initial state for L2 learning, and is subsequently trained on L2 input. To model the extreme version of a single phonological space, the same output layer could be used for L1 and L2 input, and it may also be necessary to continue presenting L1 as well as L2 input. Such an approach could, however, lead to an excessive influence of L2 learning on L1 perception. To reduce but not eliminate this influence, the network has separate but parallel output layers for the L1 and the L2. Because the topological map has been warped in the process of learning the L1, the existing L1 system interferes with the learning of the L2 categories. Because learning the L2 gradually modifies the topographical map, and this topographical map is still a component of the L1 perception system, learning the L2 has a mediated effect on L1 perception.

1.3 L1-Spanish Speakers' Perception and Production of English /i/ and /ɪ/

1.3.1 Theories accounting for duration-based perception

L1-Spanish L2-English speakers often have difficulty perceiving and pronouncing the English /i/–/ɪ/ contrast.[6] Instances of both English vowels are typically assimilated to Spanish /i/ (Álvarez González, 1980, chap. 5; Escudero, 2005, §1.2.2; Flege, 1991; Imai, Flege, & Wayland, 2002), L1-Spanish L2-English listeners misidentify natural English /ɪ/ productions as English /i/ and vice versa (Møller Glasbrenner, 2005), and in production L1-Spanish L2-English speakers tend to substitute a Spanish-/i/-like vowel for both English vowels (Brennan & Brennan, 1981; Hammond, 1986; Flege, Bohn, & Jang, 1997; Morrison, 2005b). Using synthetic stimuli, Flege Bohn, & Jang (1997) found that, whereas L1-English listeners relied almost exclusively on spectral differences, L1-Spanish listeners had a greater

[6] This English contrast is also problematic for L2-English speakers with other L1s; for example, Mandarin & Korean (Flege, Bohn, & Jang, 1997; Ingram & Park, 1997; Wang & Munro, 2004), Japanese (Kewley-Port, Akahane-Yamada, & Aikawa, 1996; Morrison, 2002a, 2002b), Italian (Flege & MacKay, 2004), and Catalan (Cebrian, 2003).

preference for using vowel duration differences to perceptually distinguish English /i/ and /ɪ/. This result was replicated by Escudero & Boersma (2004) and Morrison (2002a).

The finding that L1-Spanish L2-English listeners use duration cues to distinguish English /i/ and /ɪ/ is noteworthy because Spanish does not have a vowel-duration contrast, and spectral properties are the primary perceptual cues for L1-English listeners. Although, all else being equal, L1-English speakers do tend to produce English /i/ as a longer vowel than English /ɪ/, the large overlap of the distributions of durations for instances of each category make duration a poor predictor of vowel identity (see Morrison 2005b). Several hypotheses have been advanced to account for L2-English learners' use of duration:

– Several authors (e.g., Flege, Bohn, & Jang, 1997; Morrison, 2005b; Wang & Munro, 1999) have noted that L2-English learners are typically taught that the difference between English /i/ and /ɪ/ is that /i/ is long and /ɪ/ is short. L2-English learners may therefore use duration to distinguish English /i/ and /ɪ/ because they are taught to do so. However, this raises the question of why teachers should teach that the difference between English /i/ and /ɪ/ is duration. At one level this could be a vocabulary problem: Phonetically naïve students and teachers are unlikely to have the vocabulary to describe spectral differences, but will understand the terms *long* and *short*, and therefore teachers may use these terms even if they know that duration is not the most important perceptual difference between the vowels. However, a theory is still needed to account for why students do not immediately notice that English /i/ and /ɪ/ differ in spectral properties as well as duration.[7]

– Bohn's (1995) Desensitisation Hypothesis proposed that Spanish listeners do not use spectral cues because, since their L1 does not expose them to phoneme-distinguishing spectral differences in the low-F1–high-F2 part of the vowel space, they are "linguistically desensitised" to spectral differences in this part of the vowel space. This would be compatible with the perceptual magnet effect (Grieser & Kuhl, 1989; Kuhl & Iverson, 1995). However, since Spanish does not expose L1-Spanish listeners to phoneme-distinguishing duration differences either, it is not clear why L1-Spanish listeners should not also be desensitised to duration differences, and desensitisation to spectral differences alone appears

[7] Another education-induced explanation arises if the teacher speaks English as a second language and produces a duration but no spectral difference between English /i/ and /ɪ/. In this case the students may accurately learn the L2-accented English to which they are exposed (see Flege & Eefting, 1987).

to be an arbitrary stipulation. In addition, no direct evidence has been presented in support of the hypothesis that L1-Spanish listeners are desensitised to spectral differences in this part of the vowel space; González Álvarez & Cervera Crespo (2001) failed to find spectral desensitisation in the vicinity of Spanish /i/.

– Escudero & Boersma (2004) proposed a theory related to the L2LP model: Since L1-Spanish listeners already have L1 vowel categories in the spectral dimension, this impedes their ability to form L2 categories on the basis of the distribution of the spectral properties in the L2 input. They proposed that L1-Spanish listeners have no categories in the duration dimension, and hence no impediment to learning English categories on the basis of the distribution of the duration properties in the L2 input. However, since all physical productions of a vowel must have some duration, all vowel categories must have some distribution of duration properties, and the claim that L1-Spanish listeners have no categories in the duration dimension is an arbitrary stipulation.

– Another idea is that vowel duration differences are somehow easier to perceive than vowel spectral differences, duration being viewed as a simple easily extractable one dimensional cue in contrast to multidimensional hard-to-extract formant cues. However, by definition, a just noticeable difference (JND) for duration is no more perceptible than a JND for formant values, and what would need to be considered is whether the difference between English /i/ and /ɪ/ productions is larger in terms of duration JNDs than spectral JNDs.

– Morrison (2005b) proposed a theory related to the multidimensional distribution of acoustic properties with no dimension-specific stipulations: Instances of English vowels which fall to the /i/ side of the Spanish /i/–/e/ boundary in a multidimensional acoustic space are assimilated to Spanish /i/, and instances which fall to the /e/ side are assimilated to Spanish /e/. This results in an upper limit on the distance in the higher-F1 direction from the Spanish /i/ prototype to an instance of an English vowel which is assimilated to Spanish /i/. There is no Spanish vowel category that has lower F1 and is longer than Spanish /i/, so there is no upper limit on the distance in the lower-F1–longer-duration direction from the Spanish /i/ prototype to an instance of an English vowel which is assimilated to Spanish /i/. Distance away from the English /i/ prototype could be measured in terms of JNDs, or in terms of the multidimensional probability density function (PDF) for Spanish /i/. Instances of English vowels which have lower F1 and longer duration (mostly English /i/) can be far out on the tail of the Spanish /i/ PDF but still assimilated to the Spanish /i/ category. A PAM-type prediction is that such stimuli will be noticed as deviant members of the Spanish /i/ category. In contrast instances of English vowels which have higher F1 and similar duration (mostly English /ɪ/) and are still assimilated to the Spanish /i/ cannot be as far out on the tail of the Spanish /i/ PDF, and are therefore less likely to be noticed as deviant members of the Spanish /i/ category. L1-Spanish listeners are therefore more likely to notice the duration difference and attempt to use duration to distinguish English /i/ and /ɪ/.

1.3.2 Hypothesised development of English /i/– /ɪ/ perception

L1-Spanish L2-English speakers who have lived in an English speaking environment for longer, have been found to perceive and produce English /i/ and /ɪ/ in a more L1-English-like manner (e.g., Flege, Bohn, & Jang, 1997; Møller Glasbrenner, 2005; Morrison, 2002a). Escudero (2000) proposed a hypothetical developmental sequence for English /i/ and /ɪ/ learning:

- Stage 0: L1-Spanish learners of English are at first unable to perceptually distinguish English /i/ and /ɪ/.
- Stage 1: Next they distinguish English /i/ and /ɪ/ using duration cues.
- Stage 2: Next they use a mixture of duration and spectral cues but still make more use of duration than L1-English listeners
- Stage 3: Finally they use spectral cues in an L1-English-like manner.

Morrison (2005b) found an additional pattern for English /i/ and /ɪ/ perception that did not fit any of the stages hypothesised by Escudero. This led him to hypothesise an additional stage between Escudero's Stages 0 and 1:

- Stage ½: L1-Spanish learners of English distinguish English /i/ and /ɪ/ via a category-goodness-difference assimilation to Spanish /i/ using both duration and spectral cues.

Stimuli which had low F1 and short duration were more Spanish-/i/ like and were identified as English /ɪ/, and stimuli which had higher F1 and longer duration were less Spanish-/i/ like and were identified as English /i/. The choice of English /ɪ/ to label stimuli which are more Spanish-/i/ like may have been related to orthography: In English orthography, English /ɪ/ is represented by the letter 'i', which in Spanish orthography represents Spanish /i/.

1.4 Aspects of the Present Study

The present study reexamines L1-Spanish speakers' acquisition of English /i/ and /ɪ/ via the collection and analysis of cross-sectional vowel perception and production data from L1-Spanish L2-English speakers who have lived for different periods of time in an Anglophone region of Canada. The study also directly examines development of L2-English /i/ and /ɪ/ perception and production via longitudinal case studies of the changes in L1-Spanish L2-English speakers' vowel perception and production during their first seven months of residence in Canada (substantial changes can occur over this time period, Morrison, 2002a, see also Flege & Liu, 2001).

Although L1-Spanish listeners have primarily been reported to identify instances of English /i/ and /ɪ/ as Spanish /i/, the situation is much more complicated. As illustrated in Figure 1.5, L1-Spanish listeners also identify instances of English /ɪ/ as Spanish /e/ (Álvarez González, 1980, chap.5; Escudero, 2005, §1.2.2; Flege, 1991; Højen & Flege, in press; Imai, Flege, & Wayland, 2002) and as English /ɛ/ (Møller Glasbrenner, 2005).[8] Instances of English /ɛ/ are also typically identified as Spanish /e/ (Álvarez González, 1980, chap. 5; Flege, 1991; Imai, Flege, & Wayland, 2002). Instances of Spanish /e/ are in turn typically identified as English /e/ by L1-English listeners (Morrison, 2003), and L1-Spanish listeners primarily identify instances of English /e/ as Spanish /e/ (Højen & Flege, in press; Imai, Flege, & Wayland, 2002).[9] Impressionistically, one would also expect instances of Spanish /ei/ to be perceived as English /e/ by L1-English listeners.[10] The present study therefore

[8] The rate of assimilation of English /ɪ/ to Spanish /e/ versus Spanish /i/ is dependent on the English dialect, e.g., compared to English speakers from the south of England, Scottish-English speakers produce English /ɪ/ with greater spectral and less duration separation from English /i/. English /ɪ/ has higher F1 resulting in a greater rate of assimilation of Scottish-English /ɪ/ to Spanish /e/ (Escudero, 2005, §1.2.2; see also Escudero & Boersma, 2004).

[9] On the basis of the assimilation results for Southeastern US English in Højen & Flege (in press), L1-Spanish listeners would be expected to have more difficulty distinguishing English /ɪ/ and /e/ than distinguishing English /i/ and /ɪ/.

[10] Additional studies of L1-Spanish speakers' perception and production of English vowels include: Blankenship (1991), Bradlow (1995), Contreras Oller (1997), Flege, Munro, & Fox (1994), Fox, Flege, & Munro (1995), García Bayonas (2004), Højen (2005).

investigates all the non-low front vowels of English and Spanish: English /i/, /ɪ/, /e/, /ɛ/, and Spanish /i/, /e/, /ei/.[11]

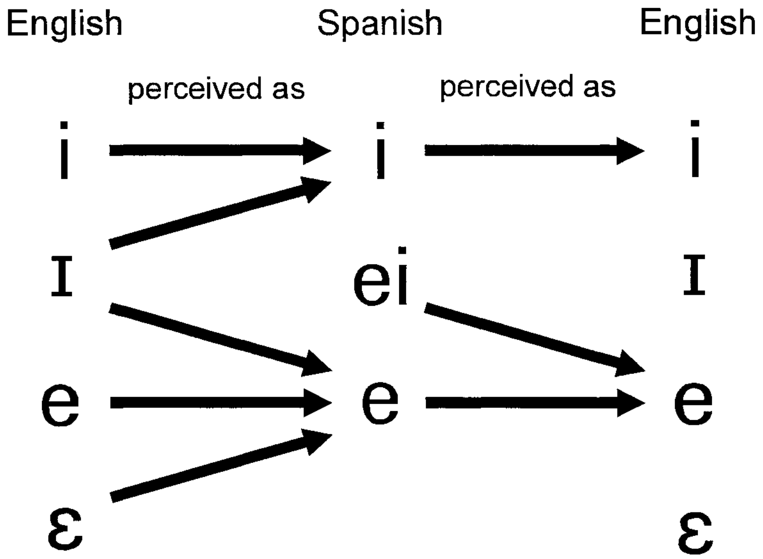


Figure 1.5 Previously observed assimilations of instances of English vowel categories to Spanish vowel categories, and instances of Spanish vowel categories to English vowel categories. Arrows are not proportionally weighted.

Most earlier L2 vowel studies considered only the duration and steady-state spectral properties of the vowels under investigation; however, the Canadian English phonetic diphthong /e/, and nominal monophthongs /ɪ/ and /ɛ/, have substantial vowel inherent spectral change. The proposed study therefore investigates the effect of VISC on Spanish speakers' vowel perception, testing listeners' perception of a synthetic speech continuum with three varying dimensions: initial spectral properties, VISC, and duration. This is one of the first studies to systematically investigate L2 VISC perception. Møller Glasbrenner (2005) investigated the effect on L1-Spanish listeners' perception of removing VISC or duration from STRAIGHT resynthesised English vowel productions (Kawahara, Matsuda-Katsue, & Cheveigné, 1998). Removing duration significantly reduced the correct-

[11] Although English /e/ and /ɛ/ are primarily identified as Spanish /e/, they are also sometimes identified as Spanish /a/ (Flege, 1991; Imai, Flege, & Wayland, 2002). The latter Spanish vowel is not included in the present study. The synthetic vowels in the present study have relatively high F2 values and are not expected to result in /a/ perception.

classification rate for English /ɪ/ for low-proficiency L2-English listeners, but not high-proficiency L2-English listeners or L1-English listeners, indicating that the low-proficiency group had a greater reliance on duration when identifying this vowel. All three groups had a significantly reduced correct-classification rate for English /e/ and /æ/ when VISC was removed, indicating that all three groups made use of VISC when identifying these vowels, the high-proficiency L2-English listeners also had a significantly reduced correct-classification rate for English /ɑ/.[12]

Because it is based on multidimensional-category-goodness-difference assimilation to Spanish /i/, Morrison's (2005b) extension of Escudero's (2000) hypothesised developmental stages for English /i/– /ɪ/ perception learning by L1-Spanish listeners, can be easily extended to additional dimensions such as VISC. Since Spanish has no converging diphthong (rising F1 and falling F2) resembling English /ɪ/, stimuli with the VISC pattern of English /ɪ/ are more likely to be noticed as deviant members of the Spanish category to which they are assimilated. In Morrison (2005b), L1-Spanish listeners assigned to the category-goodness-assimilation stage of English learning, labelled resynthesised stimuli that had Spanish-/i/-like properties (stimuli with low F1 and short duration) as English /ɪ/ and stimuli which were less Spanish-/i/ like (stimuli that had higher F1 or longer duration or both) were labelled as English /i/. All else being equal, synthetic stimuli in the present study which are assimilated to Spanish /i/ are therefore hypothesised to be more likely to be labelled as English /i/ if they have converging VISC, an English-/ɪ/-like but not a Spanish-/i/-like acoustic property.

Most L2 perception research has focussed on cases such as L1-Spanish speakers learning English where the number of phoneme contrasts in the L2 is greater than that in the L1 (Boersma & Escudero, 2004; Escudero, 2005, chap. 6; Escudero & Boersma, 2002; and Morrison, 2003, being exceptions). However, L1-English learners of Spanish typically also have serious pronunciation problems. The present study therefore also investigates the perception and production of vowels by L1-English L2-Spanish speakers, including participants who have studied Spanish only in the classroom and participants who have lived

[12] Other L2 studies with VISC components include Cebrian (2003), Flege, Schirru, & MacKay (2002), Munro (1993).

in Spanish-speaking countries.

In order to understand the results of L2 perception and production research, one must first have a good understanding of L1 perception and production (Escudero, 2005, §3.2–3.3; Højen & Flege, in press; Morrison, 2006; Rochet, 1995). The present study therefore collects data on the perception and production of English and Spanish vowels by monolingual English and Spanish participants as well as by bilingual participants.

Discriminant analysis and logistic regression, which are established statistical techniques which have been successfully used for modelling L1 production and perception data. Discriminant analysis and logistic regression will be used to build models of L1-Spanish and L1-English production and perception. Models trained on one L1 data from one language will be used to classify data from the other language and thus make predictions regarding the assimilation of instances of L2 vowels to L1 vowel categories. The same types of models will be fitted to L2 production and perception data in order to quantify individual L2 learners' perception and production patterns. These static L2 models are akin to snapshots which will be compared in search of patterns of change across apparent (cross-sectional) or real (longitudinal) time. The aim is to establish a thorough description of the patterns of L2 learning in the data.

2. Methodology

Experiment presentation, stimulus synthesis, acoustic analysis, and statistical analysis were performed using software programmed in Matlab (MathWorks, 2001, 2004) by the author, incorporating pre-existing components programmed by Terrance M. Nearey (spectrogram, formant tracker, logistic regression analysis), and incorporating a precompiled version of the Klatt synthesiser (Klatt & Whalen, 1985). Repeated-measures ANOVAs were performed using SPSS (SPSS, 2002).

2.1 Participants

Potential participants completed a brief language background questionnaire. One section of the questionnaire asked potential participants to list the languages they spoke and indicate their proficiency in those languages, the choices being: *a-little*, *some*, *well*, and *near-native* (or in Spanish *un-poco*, *algo*, *bien*, *casi-nativo*). Potential participants who responded *well* or *near-native* for any languages other than English or Spanish were not included in the study. Potential participants who had been exposed to other languages in early childhood were also excluded, as were potential participants who reported hearing or speech impediments.

Since there can be substantial differences in vowel realisations and vowel inventory across English dialects, the participants' English dialect was controlled. All L1-English participants were speakers of General Canadian English. L2-English dialect was not as tightly controlled, but most L1-Spanish L2-English participants had not lived in any English-speaking country other than Canada. Given the pool of potential volunteers, it was not possible to control for Spanish dialect and obtain a reasonably large number of volunteers. Control over Spanish dialect was not as important as control over the English dialect since, compared to English, Spanish vowels have relatively little cross-dialectal variation, especially for stressed vowels and especially for educated urban speech.[1] L1-Spanish participants came from various Spanish-speaking countries, and L1-English L2-Spanish participants had been exposed to various Spanish dialects.

Table 2.1 Background information for monolingual Spanish participants.

ID	Gender	Place of Origin	Age	Other Languages*
ms030	M	Basque Country	49	French
ms031	M	Basque Country	50	–
ms032	M	Basque Country	43	English, French, Basque
ms033	M	Madrid	34	–
ms034	M	Basque Country	34	Basque, English
ms035	M	Basque Country	34	English
ms036	F	Basque Country	44	French, Basque
ms037	M	Basque Country	45	–
ms038	M	Basque Country	40	–
ms039	F	Basque Country	36	English
ms040	F	Basque Country	25	–
ms041	F	Leén	48	–
ms042	F	Basque Country	53	–
ms043	F	Basque Country	49	–
ms045	F	Basque Country	53	–
ms046	F	Colombia	18	–
ms047	F	Burgos	28	–
ms048	F	Navarre	44	English, Basque, French

* Participants reported speaking *a-little* or *some*.

Eighteen (18) monolingual Spanish participants were recruited from members and friends of members of a paragliding club based in Vitoria-Gasteiz in the Autonomous Region of the Basque Country, Spain. This is a traditionally Spanish speaking part of the Basque Country, and all of the participants were educated in Spanish, and were functionally

[1] Vowel variation in different Spanish dialects is discussed in Alvar (1996a, 1996b), and Vaquero de Ramírez (1996). In an acoustic study, Godínez (1978) suggested that there are some differences between Peninsular, Mexican, and Argentinian Spanish monophthong production; however, the number of speakers per group was small and no statistical tests were conducted. The similarity of Peninsular- and Mexican-Spanish speakers' vowels will be directly assessed as part of the present study.

monolingual in Spanish. Participants in the monolingual Spanish group tended to be older than those in the other groups since most residents of Vitoria-Gasteiz in their teens and twenties had been educated in Basque. Additional background information about these participants is provided in Table 2.1.

Table 2.2 Background information for monolingual English participants.

ID	Gender	Place of Origin	Age	Control Condition*
me095†	F	Alberta	21	–
me096	F	Alberta	24	2
me097	M	Alberta	54	4
me098	F	Alberta	20	3
me099	F	Saskatchewan	19	1
me100	F	Alberta	19	2
me101	F	Alberta	20	4
me102†	F	Alberta	20	–
me103	F	Alberta	18	1
me104‡	M	Alberta	21	–
me105	F	Alberta	20	4
me106	M	Alberta	27	3
me107	F	Alberta	19	1
me108	F	Alberta	18	2
me109	M	Alberta	23	4
me110†	F	Alberta	20	–
me111	F	Alberta	19	1
me112	M	Alberta	18	2
me113	M	Alberta	21	4
me115	M	Alberta	21	1
me116†	M	BC/Alberta	22	–
me117	M	Alberta	28	3
me118	F	Alberta	19	3

* In the second experiment session, participants completed one of four control conditions for the perception experiment: 1, Unaltered replication. 2, Spanish carrier sentence. 3, Isolated word with no carrier sentence. 4, Alternative voice for English carrier sentence and synthetic stimuli.

† Participant had difficulty with first production experiment, excluded from subsequent experiments.

‡ Participant was unable to return for control experiments.

Twenty-three (23) monolingual English participants were recruited from students at the University of Alberta in Edmonton, Alberta, Canada. They where all speakers of General Canadian English, had grown up in Western or Central Canada, had English-speaking parents, and had been educated in English. None reported speaking any other language, even at the *a-little* or *some* level. Additional background information about these participants is provided in Table 2.2.

Forty-one (41) L1-Spanish L2-English participants were recruited at the University of Alberta and surrounding community. They constituted a cross-sectional group in which individuals had varying degrees of exposure to English. Length of residence (LOR) in Canada is reported in Table 2.3 as the primary indicator of exposure to English.[2] Additional background information about these participants is also provided in Table 2.3. Four (4) of the L1-Spanish L2-English speakers also participated in longitudinal case studies.

Table 2.3 Background information for L1-Spanish L2-English participants, including length of residence (LOR) in Canada.

ID	Gender	Place of Origin	Age	LOR in Canada	Other Languages*
bs001	F	Mexico	25	3 months	–
bs002	F	Bolivia	30	2 years	Portuguese
bs003	M	Spain	29	1 year	–
bs016	M	Colombia	30	5 years	–
bs017	F	Chile	30	24 years	–
bs019	F	Mexico	36	2 years	–
bs023	M	Panama	21	6 months	–
bs028	F	Mexico	43	1 year (+ 6 months Ohio, 1 year Manchester, UK)	–
bs049	M	Venezuela	28	4 years	–
bs051	F	Mexico	23	6 months	–

continued on next page...

[2] LOR may be a relatively poor predictor of actual exposure to the L2 since different individuals may spend more or less time communicating in the L2 on a daily basis. Flege & Liu (2001) found greater effects for LOR on L2 perception for participants whose occupations required extensive oral-aural interaction in the L2, i.e., students, than for participants whose occupations required relatively little oral-aural interaction in the L2, e.g., biomedical laboratory researchers. The vast majority of the L2 participants in the present study were students, and most of the remainder also had occupations which required extensive oral-aural interaction in the L2.

Table 2.3 ...continued from previous page.

ID	Gender	Place of Origin	Age	LOR in Canada	Other Languages*
bs052	F	El Salvador	25	6 months	–
bs056	M	Colombia	26	1 year (+ 1 year Wisconsin, 6 months New Mexico)	–
bs057	F	Peru	48	15 years	–
bs058	F	Spain	50	4 years (but for first 3 worked and socialised only in Spanish)	–
bs059	M	Colombia	30	5 years	French
bs061	F	Colombia	26	3 years	French
bs062	F	Mexico	26	2 weeks (+ 1.5 years in Texas)	–
bs063	F	Mexico	21	2 weeks	French
bs064	F	Mexico	28	4 months	–
bs065	F	Mexico	19	1 year	Hebrew
bs067	F	Argentina	27	2 weeks	–
bs068	F	Mexico	21	3 years	Japanese
bs069†	F	Peru	29	8 months (+ 4 months Boston) limited contact with English speakers	–
bs070‡	F	Mexico	27	3 years	–
bs071	M	Mexico	21	2 weeks	–
bs072	M	Mexico	25	1 month (+ 1 year Chicago)	Japanese, French
bs073	M	Mexico	21	1 month	French
bs074	M	Mexico	22	2 years	French
bs075†	M	Venezuela	26	1 month (+ 3 months in Boston)	–
bs076	F	Mexico	34	1.5 years	–
bs077	M	Peru	20	4 years (+ 6 months in Scotland, 6 months in Florida)	French
bs078	M	Mexico	19	1 year	French
bs081	M	Peru	22	1 month	–
bs082	M	Mexico	20	2 weeks (+ 1 month in Texas)	French, Italian
bs083†	M	Argentina	39	1 month	German
bs086	F	Puerto Rico	28	2 years (+ 2 years in Hawai'i)	–
bs087†	M	Mexico	30	1 month	–
bs088	F	Mexico	24	6 months	–
bs091	F	Mexico	20	3 months	–
bs114	M	Mexico	18	3 months	–

* Participants reported speaking *a-little* or *some*. † Participant in longitudinal case study.

‡ Participant was unable to return for English version of experiments.

Table 2.4 Background information for L1-English L2-Spanish participants, including highest level of Spanish language instruction taken at university (Level), and length of residence (LOR) in a Spanish speaking country.

ID	Gender	Place of Origin	Age	Level	LOR in Spanish Speaking Country	Other Languages*
be004	M	Saskatchewan	23	200	6 months Spain	–
be005	M	Alberta	22	–	2 years Argentina	–
be006	F	Alberta	19	–	1 year Argentina, 1 year Bolivia	–
be009	F	Alberta	21	300	–	–
be010	M	Saskatchewan	23	400	–	–
be011	F	Alberta	23	300	1.5 years Spanish speaking community in United States	–
be012	F	Alberta	25	400	–	–
be014	F	Saskatchewan	22	400	–	French
be015†	M	Alberta	24	400	–	–
be020	F	Alberta	20	–	1 year Mexico	French
be021	F	Alberta	20	–	1 year Ecuador	–
be022	F	Alberta	20	–	2 years Panama	–
be024	F	Alberta	20	200	–	French
be027	F	Alberta	20	300	–	–
be029	F	Alberta	20	300	–	–
be053	F	Alberta	18	200	–	–
be054	M	Alberta	23	–	2 years Guatemala	–
be055	F	Alberta	45	200	–	–
be079	M	Alberta	22	200	4 months Peru, Bolivia, Argentina, Uruguay, Costa Rica	–
be080	F	Alberta	22	300	–	Italian
be084	F	Alberta	21	400	–	–
be085	F	Alberta	27	200	–	–
be089	M	Ontario	22	–	2 years Mexico	French, Japanese
be090	F	Alberta	29	400	3 years Spain	–
be093	F	Alberta	19	300	–	French
be094	M	Alberta	21	200	1 year Mexico	–
be119	M	BC/Alberta	23	–	2 years Spanish speaking community in United States	French

* Participants reported speaking *a-little* or *some*.

† Participant had speech impediment, production data excluded.

Twenty-seven (27) L1-English L2-Spanish participants were recruited at the University of Alberta and surrounding community. There were two principle subgroups: Fourteen (14) participants had received at least two years of classroom Spanish instruction but had never lived in a Spanish-speaking country, and the other thirteen (13) participants had lived in a Spanish-speaking country for at least four months. Additional background information about these participants is provided in Table 2.4.

2.2 Stimuli & Prompts

2.2.1 Production Prompts

Production prompts consisted of written sentences. In English the prompts were:

> *The next word is BEEPA.*
> *The next word is BIPPA.*
> *The next word is BAYPA.*
> *The next word is BEPPA.*

Some participants may have been familiar with phonetic symbols, so to reduce the risk of them confusing the orthographic prompts with phonetic-symbols, the target words were written in capital letters and in a handwritten-style font. The orthographic *BEEPA*, *BIPPA*, *BAYPA*, and *BEPPA* were intended to elicit /bipə/, /bɪpə/, /bepə/, and /bɛpə/ respectively.

In Spanish the prompts were:

> *La próxima palabra es BIPA.*
> *La próxima palabra es BEIPA.*
> *La próxima palabra es BEPA.*

The Spanish carrier sentence had the same meaning as the English carrier sentence. Orthographic *BIPA*, *BEIPA*, and *BEPA* were intended to elicit /bipa/, /beipa/, and /bepa/; the Spanish orthography has a one-to-one relation with the phonemes.

In both languages the target words are nonsense words. The target word was preceded by an alveolar fricative /s/ in both languages (in the synthetic stimuli below in both language the fricative was partially voiced). Prevocalically the /b/ phoneme of English is typically pronounced as a voiceless (zero or short lag VOT) or a voiced bilabial plosive, and prevocalically the /b/ phoneme of Spanish is typically pronounced as a voiced bilabial plosive. Intervocalically the /p/ phoneme of both languages is typically pronounced as a voiceless bilabial plosive. Utterance finally the English /ə/ and Spanish /a/ are relatively similar: they are both pronounced as mid to mid-low relatively central vowels.

2.2.2 Perception Stimuli

A synthetic-stimulus continuum was designed to cover an acoustic range that would include the English vowels /i, ɪ, e, ɛ/ and the Spanish vowels /i, e, ei/. They were based on the natural productions of a male bilingual speaker who was recorded reading the production stimuli out loud.[3] Measurements were made of the duration of the vowels and /p/ closures, and F0, F1, F2 and F3 at 0%, 20%, 80%, and 100% of the duration of the vowels. The vowel duration range was approximately 20–140 ms. The F1 production range at 20% of the vowel duration was approximately 250–580 Hz and the F2 range 1650–2050 Hz.

Stimuli were synthesised at a sampling frequency of 10 kHz using a version of the Klatt synthesiser (Klatt & Whalen, 1985; see also Klatt, 1980; Klatt & Klatt, 1990). A synthetic /bip̚/ was created, based closely on the measured properties of a Spanish /bipa/ production. The synthetic /bip̚/ was spliced into the Spanish carrier sentence (downsampled to 10 kHz and including the final /pa/ from the release burst onwards), and trial and error adjustments were made to the parameter settings until the voice quality of the synthetic /bip̚/ sounded as close as possible to the natural carrier sentence, a full set of non-varying parameter values is provided in Table 2.5a. The resulting stimulus sentences were very natural sounding, and the synthetic portion did not stand out as being synthetic or as being

[3] The speaker was the author, an L1-English L2-Spanish speaker who was 35 years old when the recordings were made. The speaker had learnt Spanish in Spain. In pilot studies, L1-Spanish listeners indicated that they did not perceive a foreign accent in the Spanish carrier sentence produced by the speaker. Although he had spent most of his adult life in Canada the speaker was originally from the UK. A control condition was therefore included using the voice of a monolingual-English speaker from Edmonton.

Table 2.5 Non-varying parameters specifying voice quality for Klatt Synthesiser. See software documentation for explanation of parameters (Klatt & Whalen, 1985; see also Klatt, 1980; Klatt & Klatt, 1990).

a. Voice quality used in English and Spanish perception experiments.

SYM	V/C	MIN	VAL	MAX	SYM	V/C	MIN	VAL	MAX	SYM	V/C	MIN	VAL	MAX
sr	c	5000	10000	20000	nf	c	2	5	6	ss	c	1	2	2
du	c	30	210	1000	rs	c	1	1	99999	os	c	0	0	20
ui	c	1	5	20	sk	v	0	0	80	fp	v	200	250	500
f0	V	0	1300	5000	no	v	10	60	65	fz	v	200	250	700
f1	V	150	280	2000	b1	v	40	90	500	p1	v	30	80	1000
f2	V	150	1240	3500	b2	v	40	110	500	p2	v	40	200	1000
f3	V	150	2400	5500	b3	v	40	170	500	p3	v	60	350	1000
f4	v	1500	3300	6500	b4	v	100	400	500	p4	v	100	500	1000
f5	v	2500	3700	7500	b5	v	150	500	700	p5	v	100	600	1500
f6	v	3000	4990	9500	b6	v	200	800	2000	p6	v	100	800	4000
a1	v	0	66	80	ab	v	0	0	80	av	V	0	66	80
a2	v	0	66	80	af	V	0	0	80	bp	v	50	100	500
a3	v	0	66	80	ah	v	0	0	80	bz	v	50	100	500
a4	v	0	66	80	an	v	0	0	80	tl	v	0	24	24
a5	v	0	66	80	ap	v	0	0	80	g0	v	0	60	80
a6	v	0	0	80	at	v	0	45	80	sc	c	0	1	1

b. Voice quality used in English alternative-voice control experiment.

SYM	V/C	MIN	VAL	MAX	SYM	V/C	MIN	VAL	MAX	SYM	V/C	MIN	VAL	MAX
sr	c	5000	10000	20000	nf	c	2	5	6	ss	c	1	2	2
du	c	30	210	1000	rs	c	1	1	99999	os	c	0	0	20
ui	c	1	5	20	sk	v	0	20	80	fp	v	200	250	500
f0	V	0	1300	5000	no	v	10	60	65	fz	v	200	250	700
f1	V	150	280	2000	b1	v	40	90	500	p1	v	30	80	1000
f2	V	150	1240	3500	b2	v	40	110	500	p2	v	40	200	1000
f3	V	150	2400	5500	b3	v	40	170	500	p3	v	60	350	1000
f4	v	1500	3300	6500	b4	v	100	400	500	p4	v	100	500	1000
f5	v	2500	3700	7500	b5	v	150	500	700	p5	v	100	600	1500
f6	v	3000	4990	9500	b6	v	200	800	2000	p6	v	100	800	4000
a1	v	0	66	80	ab	v	0	0	80	av	V	0	66	80
a2	v	0	66	80	af	V	0	0	80	bp	v	50	100	500
a3	v	0	66	80	ah	v	0	0	80	bz	v	50	100	500
a4	v	0	66	80	an	v	0	0	80	tl	v	0	20	24
a5	v	0	66	80	ap	v	0	0	80	g0	v	0	60	80
a6	v	0	0	80	at	v	0	40	80	sc	c	0	1	1

produced by a different voice.

A minimalist approach to synthesis was adopted, as in Andruski & Nearey (1992), with linear interpolations between inflection points, see Figure 2.1. The release burst of the /b/ was synthesised by setting the fricative amplitude (*af*) at 75 dB at time 5 ms, 40 dB at time 10 ms, and zero elsewhere (parameter values were specified at 5 ms intervals). The nominal onset of the vowel was at time 10 ms. In order to synthesise bilabial onset and offset transitions, inflection points were established 15 ms after the vowel onset and 10 ms before the vowel offset. At the first and second inflection points, F1, F2, and F3 were set to their initial and final target values. At onset, F1, F2, and F3 were set to 80% (in Hertz) of their initial target values, and at offset they were set to 75% of their final target values. F3 at the initial and final targets were calculated according to a formula based on a linear regression of F1 and F2 Hertz values onto F3 Hertz values taken from measurements of the speaker's productions: $F3 = 4235 - 2.427 \times F1 - 0.272 \times F2$ (a similar procedure was adopted in Nearey, 1989). F0 was set to 129 Hz at vowel onset and at the first inflection point,114 Hz at the second inflection point, and 110 Hz at vowel offset. The fundamental and formant frequencies were converted to the natural logarithm of their Hertz values, and values between the onset, the two inflection points, and offset were linearly interpolated in log-Hertz values.[4] The linear interpolations were extrapolated beyond the vowel onset and offset. The voicing amplitude (*av*) was linearly interpolated (in decibels) between 78 dB at vowel onset, 80 dB at the first inflection point, 77 dB at the second inflection point, and 74 dB at vowel offset. To produce appropriate voicing levels during the /b/, the voicing amplitude was set to 60 dB at time 0, and 70 dB at time 5 ms. At 5, 10, 15, and 20 ms after the vowel offset the voicing amplitude was ramped down to 55, 20, 5, and 0 dB resulting in a dampened pseudo sine wave, simulating diminishing glottal pulsing during the beginning of the consonant closure. Once the stimuli had been synthesised, their actual amplitude levels were adjusted so that their apparent loudness was appropriate for the carrier sentence (this involved normalising each stimulus's amplitude by multiplying by a factor adjusted for the measured mean absolute amplitude of the portion of the vowel between the initial and final

[4] Log Hertz was chosen as a scale having a more linear relationship to human frequency perception than raw Hertz. Log Hertz was also used in Nearey & Assmann (1986), Gottfried, Miller, & Meyer (1993), and others for calculation of initial and final targets, slope, and direction.

targets). The duration of the /p/ closure was 90 ms, which was silent apart from the aforementioned synthetic glottal pulsing.

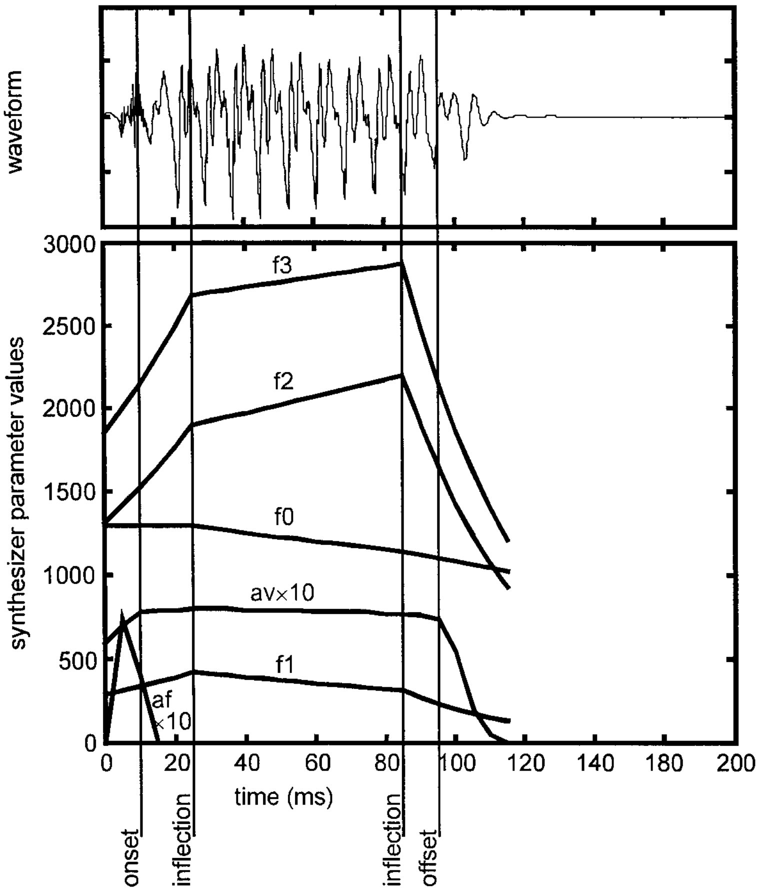


Figure 2.1 Sample variable parameter values used in the Klatt synthesiser. *f1*,*f2*,*f3* are first, second, and third formants with values in Hertz, *f0* is the fundamental frequency with values in tenths of Hertz, *av* and *af* are voicing and fricative amplitude with values in decibels (plotted at ten times the decibel values).

A large stimulus space (1464 stimuli) was initially synthesised, factorially combining duration, F1, and F2 values covering the range of values obtained when measuring the model speaker's productions. This included the magnitude of F1 and F2 formant movement. Direction of formant movement was constrained to reflect patterns in production: if F1 increased F2 decreased (converging VISC), and if F1 decreased F2 increased (diverging VISC). Pilot studies were conducted with L1-English and L1-Spanish listeners using *method of adjustment* (applied in a manner similar to that of Johnson, Flemming, & Wright, 1993) in order to discover a subset of the stimulus space in which listeners heard reasonably good examples of each of the target vowels in their native language. Eventually, the stimulus space was reduced to a set of 90 stimuli. Initial targets were synthesised as ten points equally spaced along a straight line in the F1–F2-Hertz space, see Table 2.6 for a full set of initial target values. Each initial-target point was combined with the three final-target values given in Table 2.7, and with three durations of 80, 95, and 110 ms for the vowel (55, 70, and 85 ms for the portion of the vowel between the initial and final targets).

Table 2.6 Initial-target values of synthetic stimuli.

Initial Target	
F1 (Hz)	F2 (Hz)
283	2090
316	2050
349	2010
382	1970
415	1930
448	1890
481	1850
514	1810
547	1770
580	1730

Table 2.7 Final-target values relative to initial-target values of synthetic stimuli.

Formant Movement (VISC)	
F1 (Hz)	F2 (Hz)
-99	120
0	0
99	-120

During the perception experiments, the synthetic stimuli were spliced into the English and Spanish natural-speech carrier sentences *The next word is* ____ , and *La próxima palabra es* ___ . In both languages, the synthetic /bVp˺/ stimuli were followed by a natural Spanish /pa/ (the Spanish /pa/ was not noticeably non-English like in the English context). A general acoustic difference between the speaker's Spanish and English pronunciation was that his Spanish had a steeper negative spectral tilt than his English. A filter was applied to change the spectral tilt of the English carrier sentence to match the Spanish-voice-quality-matched synthetic stimuli. The amplitude of the English carrier sentence was also adjusted to match the apparent loudness of the Spanish carrier and synthetic stimuli.

In the electronic (pdf) version of this document, Figure 2.2 includes embedded sound files of the stimuli.

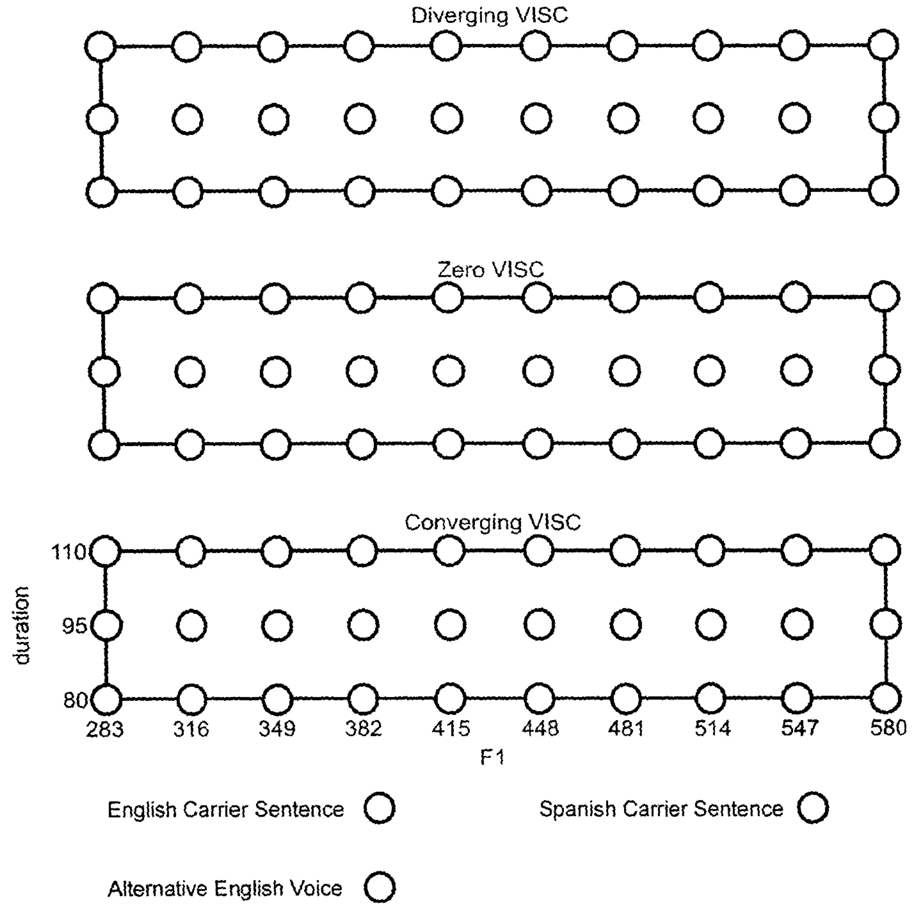


Figure 2.2 Synthetic stimulus properties. Click on a dot to hear the stimulus word with the corresponding properties.

2.3 Procedures

Monolingual-Spanish speakers were tested once in Spanish. Members of the cross-sectional bilingual groups were tested twice, once in their L1, and once in their L2. The L1 test was conducted one day and the L2 test on another day. Most Monolingual-English speakers were tested twice, on the first occasion they completed the standard English experiments and on the second occasion they completed one of four control conditions, details of which are described below. The L1-Spanish L2-English participants taking part in the longitudinal case studies were tested at four points in time. They were tested in Spanish and English when they had spent less than one month in Canada, in English when they had spent approximately three and five months in Canada, and in English and Spanish when they had spent approximately seven months in Canada.

Participants were tested one at a time in a sound booth. Monolingual Spanish participants were tested in the Phonetics Laboratory at the University of the Basque Country, and the remaining participants in the Centre for Comparative Psycholinguistics at the University of Alberta. Recording and playback of stimuli were to and from computer via a Sennheiser HMD 280 PRO headphone-microphone set and a Roland ED UA-30 USB Audio Interface, with a Rolls MP13 Mini-Mic Preamp for recording. The sampling frequency for recordings was 44.1 kHz. Playback volume was set at a comfortable level. The headphones attenuated outside noise, so, in order to prevent the Lombard effect, the signal from the audio interface was fed back into the headphones during recording.

All communication between the researcher and the participant, all the instructions, and the stimulus sentences, were in the language being tested. This was designed to prime the participants to perceive the stimuli in the appropriate language (see Bohn & Flege, 1993; Escudero & Boersma, 2002).

2.3.1 Production experiment

Participants saw instructions for the production experiment written on a computer screen and heard them read out loud. Participants saw the production prompts written on the screen and practised reading them out loud. In the English version of the experiment, participants were first introduced to real English words containing the same vowel sounds, including representations of the vowel sounds using the same spellings:

sleep	keep	meet	beeper	BEEPA
bit	sit	pick	zipper	BIPPA
say	play	day	paper	BAYPA
pet	neck	get	pepper	BEPPA

The researcher monitored the participants' pronunciation to ensure that they had understood the phonemes represented by the orthographic prompts. During this practice period, the researcher also adjusted the microphone position and recording level. The written prompts were then presented ten times in ten blocks, each sentence occurred once in each block with the order of presentation randomised in each block, and the first sentence of a block was never the same as the last sentence of the preceding block. Randomisations were generated on the fly so that each participant received different randomisations. During the experiment proper, the computer program played a beep, started recording sound, then displayed a stimulus sentence in the middle of the screen. The participant read the sentence out loud, then the researcher pressed a button to stop the recording. Periods of silence were automatically stripped from the beginning and end of the recording,[5] and a raw waveform was displayed for the researcher. The researcher then had the options of listening to the recording, accepting it, or rejecting it. Recordings were rejected if the participant misspoke or did not read the sentence fluently, if the recording included extraneous noise, or if there were recording problems such as clipping. The program presented a new stimulus 500 ms after the researcher accepted or rejected a recording. If a recording was rejected, the program added the corresponding stimulus to the next stimulus block and randomised the block so that the repeated stimulus was not adjacent to another instance of the same stimulus. If stimuli were rejected during the last stimulus block, an additional block was added and the rejected stimuli randomised within the additional block. The program stopped the production experiment once it had recorded at least ten responses to each stimulus which had been accepted by the researcher.

[5] The recording was truncated 20 ms before and after the first and last peak which was greater than 5% of the maximum peak amplitude of the whole recording.

In the first experiment session the monolingual English speakers completed the English production experiment described above, and in the second session they completed a control condition. In the control condition, monolingual English participants were presented with, and asked to read aloud, words in isolation (i.e., without the carrier sentence). These were the same CVCV words *BEEPA*, *BIPPA*, *BAYPA*, and *BEPPA* described above, plus the CVC words *BEEP*, *BIP*, *BAPE*, and *BEP* exemplifying the same set of vowel phonemes /i/, /ɪ/, /e/, and /ɛ/. This condition was included to give an indication of the effect of the carrier sentence on vowel production compared to vowel production in isolated words, and to allow for comparison with other isolated-word data sets.

2.3.2 Perception experiment

Following the production experiment, participants saw instructions for the perception experiment written on the computer screen and heard them read out loud. They also completed a practice version of the experiment in which they heard a small number of stimuli repeated in two blocks. Four stimuli were included in the English experiment, and three in the Spanish experiment. Each of the stimuli selected was expected to be a good example of one of the response categories. During the experiment, participants heard a stimulus sentence, then saw a number of rectangles on the screen, four rectangles in the English experiment, and three in the Spanish experiment, see Figure 2.3. In the English experiment, the rectangles were labelled *BEEPA*, *BIPPA*, *BAYPA*, and *BEPPA* representing /bipə/, /bɪpə/, /bepə/, and /bɛpə/ respectively. In the Spanish experiment, the rectangles were labelled *BIPA*, *BEIPA*, and *BEPA* representing /bipa/, /beipa/, and /bepa/ respectively. The participant used a mouse to click on the response which corresponded most closely to the word they heard. If they perceived the word they heard to be a good example of the response they selected, they clicked near the top of the rectangle, and if they perceived it as a poor example, they clicked near the bottom of the rectangle. They could click in any part of the rectangle to indicate how well the stimulus matched the category. The computer program stored the response category, and the category-goodness rating (a number from 1 to 100 reflecting the height at which the participant clicked on the rectangle).

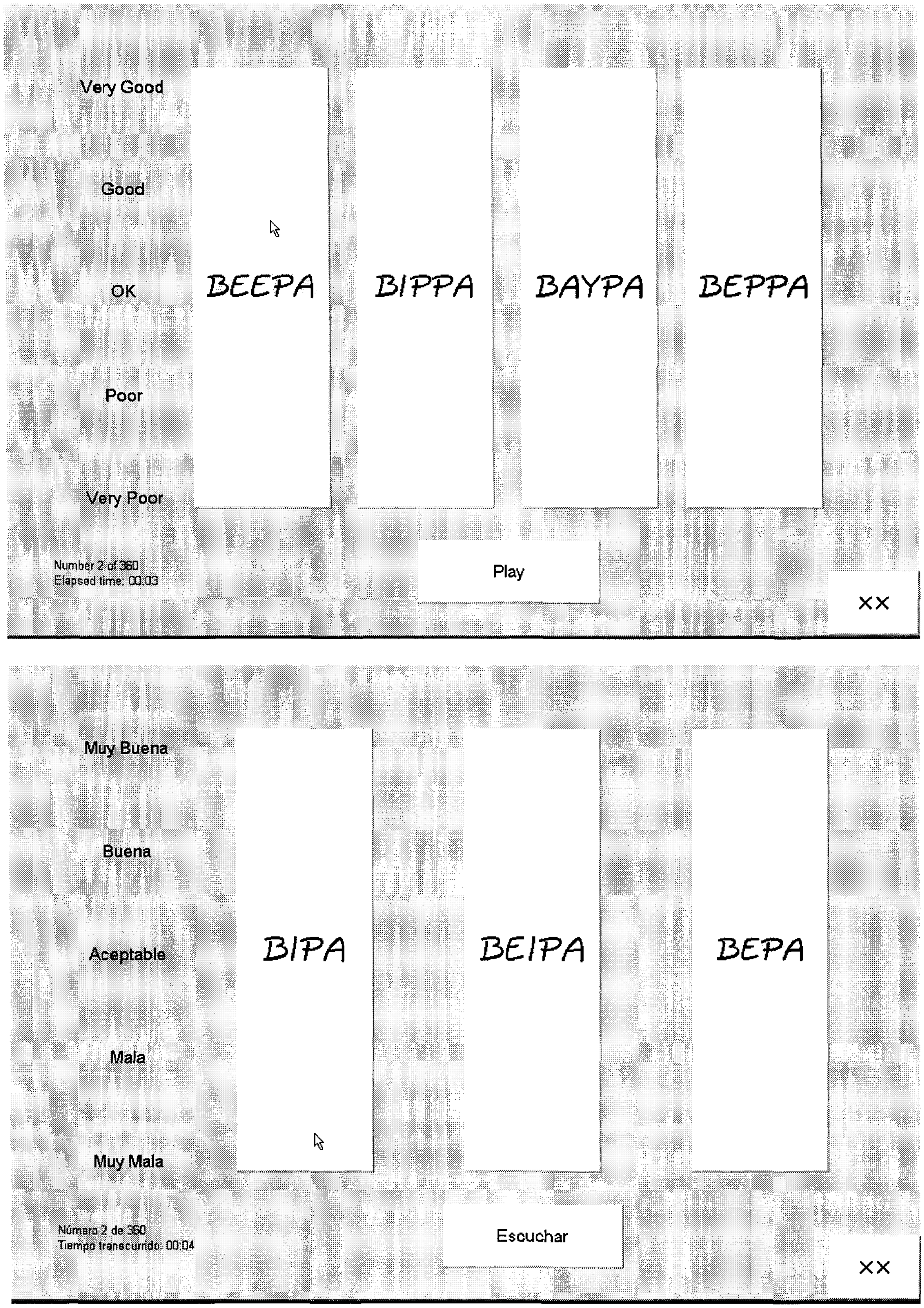


Figure 2.3 Screen shots of response options in English (top) and Spanish (bottom) perception experiments.

Participants also had the option of clicking on a play button in order to listen to the stimulus a second time before giving their response, the button then disappeared so that listening more than twice was not an option, and the researcher suggested to participants that they usually try to give a response after listening once. There was also an error button (marked "XX" in Figure 2.3) which participants could click if they had accidentally given the wrong response to the previous stimulus; they were instructed only to use this button if they had made a mousing error, not if they changed their mind after they had given a response. If a participant clicked the error button, the response to the previous stimulus was discarded, a response was not collected for the current stimulus, and the previous and the current stimuli were repeated at the end of the block. The program played a new stimulus 500 ms after the participant had given a response to the previous stimulus or pressed the error button.

The stimuli were presented in six blocks with the order of presentation randomised within each block and with the restriction that the first stimulus of a block was not the same as the last stimulus of the preceding block. Randomisations were generated on the fly so that each participant received different randomisations. Participants may find long perception experiments tiresome, and it was therefore important to keep the experiment short, especially since most participants had been asked to participate in more than one session. In order to reduce the number of responses given by each participant without adversely affecting the quality of the data, an efficient sampling procedure was developed, details of this procedure are given in Appendix 2. All 90 stimuli were presented in each of the first two blocks, and 45 stimuli were selected for presentation in each of the subsequent blocks. The 360 responses took approximately half an hour to collect, and an entire experimental session lasted a little less than an hour.

Following the perception experiment, participants were shown a graphical representation of their results, and the researcher gave them an explanation of their perception pattern.

Each monolingual English speaker (except as noted in Table 2.2) participated in two versions of the perception experiment: the first version was the English experiment described above, and the second version was one of four control conditions. The first control condition was an exact replication of the original English perception experiment; this was included to provide an indication of intralistener variability on repetitions of the perception experiment, and hence a lower limit on the size of meaningful interlistener or intercondition differences. In the second control condition, the carrier sentence was in Spanish with the response options still in English; this was included to provide an indication of differences in responses that may result because of non-language-mode differences in the carrier sentences. Since the listeners were monolingual, any differences in responses between the two versions of the experiment would not be because they were listening in English mode in one version and in Spanish mode in the other, and if similar differences in bilinguals were found they would therefore not be attributable to language-mode differences. Differences across the two versions of the experiment could be due to factors such as acoustic contrast effects due to formant frequency differences in the last vowels of each carrier sentence, or a foreign-accent-mode effect if listeners were familiar with Spanish accented English and adjusted their perception accordingly (see Flege & Hammond, 1982, on foreign-accented production). The third control condition used the original synthetic stimuli and final natural /pə/ but did not include a carrier sentence; this was included to give an indication of differences that may arise because of contrast effects with the carrier sentence (and other contextual effects).[6] In the fourth control condition the carrier sentence was in English but spoken by a different speaker, a 22-year-old male who had grown up in Edmonton. The voice quality properties of the synthetic stimuli were matched to the speaker's voice (the set of non-varying parameter values for this voice is provided in Table 2.5b), but the remaining properties were identical to those in the original experiment. This condition was included to provide an indication of whether there may be any distortions in perception because the listeners may have noticed that the original speaker was not a first-dialect speaker of their dialect of English.

[6] The carrier sentences had originally been included to foster English and Spanish mode perception by the bilingual participants.

3. L1 Production Results & Discussion

This section will begin with an acoustic and statistical analysis of L1-Spanish and L1-English speakers' L1 vowel productions. The production data will then be used to build models of L1-Spanish and L1-English vowel production. The L1-Spanish and L1-English models will be compared and used to make predictions as to the perception of English vowels by L1-Spanish speakers just beginning to learn Canadian English, and the perception of Spanish vowels by L1-English speakers just beginning to learn Spanish.

Models of vowel production (and models of vowel perception in Section 4) will be parameterised in terms of the *dual-target* hypothesis for vowel inherent spectral change (VISC, Nearey & Assmann, 1986; see also Gottfried, Miller, & Meyer, 1993). The dual-target hypothesis posits that the relevant acoustic properties for VISC perception are the initial and final formant values of the vowel, taking measurements at points such as 25% and 75% of the vowel duration reduces the influence of consonant transitions at the extremes of the vowel. Although not conclusive, the weight of evidence suggests that the dual-target hypothesis is superior to the competing *target plus direction* and *target plus slope* hypotheses (these hypotheses were directly compared in Nearey & Assmann, 1986; and Gottfried, Miller, & Meyer, 1993), and models parameterised according to the dual-target hypothesis are not outperformed by triple-target models or more sophisticated curve-fitting models (Hillenbrand et al., 1995; Hillenbrand, Clark, & Nearey, 2001). See Appendix 1 for a more detailed review.

3.1 Acoustic analysis

L1 production data were available from 18 monolingual Spanish speakers (see Table 2.1), 41 bilingual L1-Spanish speakers (see Table 2.3), 23 monolingual English speakers (see Table 2.2), and 26 bilingual L1-English speakers (see Table 2.4). Ten (10) recordings of each vowel were usually available from each participant. The online screening during data collection (described in Section 2.3.1), occasionally resulted in extra recordings for some of the vowels. Despite the online screening, a few recordings contained obvious errors (e.g., clipping, noise during vowel production, speaker disfluency or speaker having misspoken), and these were excluded during acoustic analysis. Each individual participant's data were screened for outliers prior to statistical analysis, outliers were reexamined and remeasured, and excluded if they proved to be problematic (e.g., if acoustic measurements and auditory perception clearly indicated that the speaker had accidentally read the wrong vowel phoneme in response to the stimulus).

For each recording, vowel duration and F1 and F2 formant tracks were measured between the end of the /b/ release burst and the drop in intensity and disappearance of formants at the onset of the /p/. The beginning and end of each vowel were manually marked based on visual examination of the raw waveform and spectrogram, with audio playback to confirm recording quality. Formant tracks were measured using the automated technique described by Nearey, Assmann, & Hillenbrand (2002):

– Try several frequency cutoff values in the expected range of the midpoint between F3 and F4 (e.g., eight values between 3000 and 4000 Hz).

– For each cutoff, apply linear predictive coding (LPC) with 9 coefficients (allows for a maximum of four formants), and track formants using a variant of Markel & Gray's (1976, p. 176–180) algorithm.

– Apply heuristics to calculate a goodness score for each trackset.

Each trackset was overlayed on the spectrogram, with the suggested best trackset indicated. The researcher manually selected the best trackset on the basis of visual match to the spectrogram (usually but not always the best trackset suggested by the algorithm), and occasionally made manual corrections when the best trackset clearly deviated from the spectrogram. Formant measurements were obtained every 2 ms using a 100 ms power-four-cosine window. Formant values at 25 and 75% of the duration of the vowel were used in statistical analyses, these values were obtained via linear interpolation (in Hertz) between the values measured at the two nearest time points.

Statistical tests were conducted on five variables: F1, ΔF1, F2, ΔF2, and duration. Prior to statistical analysis, all measures were subjected to a natural logarithm transform (a standard practice that typically results in statistically better behaved values for vowel formant and duration measures), formant values were therefore entered in log Hertz, and vowel duration in log milliseconds. F1 and F2 were measured at 25% of the duration of the vowel, and ΔF1 and ΔF2 were calculated as the difference between the log Hertz values at

25% and 75% of the duration of the vowel.

Univariate repeated-measures analyses of variance (ANOVAs) were conducted in order to determine whether there were differences between monolingual and bilingual speakers' productions, and between Spanish dialects. Each acoustic variable was tested independently with language group, gender, and vowel treated as fixed factors, and speaker as a random factor nested within language group and gender. Results of these tests are reported in Appendix 4 Tables A4.1–15.

For monolingual versus bilingual L1-English speakers there were significant (α = .05) group by vowel interactions for F1, F2, and duration; however, only in the cases of F2 for English /i/ and for English /e/ was it possible to isolate a significant (nominal α = .05, no correction for multiple comparisons) between-group difference (the bilingual group's F2 values were approximately 2% and 3% greater respectively). For all individual vowels and all acoustic variables, the magnitudes of the between-group differences in the marginal means were relatively small. Given the small size of the differences, subsequent statistical tests on L1-English production will be based on data pooled across monolingual and bilingual speakers.

For monolingual versus bilingual L1-Spanish speakers there was a significant group main effect for duration; the bilingual group's vowels were approximately 10% longer than the monolingual group's vowels. There were significant group by vowel interactions for F1, ΔF1, and ΔF2; the bilingual group had significantly higher F1 for Spanish /e/ (approximately 4% higher), and significantly smaller ΔF1 magnitude for Spanish /ei/ (approximately 12% smaller).

For Peninsular versus Mexican L1-Spanish speakers there was a significant group main effect for duration; the Mexican group's vowels were approximately 10% longer than the monolingual group's vowels. There were significant group by vowel interactions for F2 and ΔF1; the Mexican group had significantly smaller ΔF1 magnitude for Spanish /ei/ (approximately 12% smaller), and for Spanish /e/ a ΔF1 of +9 Hz compared to – 1 Hz for the Peninsular group (however the +9 Hz ΔF1 was only a 2% shift from the mean initial F1 for Spanish /e/ of 523 Hz).

The largest differences between both the monolingual versus bilingual and the Peninsular versus Mexican L1-Spanish groups were the longer vowels and smaller VISC magnitude in /ei/ for the bilingual and the Mexican groups. Since most monolingual participants were speakers of a Peninsular dialect, and all Mexican participants were bilingual, it is possible that these differences in production are due either to dialect differences or to the effect of learning English. The vowel duration difference could potentially be related to dialect: In the Spanish-speaking world, Spaniards have a reputation for speaking quickly. Since English /e/ has a smaller VISC magnitude than Spanish /ei/, the smaller VISC magnitude could potentially be an effect of L2-English learning. However, no firm conclusions can be drawn. Whether the cause is bilingualism or dialect differences, L1and L2 learners of Spanish will be assumed to be exposed to this range of variation, and since the ultimate goal is to model L1- and L2-Spanish speakers' behaviour, subsequent statistical tests involving L1-Spanish production will be based on data pooled across monolingual and bilingual speakers. Note that the size of the ΔF1 differences between monolingual Spanish versus bilingual Spanish groups (−129 Hz, 25%, change from a initial F1 of 510 Hz versus −113 Hz, 22%, change from a initial F1 of 509Hz) and between Peninsular versus Mexican groups (−126 Hz, 25%, change from a initial F1 of 506 Hz versus −114 Hz, 22%, change from a initial F1 of 513Hz), are small compared to the difference between L1-Spanish versus L1-English groups (−118 Hz, 23%, change from a initial F1 of 509 Hz versus −56 Hz, 11%, change from a initial F1 of 501Hz).

Geometric mean values for acoustic properties for each L1 group, pooled across monolingual and bilingual speakers, are given in Table 3.1 for male speakers, Table 3.2 for female speakers, and Table 3.3 for pooled male and female speakers (geometric means were calculated on log scales then converted back to Hertz and milliseconds). Geometric means for L1-Spanish and L1-English vowels pooled across gender and monolingual and bilingual speakers are plotted in Figure 3.1.

Table 3.1 Mean acoustic properties of L1 vowels produced by male L1-Spanish and L1-English speakers.

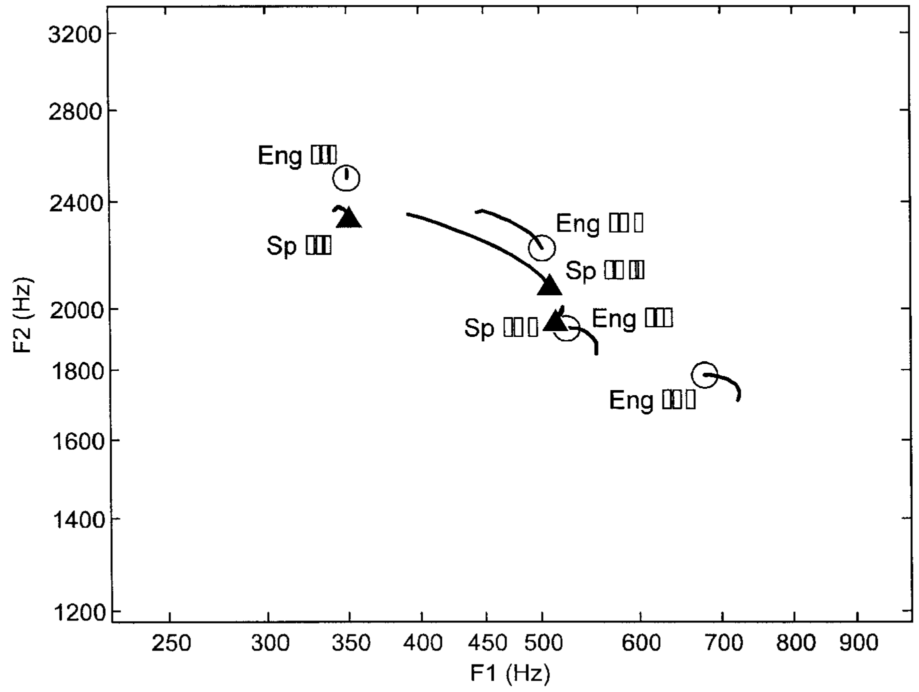
Vowel	F1 (Hz)	ΔF1 (Hz)	F2 (Hz)	ΔF2 (Hz)	duration (ms)
Sp /i/	322	-8	2078	+34	75
Sp /ei/	479	-108	1853	+238	126
Sp /e/	485	+2	1748	+20	81
Eng /i/	306	-0	2157	+21	85
Eng /ɪ/	457	+19	1707	-63	64
Eng /e/	455	-40	1921	+119	109
Eng /ɛ/	585	+38	1583	-48	82

Table 3.2 Mean acoustic properties of L1 vowels produced by female L1-Spanish and L1-English speakers.

Vowel	F1 (Hz)	ΔF1 (Hz)	F2 (Hz)	ΔF2 (Hz)	duration (ms)
Sp /i/	378	-11	2553	+47	87
Sp /ei/	536	-127	2276	+319	149
Sp /e/	538	+4	2141	+39	92
Eng /i/	376	+0	2696	+7	98
Eng /ɪ/	563	+38	2064	-89	75
Eng /e/	528	-66	2395	+149	128
Eng /ɛ/	733	+48	1901	-90	96

Table 3.3 Mean acoustic properties of L1 vowels produced by L1-Spanish and L1-English speakers, pooled across male and female speakers.

Vowel	F1 (Hz)	ΔF1 (Hz)	F2 (Hz)	ΔF2 (Hz)	duration (ms)
Sp /i/	351	-9	2324	+41	81
Sp /ei/	509	-118	2072	+279	138
Sp /e/	513	+3	1950	+30	86
Eng /i/	350	+0	2495	+13	93
Eng /ɪ/	524	+31	1932	-79	71
Eng /e/	501	-56	2218	+138	121
Eng /ɛ/	677	+44	1784	-74	91



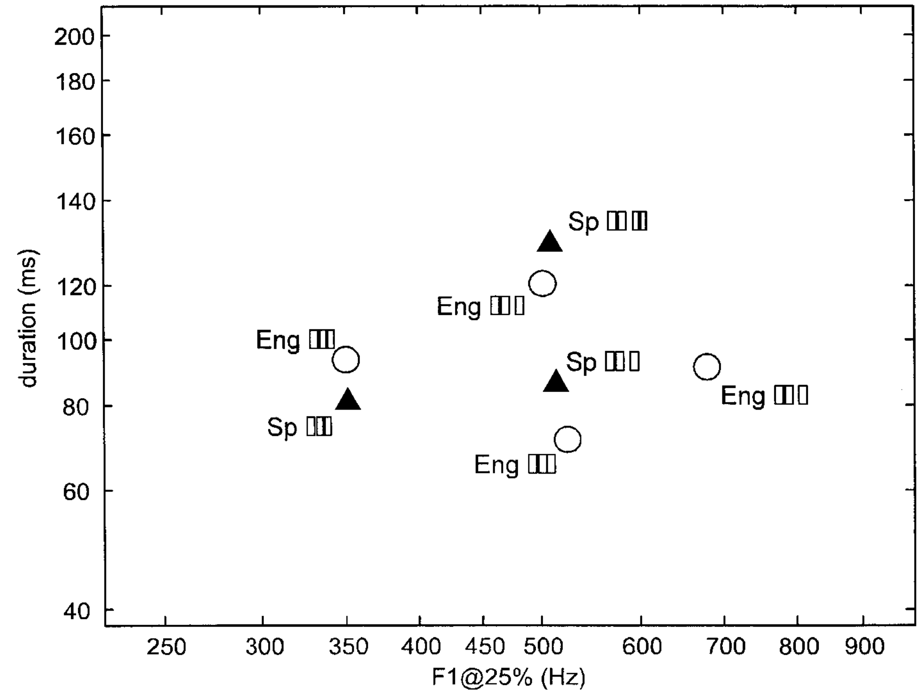


Figure 3.1 Non-normalised mean acoustic properties of L1-Spanish and L1-English vowels. Top: F1, F1, ΔF1, and ΔF2. Comet heads indicate formant values at 25% of the duration of the vowel, ends of comet tails indicate formant values at 75% of the duration of the vowel. Bottom: F1 at 25% of the duration of the vowel and vowel duration.

Multivariate Hotelling's T^2 tests, and follow-up univariate *t*-tests, were conducted on ΔF1 and ΔF2 to determine whether individual vowels had VISC magnitude significantly different from zero (results are reported in Appendix 4 Table A4.16). Univariate paired-sample *t*-tests were conducted comparing the duration of selected pairs of L1-Spanish vowels and selected pairs of L1-English vowels (results are reported in Appendix 4 Table A4.17). Univariate ANOVAs were conducted comparing selected pairs of L1-Spanish and L1-English vowels, with Speaker as a random factor nested within Language and Gender (results are reported in Appendix 4 Table A4.18–22). On the basis of visual inspection of Figure 3.1 and the statistical results reported in Appendix 4, the following observations can be made regarding L1-Spanish and L1-English production.[1]

Spanish vowels:

- Although Spanish /i/ and /e/ are traditionally described as monophthongs, they were found to have significant formant movement. However, the magnitude of the formant movement was small so it may be reasonable to regard these vowels as essentially monophthongal.[2]
  - Spanish /i/ had a significant mean ΔF1 of – 3% and a significant mean ΔF2 of +2%
  - Spanish /e/ had a significant mean ΔF2 of +2%, the mean ΔF1 was +3 Hz, less than 1% change, and not significantly different from zero
- Spanish /e/ is slightly longer than Spanish /i/, and Spanish /ei/ is substantially longer then Spanish /e/.
  - The mean duration of Spanish /e/ was 5 ms (6%) significantly longer than Spanish /i/, and the mean duration of Spanish /ei/ was 52 ms (60%) significantly longer than Spanish /e/.
- Spanish /ei/ has diverging VISC (significant mean – 23% ΔF1 and +13% ΔF2), the

[1] For brevity, the following discussion applies a nominal significance threshold of $\alpha = .05$, and accept the null hypothesis if $p > \alpha$.

[2] The small change in F2 from 25 to 75% of the duration of the vowel could be due to consonant context effects with the voiced initial consonant /b/ and the voiceless final consonant /p/, Morrison (2002c, 2006) reported higher F2 in Spanish /i/ preceding /t/ than preceding /d/.

initial and final formant values of Spanish /ei/ are relatively close to the formant values of the Spanish monophthongs /e/ and /i/ respectively (see Figure 3.1).

English vowels:

- Although introductory linguistics textbooks may transcribe English /i/ as a diphthong (e.g., Dobrovolsky, 1996, transcribes Canadian English /i/ as [ij]), English /i/ was the only vowel investigated here that did not have significant formant movement (replicating the findings of Andruski & Nearey, 1992; and Nearey & Assmann, 1986). Canadian English /i/ can therefore be regarded as a true monophthong.
    - The magnitude of VISC did not differ significantly from zero, mean ΔF1 was 0, and mean ΔF2 was +12 Hz, less than a 1% change.
- English /e/ has diverging VISC, and English /ɪ/ and /ɛ/ have converging VISC.
    - The magnitudes of VISC were significant.
        - English /e/, −11% ΔF1 and +6% ΔF2
        - English /ɪ/, +6% ΔF1 and −4% ΔF2
        - English /ɛ/, −7% ΔF1 and −4% ΔF2
- English /i/ and /ɛ/ have approximately the same duration, English /ɪ/ is shorter and English /e/ is longer.
    - English /i/ and /ɛ/ were both significantly longer than English /ɪ/ by approximately 21 ms, 30%, and significantly shorter than English /e/ by approximately 29 ms, 24%, but did not differ significantly from each other in terms of duration, 2 ms, less than 1% difference.
- The initial and final formant values of English /e/ are not close to the formant values of any other English vowels (see Figure 3.1).

Spanish versus English vowels:

- Spanish /i/ is spectrally closest to English /i/, and is slightly closer in duration to English /ɪ/.
    - However, F2 for Spanish /i/ was 7% lower than F2 for English /i/ and the two vowels had significant VISC differences.

  - – The mean duration of Spanish /i/ was 12 ms, 13%, shorter than English /i/, and 10 ms, 14%, longer than English /ɪ/. These differences were significant.[3]

– Compared to English /e/, Spanish /ei/ is substantially longer and has substantially greater VISC.

  - – Spanish /ei/ had significantly greater magnitude in both ΔF1 and ΔF2 (−23% ΔF1 and +13% ΔF2, compared to −11% ΔF1 and +6% ΔF2 for English /e/), and was significantly longer by 17 ms, 21%.[4]

– Spanish /e/ and English /ɪ/ have similar initial formant values, but English /ɪ/ is shorter, and has converging VISC.

  - – Initial F1 for English /ɪ/ was non-significantly 4% higher, and initial F2 was significantly[5] 1% lower.
  - – The mean duration of English /ɪ/ was 15 ms (17%) significantly shorter than Spanish /e/.[4]

– English /ɛ/ has substantially higher initial F1 and lower initial F2 values compared to Spanish /e/.

  - – F1 was 32% higher and F2 was 9% lower.

The L1-English production results are generally in accord with earlier reports on the acoustic properties of Canadian English vowels; however, the position of /ɪ/ and /ɛ/ relative to /i/ and /e/ in the F1–F2 space differs from that reported in Andruski & Nearey (1992) and Nearey & Assmann (1986): /ɪ/ and /ɛ/ have higher F1 and lower F2 values (compare Figure 3.1 with Figure 1.1).[6] As in the earlier studies, /ɪ/ is a more centralised vowel than /e/ and

[3] The mean Spanish /i/ duration would have been 4 ms shorter if only monolingual Spanish speakers' data had been used instead of pooling across monolingual and bilingual L1-Spanish speakers.

[4] The mean Spanish /ei/ duration would have been 6 ms shorter if only monolingual Spanish speakers' data had been used. The mean VISC would have been larger: −25% ΔF1 and +15% ΔF2.

[5] Although not significantly different after a minimum Bonferroni correction for the two tests ΔF1 and ΔF2 differences for this pair of vowels.

[6] A subset of the L1-English vowels in the present study were remeasured to confirm that the difference was not due to operator error.

VISCs run in parallel but in the opposite direction; however, whereas in the earlier studies the initial and final F1 values of /ɪ/ were similar respectively to the final and initial F1 values of /e/, in the present study the initial F1 value of /ɪ/ is similar to the initial F1 value of /e/. The differences in formant values between the present study and the earlier acoustic studies may be due to diachronic change. They are consistent with Clarke, Elms, & Youssef's (1995) *Canadian Shift* hypothesis in which /ɒ/ and /ɔ/ have merged, /æ/ is backing, and /ɪ/ and /ɛ/ are lowering (see also Boberg, 2005; Esling & Warkentyne, 1993; Hagiwara, 2006). The Nearey & Assmann data were collected no later than 1981, approximately 10 years before the Andruski & Nearey data, and the data for the present study were collected in 2005, at least 13 years after the Andruski & Nearey data. In the present study, the separation between English /ei/ and /ɪ/ was also slightly greater for female than for male speakers (compare Tables 3.1 and 3.2), consistent with gender differences previously hypothesised to be due to females leading the Canadian Shift.

Alternatively, the differences may, at least in part, be due to differences in context and measurement procedures: Whereas the present study made use of formant measurements taken at 25 and 75% of the duration of the vowel in /bVp/ context in a carrier sentence (10 replications by 49 speakers), Nearey & Assmann (1986) measured formant values as early and as late as possible in isolated vowels (2 replications by 10 speakers), and Andruski & Nearey (1992) measured formant values 40 ms after the release and before the closure of the consonants in /bVb/ context (6 speakers, as well as in isolated vowels produced by 2 speakers).[7] For the longer English /e/ (mean duration 121 ms), the present study therefore measured formant values closer to the centre of the vowel compared to Nearey & Assmann, but there was probably less difference with respect to the shorter /ɪ/ and /ɛ/ (mean durations of 71 and 91 ms). To explore the potential effect of measurement point, formant measurements from the recordings in the present study were extracted at 10 and 90% of the duration of the vowel, and at 10 ms from vowel edges. Measurements taken at these points did not result in substantially different F1–F2 relationships between English /e/ and /i/ compared to measurements taken at 25 and 75% of the duration of the vowel.

[7] Each vowel was produced once, and all vowels were substantially longer than 80 ms (T. M. Nearey, personal communication, 1 February 2006).

To explore the potential effects of diachronic change and of consonant context, formants of vowels in an unpublished database of Western Canadian English speech recorded by T. M. Nearey circa 1992 were measured using the same procedures as in the present study. One recording of each nominal monophthong and phonetic diphthong of Canadian English was available from five male and five female speakers in /bVb/ and /bVp/ isolated-word contexts. The first context matches that of Andruski & Nearey (1992), and the second that of the present study. In /bVb/ context, the relative difference in the location of English /e/ and /ɪ/ in the F1–F2 space was similar to the results reported in Andruski & Nearey, and in /bVp/ it was similar to the results of the present study. The differences between the results of Andruski & Nearey and those of the present study therefore appear to be due to the contextual difference of voiced versus voiceless postvocalic consonant (see Summers, 1987). The differential influence of the consonant context on English /e/ and /ɪ/ may be related to intrinsic vowel duration differences.

English vowel production data was also collected in isolated /bVpə/ and /bVp/ words. Results from these contexts are reported in Appendix 5.

3.2 L1-Spanish Production Model

Canonical discriminant function analyses (CDFAs, see Johnson, 1998; Klecka, 1980; Tatsuoka, 1970) can lead to insightful summaries of potentially complex multivariate patterns. In the present study, CDFAs allowed the five-dimensional acoustic space to be summarised in two dimensions facilitating graphical representations of the data. Prior to the analyses, formant values were normalised using a variant of log mean normalisation (Nearey 1978, 1989; Nearey & Assmann, in press; Morrison & Nearey, in press). Vowel duration was independently normalised using the same procedure. Details of the normalisation procedure are given in Appendix 6. Compared to CDFA models trained on non-normalised data, the versions trained on normalised data had higher correct-classification rates, and the English production model was more similar to the English perception model in Section 4.[8]

A CDFA was fitted to the L1-Spanish speakers' L1 vowel data (fitted to individual

[8] Data were not normalised prior to the tests in Appendix 4 because the adjustments for inter-speaker differences implied by log mean normalization were subsumed by the Speaker and Gender factors in repeated measures ANOVA or avoided altogether via within-speaker pairings.

vowel productions). Summary statistics from the derivation of the canonical discriminant functions are given in Table 3.4, unstandardised coefficients and total structure coefficients are given in Table 3.5 (total structure coefficients are univariate correlations between each of the original variables and the discriminant functions, see Klecka, 1980, p. 31), and a plot of the data transformed by the first and second functions is given in Figure 3.2.

Table 3.4 Summary statistics from derivation of canonical discriminant functions for L1-Spanish vowels (Wilks's Λ before the corresponding function was derived). Significance levels in χ^2 tests are unlikely to be accurate because of heterogeneity in the data due to pooling across speakers.

Function	Eigen values	Relative percentage	Canonical correlation	Wilks's Λ	χ^2	*df*	*p*
1	8.770	66.5	.947	.019	6993.336	10	.000
2	4.418	33.5	.903	.185	2977.231	4	.000

Table 3.5 Unstandardised canonical discriminant function coefficients and total structure coefficients from the CDFA trained on L1-Spanish vowels.

	Unstandardised coefficients		Total structure coefficients	
Original variables	Function 1	Function 2	Function 1	Function 2
Constant	277.695	223.167		
F1	-22.989	-29.629	-.737	-.655
ΔF1	12.324	-24.705	.840	-.454
F2	-10.487	-7.335	.400	.694
ΔF2	-28.513	14.804	-.834	.350
duration	-12.676	1.684	-.912	.249

A leave-one-speaker-out cross-validation was conducted (the vowel productions of each speaker were classified on the basis of a CDFA trained on the remainder of speakers' vowel productions), and the resulting confusion matrix is given in Table 3.6. Following conversion of each case to canonical discriminant function values, classification can be performed by assigning the case to the nearest group centroid (mean values calculated on the basis of known group membership) using Euclidian distance. Overall correct classification was 99.1% (proportional reduction in error τ = .991). The CDFA was therefore highly successful at classifying the Spanish vowel productions.

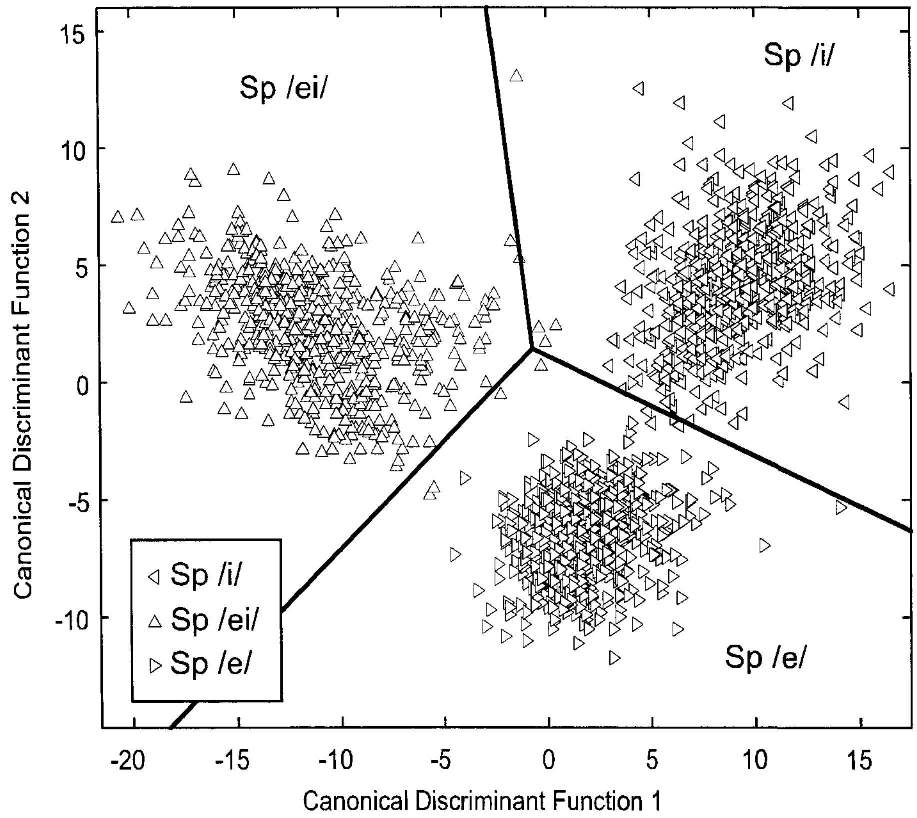


Figure 3.2 Location of L1-Spanish speakers' vowel productions and linear boundaries in the Function 1 – Function 2 space of a linear CDFA trained on L1-Spanish speakers' vowel production data.

Table 3.6 Leave-one-speaker-out cross-validation confusion matrix for the linear CDFA trained on L1-Spanish vowels. Values are percentages summing to 100 along each row, blank cells have values of zero.

	Predicted		
Produced	Sp /i/	Sp /ei/	Sp /e/
Sp /i/	99.3		0.7
Sp /ei/	1.0	98.3	0.7
Sp /e/	0.3		99.7

An alternative classification method involves calculating the a posteriori probability of membership of each group on the basis of multivariate probability density functions, and assigning each case to the group for which it has highest predicted probability. This method can be applied to the original variable values, and will give the same results as the canonical discriminant function method with Euclidian distances if all of the canonical discriminant function values are used, and the same pooled-across-groups estimate of the covariance matrix is used to classify the stimuli as is used to derive the canonical discriminant functions. Classification on the basis of a single pooled covariance matrix is known as *linear* discriminant analysis, a variant is *quadratic* discriminant analysis which uses a different covariance matrix for each group (see Hastie, Tibshirani, & Friedman, 2001, p. 88). In order to quadratically classify data and produce the two dimensional graphical representations of quadratic classification below, production data were transformed to linear canonical discriminant functions, and classification boundaries in the transformed space were calculated using a quadratic classifier. For convenience, this will be referred to as quadratic CDFA. This procedure was recommended in Gnanadesikan, (1977, §4.2.1).

Table 3.7 Leave-one-speaker-out cross-validation confusion matrix for the quadratic CDFA trained on L1-Spanish vowels. Values are percentages summing to 100 along each row, blank cells have values of zero.

	Predicted		
Produced	Sp /i/	Sp /ei/	Sp /e/
Sp /i/	100.0		
Sp /ei/	0.3	99.7	
Sp /e/	0.5	0.2	99.3

The CDFA with linear classification misclassified a number of /ei/ productions as /i/ and /e/, but no /e/ or /i/ productions were misclassified as /ei/. The linear boundaries therefore appear to be misplaced. A visual inspection of Figure 3.2 indicates that the /ei/ productions have a greater variance in the first canonical function dimension than do the other two vowels, and Box's test of equality of covariance matrices indicated that there was a significant difference between the covariance matrices of the three vowel categories: M = 393.310, F(6, 74926079.5) = 65.442, $p < .05$. A new classification model was calculated, using a quadratic rather than a linear classifier on the canonical function values. The location of the quadratic boundaries is shown in Figure 3.3, and the classification confusion matrix is given in Table 3.7. Use of the quadratic CDFA resulted in a slight increase in the leave-one-participant-out correct-classification rate for the L1-Spanish speakers' production data: a correct classification rate of 99.7% (τ = .997) compared to 99.1% for the linear CDFA. The quadratic CDFA model will be used in subsequent analyses of Spanish production data because it also increased the similarity between the production-based model and the perception-based model discussed in Section 4.

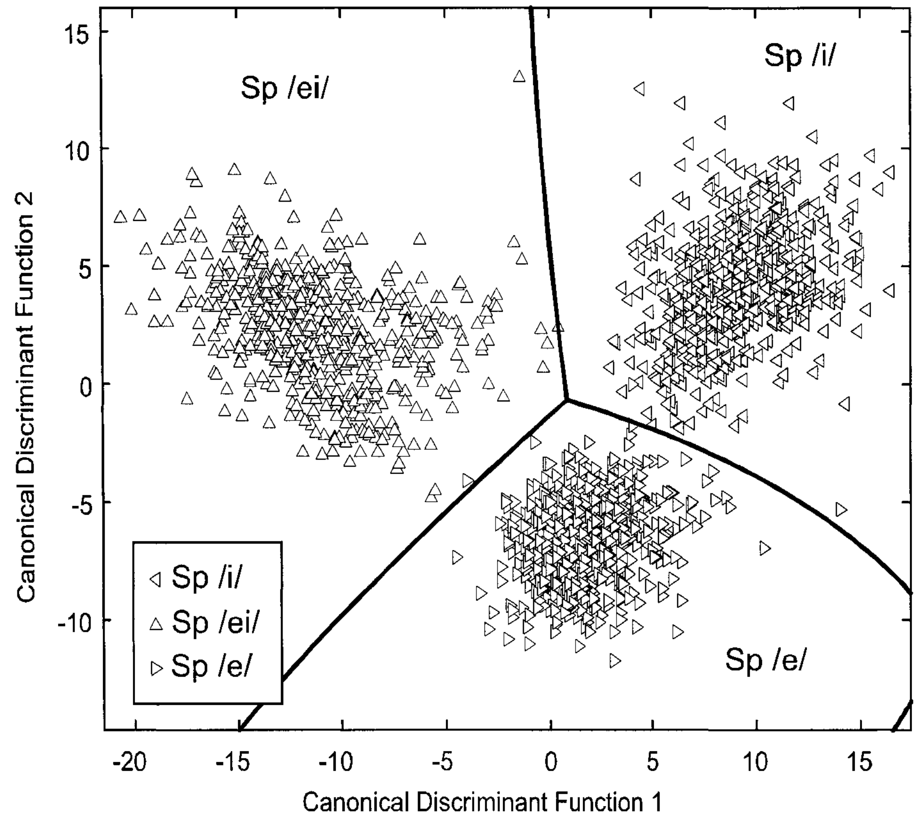


Figure 3.3 Location of L1-Spanish speakers' vowel productions and quadratic boundaries in the Function 1 – Function 2 space of a quadratic CDFA trained on L1-Spanish speakers' vowel production data.

3.3 L1-English Production Model

A CDFA was fitted to the L1-English speakers' L1 vowel data. Summary statistics from the derivation of the canonical discriminant functions are given in Table 3.8, unstandardised coefficients and total structure coefficients are given in Table 3.9, and a plot of the data transformed by the first and second functions is given in Figure 3.4.

A leave-one-speaker-out cross-validation was conducted using the linear classifier. The classification confusion matrix is given in Table 3.10. Overall correct classification was 98.7% (τ = .987).[9] The CDFA was therefore very successful at classifying the English vowel categories.

Table 3.8 Summary statistics from derivation of canonical discriminant functions for L1-English vowels (Wilks's Λ before the corresponding function was derived). Significance levels in χ^2 tests are unlikely to be accurate because of heterogeneity in the data due to pooling across speakers.

Function	Eigen values	Relative percentage	Canonical correlation	Wilks's Λ	χ^2	*df*	*p*
1	13.821	71.5	.966	.008	9432.809	15	.000
2	5.224	27.0	.916	.126	4101.411	8	.000
3	0.279	1.4	.467	.782	485.861	3	.000

Table 3.9 Unstandardised canonical discriminant function coefficients and total structure coefficients from the CDFA trained on L1-English vowels.

Original variables	Unstandardised coefficients			Total structure coefficients		
	Function 1	Function 2	Function 3	Function 1	Function 2	Function 3
Constant	-75.071	307.919	-90.257			
F1	-16.881	-21.914	-1.259	-.881	-.461	-.086
ΔF1	-11.670	5.755	20.109	-.553	.562	.463
F2	20.963	-14.560	4.944	.952	.138	.144
ΔF2	33.124	-32.187	-36.896	.669	-.553	-.341
duration	4.600	-14.752	13.685	.474	-.806	.315

[9] The quadratic classifier gave a correct-classification rate of 98.6% (τ = .986). The reduction in correct-classification rate compared to the linear model indicates that the quadratic model is overfitted to the data sample. The quadratic model was also less similar to the perception model than the linear model.

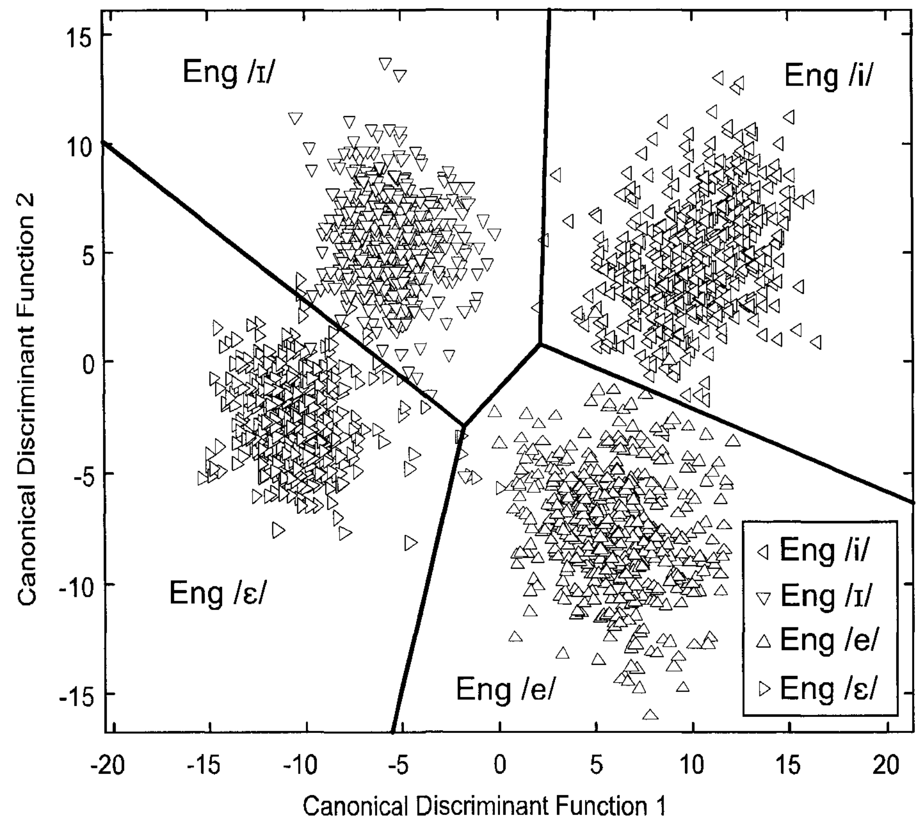


Figure 3.4 Location of L1-English speakers' vowel productions and quadratic boundaries in the Function 1 – Function 2 space of a linear CDFA trained on L1-English speakers' vowel production data.

Table 3.10 Leave-one-speaker-out cross-validation confusion matrix for the linear CDFA trained on L1-English vowels. Values are percentages summing to 100 along each row, blank cells have values of zero.

	Predicted			
Produced	Eng /i/	Eng /ɪ/	Eng /e/	Eng /ɛ/
Eng /i/	99.8		0.2	
Eng /ɪ/		98.0	0.2	1.8
Eng /e/	0.6		99.4	
Eng /ɛ/		1.4	0.8	97.8

3.4 Comparison of L1-Spanish and L1-English Production Models

The quadratic CDFA trained on L1-Spanish vowels was used to classify the L1-English vowels; the resulting confusion matrix is given in Table 3.11 and the Function 1 – Function 2 plot in Figure 3.5.

Table 3.11 Confusion matrix for classification of L1-English vowels by the quadratic CDFA trained on L1-Spanish vowels. Values are percentages summing to 100 along each row, blank cells have values of zero.

	Predicted		
Produced	Sp /i/	Sp /ei/	Sp /e/
Eng /i/	99.8		0.2
Eng /ɪ/	1.0		99.0
Eng /e/	3.0	82.4	14.6
Eng /ɛ/			100.0

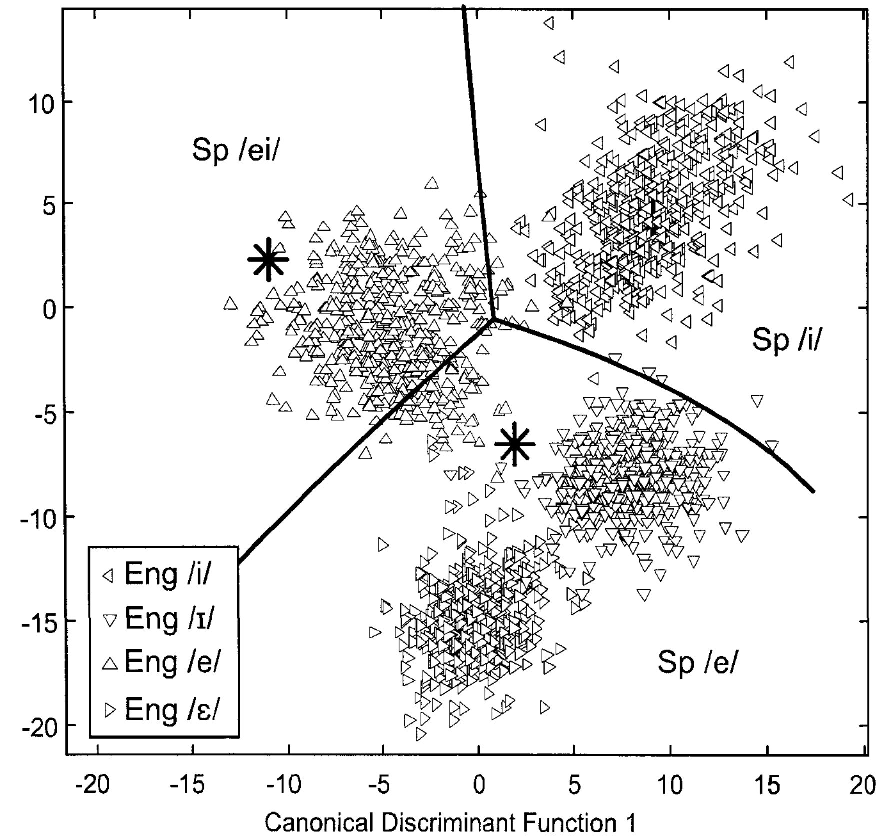


Figure 3.5 Location of L1-English speakers' vowel productions in the Function 1 – Function 2 space of a CDFA trained on L1-Spanish speakers' vowel production data. Stars represent centroids of L1-Spanish data.

The CDFA trained on L1-English vowels was used to classify the L1-Spanish vowels; the resulting confusion matrix is given in Table 3.12 and the Function 1 – Function 2 plot in Figure 3.6.

Table 3.12 Confusion matrix for classification of L1-Spanish vowels by the CDFA trained on L1-English vowels. Values are percentages summing to 100 along each row, blank cells have values of zero.

	Predicted			
Produced	Eng /i/	Eng /ɪ/	Eng /e/	Eng /ɛ/
Sp /i/	99.3	0.3	0.3	
Sp /ei/			100.0	
Sp /e/	0.7	55.0	30.8	13.5

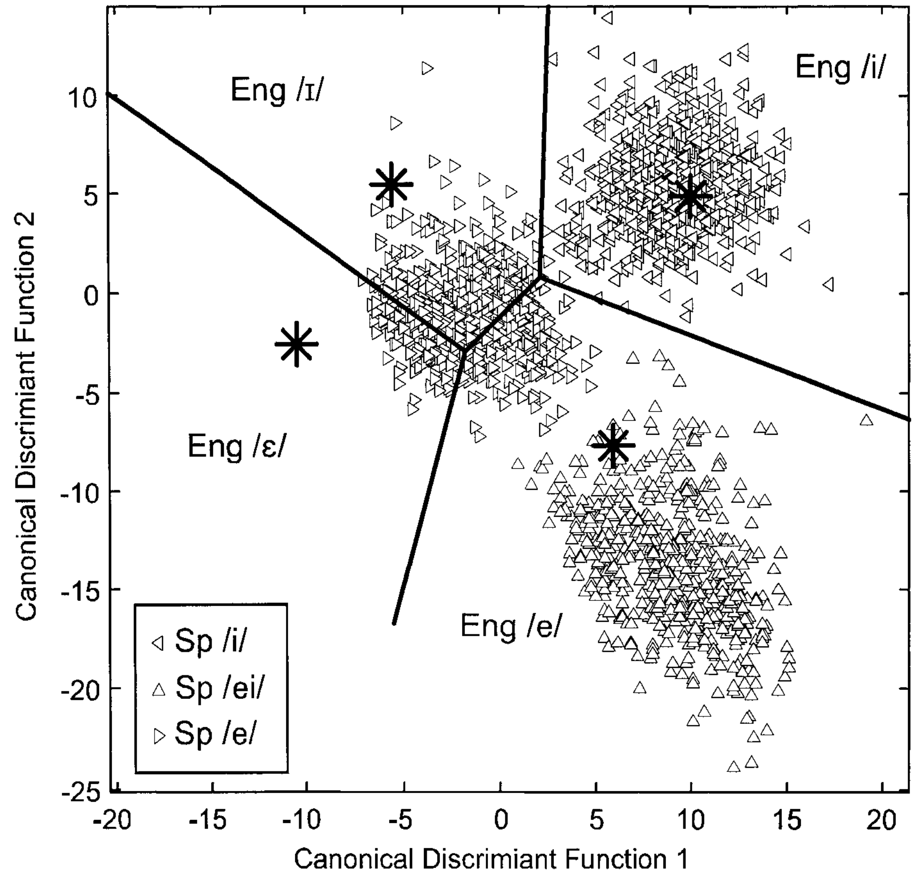


Figure 3.6 Location of L1-Spanish speakers' vowel productions in the Function 1 – Function 2 space of a CDFA trained on L1-English speakers' vowel production data. Stars represent centroids of L1-English data.

Comparing L1-English and L1-Spanish vowels on the basis of the CDFA models:

– The acoustic properties of Spanish /i/ and English /i/ are similar. The L1-Spanish production model classified almost all instances of L1-English /i/ as Spanish /i/ and vice versa.

– English /e/ and Spanish /ei/ are similar. All instances of Spanish /ei/ were classified as English /e/ by the L1-English production model. Most instances of Spanish /ei/ are more extreme than English /e/, being further away from other English vowels than most instances of English /e/. Most instances of English /e/ were classified as Spanish /ei/ by the L1-Spanish production model, but some instances were classified as Spanish /e/; most instances of English /e/ are less extreme than Spanish /ei/, being closer to Spanish /e/ than most instances of Spanish /ei/.

– Spanish /e/ is intermediate between English /ɪ/, /e/, and /ɛ/, but closest to English /ɪ/. The L1-English production model classified more than half the instances of Spanish /e/ as English /ɪ/, more than a quarter as English /e/, and a substantial number as English /ɛ/. Almost all instances of English /ɪ/ were classified as Spanish /e/ by the L1-Spanish production model.

– English /ɛ/ is the furthest English vowel from any Spanish vowel category, but the nearest Spanish vowel category is Spanish /e/. All instances of English /ɛ/ were classified as Spanish /e/ by the L1-Spanish production model.

3.5 Predictions

The results of discriminant analyses of production data have been found to correlate with L1-listeners' perception (e.g., Andruski & Nearey, 1992; Assmann, Nearey, & Hogan, 1982; Hillenbrand & Nearey, 1999; Nearey & Assmann, 1986; Parker & Diehl, 1984). Under the hypothesis that the CDFA models above have captured factors that are close to those of monolingual perception, the following predictions as to the initial state of L2 learning would seem reasonable.

L1-Spanish speakers just beginning to learn Canadian English are predicted to make the patterns of assimilation of English vowels to Spanish vowels shown in Figure 3.7. The following predictions for assimilation of English vowels to Spanish vowels are made for L1-Spanish speakers just beginning to learn Canadian English:

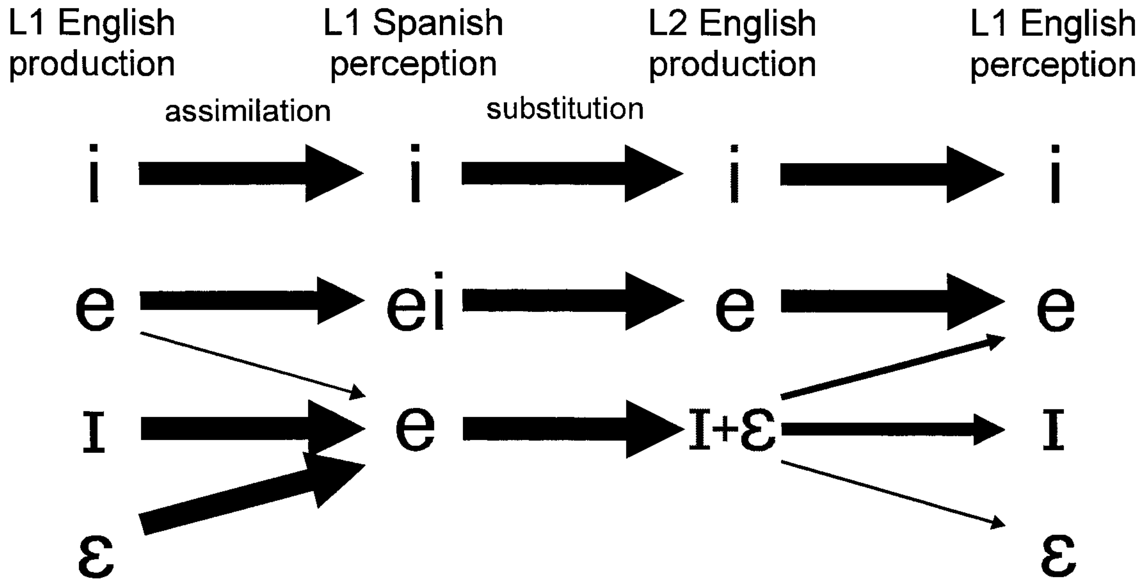


Figure 3.7 Predicted pattern of assimilation of L1-English vowel productions to L1-Spanish vowels, substitution of L1-Spanish vowels for L2-English production, and perception of L2-English productions by L1-English listeners. Predictions made on the basis of comparison of L1-English and L1-Spanish vowel production models. Arrow thickness indicates percentage of instances of vowel category on left perceived as vowel category on right (predicted percentages of less than 5% not shown).

– They will assimilate almost all instances of English /i/ to Spanish /i/.

– They will assimilate almost all instances of English /ɪ/ and /ɛ/ to Spanish /e/, although the differences in duration and VISC might make the English vowels noticeably poor versions of Spanish /e/.

– Most instances of English /e/ will be assimilated to Spanish /ei/, but some will be assimilated to Spanish /e/ causing some perceptual confusion.

In production, L1-Spanish speakers beginning to learn Canadian English are predicted to make the patterns of substitution of Spanish vowels for English vowels shown in Figure 3.7. The following predictions for substitution of Spanish vowels for English vowels in production are made for L1-Spanish speakers just beginning to learn Canadian English:

– They will substitute Spanish /i/ for English /i/. Since most instances of English /i/ are assimilated to Spanish /i/ in perception, the English /i/ category is equated with the Spanish /i/ category, and the Spanish /i/ category is therefore used to produce instances of L2-English /i/. Substitution of Spanish /i/ for English /i/ will not create perception problems for L1-English listeners.

– They will substitute Spanish /e/ for both English /ɪ/ and English /ɛ/. This will cause serious perception problems for L1-English listeners: If an L1-English listener hears an L1-Spanish speaker say either English /ɛ/, /ɪ/, or /e/, this could correspond to any one of an intended English /ɪ/, /e/, or /ɛ/.

– They will substitute Spanish /ei/ for English /e/ because most instances of English /e/ are assimilated to Spanish /ei/. L1-English listeners will perceive these productions as accented versions of English /e/.

L1-Canadian-English speakers just beginning to learn Spanish are predicted to make the patterns of assimilation of Spanish vowels to English vowels shown in Figure 3.8. The following predictions for assimilation of Spanish vowels to English vowels are made for L1-Canadian-English speakers just beginning to learn Spanish:

– They will assimilate almost all instances of Spanish /i/ to English /i/.

– They will assimilate all instances of Spanish /ei/ to English /e/, although many instances of Spanish /ei/ will be noticeably exaggerated versions of English /e/.

– Spanish /e/ could prove problematic perceptually since some instances would be assimilated to English /ɪ/, some to English /e/, and some to English /ɛ/.

In production, L1-Canadian-English speakers just beginning to learn Spanish are predicted to make the patterns of substitution of English vowels for Spanish vowels shown in Figure 3.8. The following predictions for substitution of English vowels for Spanish vowels in production are made for L1-Canadian-English speakers just beginning to learn Spanish:

– They will substitute English /i/ for Spanish /i/, without creating perception problems for L1-Spanish listeners.

– They will substitute English /e/ for Spanish /ei/, with some perception problems for L1-Spanish listeners. L1-Spanish listeners will usually correctly perceive L1-English speakers' intended Spanish /ei/, but will sometimes misperceive them as Spanish /e/.

– They will substitute English /ɪ/ for Spanish /e/. Since they assimilate Spanish /ei/ and some instances of Spanish /e/ to English /e/, and assimilate some instances of Spanish /e/ to English /ɪ/, they perceive an English /e/–/ɪ/ contrast and will substitute the two English vowels when producing the contrast. Although the use of the English

/e/–/ɪ/ boundary leads to differentiating only around seventy percent of instances of Spanish /e/ from Spanish /ei/, most instances of vowels assimilated to English /e/ are Spanish /ei/ and the only instances of Spanish vowels assimilated to English /ɪ/ are instances of Spanish /e/, therefore English /e/ is the predicted substitute for Spanish /ei/, and English /ɪ/ is the predicted production substitute for Spanish /e/. English /ɪ/ rather than English /ɛ/ will be substituted for Spanish /e/ because more instances of Spanish /e/ are assimilated to English /ɪ/ than to English /ɛ/.

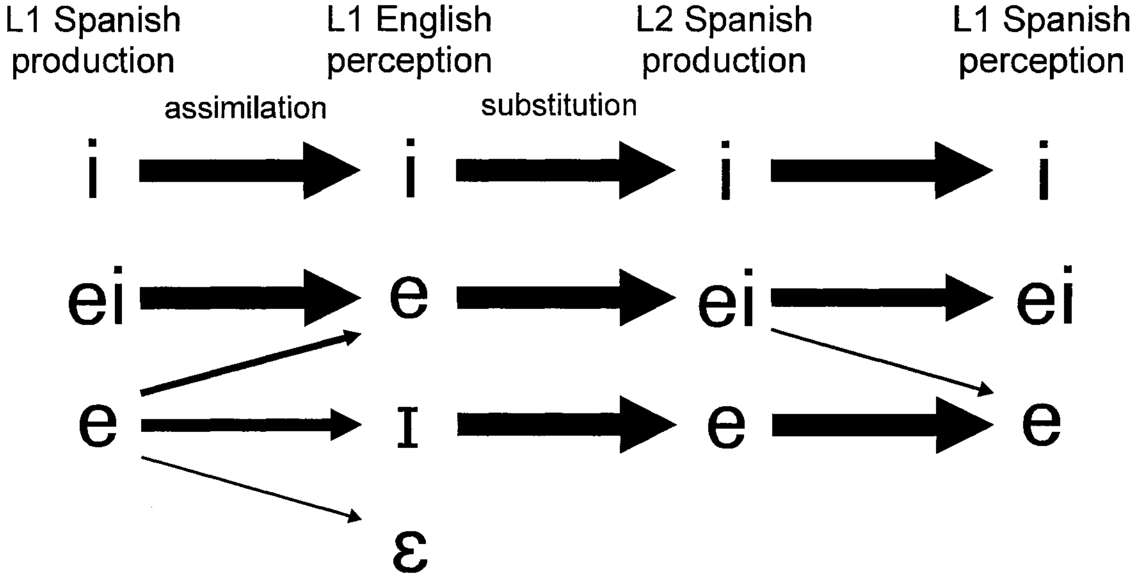


Figure 3.8 Predicted pattern of assimilation of L1-Spanish vowel productions to L1-English vowels, substitution of L1-English vowels for L2-Spanish production, and perception of L2-Spanish productions by L1-Spanish listeners. Predictions made on the basis of comparison of L1-English and L1-Spanish vowel production models. Arrow thickness indicates percentage of instances of vowel category on left perceived as vowel category on right (predicted percentages of less than 5% not shown).

These predictions as to the initial state of L2 learning are purely a priori, based on an extension of L1 production models. In Section 4, a parallel set of predictions will be made for synthetic stimuli on the basis of L1 perception models. The L1 production and perception models will be compared in Section 5, and in Section 6 L1-English listeners assimilation of L1-Spanish vowels will be directly tested in a natural vowel perception experiment.

4. L1 Perception of Synthetic Vowels Results & Discussion

This section will begin with a statistical analysis of L1-Spanish and L1-English listeners' L1-vowel identification responses for the synthetic speech continuum. The perception data will be used to build models of L1-Spanish and L1-English vowel perception. The L1-Spanish and L1-English perception models will be compared with each other in order to make predictions as to how L1-Spanish listeners just beginning to learn English will perceive the synthetic stimuli in terms of English categories, and how L1-English listeners just beginning to learn Spanish will perceive the synthetic stimuli in terms of Spanish categories. In Section 5, the L1 perception models will be compared with the L1 production models from Section 3.

4.1 Results

L1 perception data were available from 18 monolingual Spanish speakers (see Table 2.1), 41 bilingual L1-Spanish speakers (see Table 2.3), 18 monolingual English speakers (see Table 2.2), and 27 bilingual L1-English speakers (see Table 2.4). Logistic regression models were fitted to the vowel identification data; the application of this type of model to speech perception data is discussed in detail in Appendix 7 (see also Morrison, 2005a, 2005b; Nearey, 1990, 1997), and general introductions to applied logistic regression include Hosmer & Lemeshow (2000), Menard (2001), and Pampel (2000). Models are described here using the following notation:

V – a set of bias coefficient on each vowel category

expands to $i + ɪ + e + ɛ$ for English model and $i + ei + e$ for Spanish model

F1×V, ΔF1×V, dur×V

– a set of stimulus-tuned coefficients on each vowel category (F1-tuned, ΔF1-tuned, duration-tuned)

expands to F1×i + F1×$ɪ$ + F1×e + F1×$ɛ$, etc.

i, ɪ, e, ɛ, ei – bias coefficient on individual vowel categories

F1×*i*, etc. – stimulus-tuned coefficients on individual vowel categories

F1×(*i*-*ɪ*), etc. – contrast coefficients: difference between stimulus-tuned coefficients on individual vowel categories

In the synthetic stimuli, F1 and F2 were 100% correlated (see Tables 2.6 and 2.7), and therefore only one is referenced in the model. F1 and duration stimulus properties were entered as continuous variables in just-noticeable difference (JND) units. The use of JND units will facilitate the comparison of the magnitude of F1- and duration-tuned logistic regression coefficient values, which will be important when L2 perception results are analysed in Sections 8 and 9.[1] Kewley-Port (2001) reported a JND of 0.3 bark for formant frequencies in normal discourse. The F1 and F2 properties of the synthetic stimuli (see Table 2.6) were converted to bark (B1 and B2, using the inversible formula from Traunmüller, 1990), then the B1–B2 space Euclidian distance from the first stimulus [$B1_0$ = bark(283Hz), $B2_0$ = bark(2090Hz)] was calculated and divided by the JND of 0.3 bark.

$$B = \left(26.81F/(1960+F)\right) - 0.53$$

$$F1_{JND} = \sqrt{(B1-B1_0)^2 + (B2-B2_0)^2}\,/0.3$$

Noteboom & Doodeman (1980) reported a vowel duration JND of approximately 5 ms for a base duration of 90 ms. The duration properties of the synthetic stimuli (80, 95, 110 ms) were converted to JNDs using the following formula, with the zero point set to the shortest stimulus value (dur_0 = 80 ms).

$$dur_{JND} = \log_{1+(5/90)}(dur/90) - \log_{1+(5/90)}(dur_0/90)$$

A similar Weber fraction of 0.5 was used by Smits, Sereno, & Jongman (2004). For convenience, the JND subscripts in $F1_{JND}$ and dur_{JND} will be dropped in subsequent discussion.

[1] Using JNDs rather then Hertz and milliseconds resulted in approximately the same goodness of fit: SAEP of 5.11 vs 5.16 for model 5 fitted to monolingual Spanish data, SAEP of 6.40 vs 6.21 for model 5 fitted to monolingual English data, and no change in modal agreement for either (SAEP and modal agreement are defined below).

ΔF1 was entered as three discrete levels. Results of a VISC perception study by Morrison & Nearey (2005) indicated that VISC perception is not linear when measured as as a change in log Hertz (further research is in progress to determine the exact nature of the non-linear relationship between perception and physical acoustic measures of VISC).[2] Dummy-coding each VISC value in the stimuli as a discrete level allows for an arbitrary non-linear relationship between VISC and the predicted probability of identification of each vowel, and led to a substantial improvement in goodness-of-fit over models in which VISC was coded as a continuous variable.[3] Coding the three levels of VISC requires two discrete-level variables: $\Delta F1 = [\Delta F1_{-}\ \Delta F1_{+}]$, −99 Hz = [1 0], 0 Hz = [0 0], and +99 Hz = [0 1].

4.1.1 L1-Spanish speakers' perception

A series of nested models were fitted to the monolingual Spanish listeners' pooled vowel identification data, the models and corresponding goodness-of-fit measures are given in Table 4.1. Quasi-likelihood *F* tests (see McCullagh & Nelder, 1989, and Nearey, 1990, 1997), testing the improvement in goodness-of-fit are given in Table 4.2.[4] G^2 is the deviance statistic used to fit the logistic regression model, the other two measures are more intuitive indicators of the goodness-of-fit between the model and the raw data: SAEP is the sum of absolute errors in proportions, the difference between the listeners' proportion of responses for each vowel category for each stimulus and the proportion predicted by the model for each

[2] Both the linear and quadratic versions of the L1-Spanish production model had very high correct-classification rates (99.1 and 99.7%), so at least around the mean VISC properties in the L1-Spanish production data a linear or a quadratic model is a good fit. However, when used as a classifier for stimuli with VISC properties atypical of the training data, such as L1-English vowel productions, the L1-Spanish production model may not be a good predictor of VISC perception because it lacks the appropriate non-linearities.

[3] For model 5 with F1 entered in Hertz and duration entered in milliseconds, there was a decrease in SAEP of more than 2.5 percentage points, and an increase in MA of more than 5 percentage points, when ΔF1 was entered as three discrete levels rather than as a continuous variable in Hertz (SAEP and MA are defined below).

[4] The quasi-likelihood procedure is one approach to dealing with the heterogeneity introduced by using data pooled across listeners. Another approach described in Gumpertz & Pantula (1989) applies second-stage multivariate tests on the sets of coefficients from models fitted to individual listeners' responses. Both methods were applied in Nearey (1997).

vowel category for each stimulus, summed over all stimuli, it may be expressed as a percentage of the number of stimuli (this measure is described in detail in Appendix 3). MA is the modal agreement, the number of times that the most probable response category to a stimulus predicted by the model is the same as the most frequent vowel category response given by listeners for that stimulus, summed over all stimuli, it may be expressed as a percentage of the number of stimuli (the modal agreement given here is the agreement with the data pooled over participants, not the sum of the modal agreements with each individual's data). As goodness-of-fit improves, G^2 and SAEP decrease, and MA increases.

Table 4.1 Goodness-of-fit measures for logistic regression models fitted to monolingual Spanish speakers' vowel identification data.

Model	G^2	*df*	%SAEP	%MA
1. V	10345	3238	48.37	50.0
2. $V + \text{F1}\times V$	5785	3236	24.00	81.1
3. $V + \text{F1}\times V \quad + \text{dur}\times V$	5758	3234	23.94	82.2
4. $V + \text{F1}\times V + \Delta\text{F1}\times V$	3164	3232	5.56	97.8
5. $V + \text{F1}\times V + \Delta\text{F1}\times V + \text{dur}\times V$	3125	3230	5.11	96.7

Table 4.2 Quasi-likelihood *F* test for differences in goodness-of-fit between nested logistic regression models fitted to monolingual Spanish speakers' vowel identification data.

Models compared	Additional term	ΔG^2	heterogeneity	*F*	*df*	*p*
2 vs 1	$\text{F1}\times V$	4557	2.897	786.96	2, 3236	.000
3 vs 2	$\text{dur}\times V$	27	1.881	7.19	2, 3234	.001
4 vs 2	$\Delta\text{F1}\times V$	2626	1.881	348.91	4, 3232	.000
5 vs 3	$\Delta\text{F1}\times V$	2638	1.881	350.58	4, 3230	.000
5 vs 4	$\text{dur}\times V$	39	2.880	6.79	2, 3230	.001

For the monolingual Spanish participants' data, adding F1-tuning to the baseline model containing only biases resulted in a substantial improvement in goodness-of-fit. Adding ΔF1-tuning to a smaller model also resulted in a substantial improvement in goodness-of-fit. Adding duration-tuning to a smaller model resulted in a significant[5] but insubstantial improvement in goodness-of-fit (SAEP decreased by less than half a percentage point). This indicates that although monolingual Spanish listeners' L1 vowel identification was affected by spectral and VISC differences in the stimuli, the effect of duration was negligible; however, since prior research suggests that duration plays an important role in L2 perception, the full model (model 5) including duration will be used to model both L1 and L2 data. The coefficient values from model 5 fitted to the pooled monolingual Spanish perception data are given in Table 4.3.

A leave-one-participant-out analysis was conducted, obtaining the goodness-of-fit of each monolingual Spanish participant's raw data to a logistic regression model based on all the other monolingual Spanish participants' data, mean modal agreement was 86.9%.[6] Individual bilingual L1-Spanish speakers' raw L1 vowel identification data were compared for goodness-of-fit to the logistic regression model based on all the monolingual Spanish speakers' data, mean modal agreement was 79.3%. The nature of the difference between the monolingual and bilinguals' perception was qualitatively assessed via visual comparison of territorial maps based on the predicted probabilities from logistic regression models based on each groups' pooled vowel identification data (the internal goodness-of-fit-measures for the model fitted to the L1-Spanish bilinguals' data were SAEP 5.33%, MA 96.7%). The territorial maps are given in Figures 4.1 and 4.2, they divide the stimulus space into regions where different categories are predicted to be the most probable (or modal) response. For the monolingual Spanish group, the modal /ei/ response area was confined almost exclusively to the diverging-VISC stimulus subspace. For the L1-Spanish bilingual group, the modal /ei/ response area was larger, this was at the expense of the modal /e/ response area, shifting the /ei/–/e/ boundary towards higher F1 values, and also including part of the zero- and the converging-VISC stimulus subspaces.

[5] significant at $\alpha = 0.1$, equivalent to $\alpha = .05$ after a Bonferroni correction for five tests.

[6] MA is affected by boundary location only. SAEP can be affected by both boundary location and boundary crispness, SAEP will decrease if the individual's boundary is either crisper or fuzzier than the reference model. If all individuals have crisp boundaries but boundary location varies across individuals, then the model based on pooled data will derive an average boundary location but will have a fuzzier boundary than any of the individuals. Therefore, only MA will be reported for comparisons of individuals' data with pooled models.

Table 4.3 Estimated coefficient values from logistic regression model 5 fitted to pooled monolingual Spanish perception data.

Coefficient	Value	Standard Error	Wald χ^2	*df*	*p*
i	4.0098	0.1096	1338.74	1	.0000
ei	-0.3791	0.0870	18.97	1	.0000
e	-3.6307	0.1195	923.59	1	.0000
V			1375.63	2	.0000
F1×*i*	-0.9606	0.0239	1609.33	1	.0000
F1×*ei*	0.0226	0.0157	2.07	1	.1498
F1×*e*	0.9380	0.0230	1669.63	1	.0000
F1×*V*			1846.08	2	.0000
$\Delta F1_{+} \times i$	-1.3979	0.0807	300.15	1	.0000
$\Delta F1_{+} \times ei$	0.1582	0.0688	5.30	1	.0214
$\Delta F1_{+} \times e$	1.2396	0.0778	254.01	1	.0000
$\Delta F1_{+} \times V$			343.18	2	.0000
$\Delta F1_{-} \times i$	1.3778	0.0827	277.70	1	.0000
$\Delta F1_{-} \times ei$	1.4471	0.0652	491.95	1	.0000
$\Delta F1_{-} \times e$	-2.8249	0.0979	833.15	1	.0000
$\Delta F1_{-} \times V$			890.98	2	.0000
dur×*i*	-0.0458	0.0122	14.07	1	.0002
dur×*ei*	0.0587	0.0096	37.18	1	.0000
dur×*e*	-0.0128	0.0118	1.19	1	.2757
dur×*V*			38.61	2	.0000

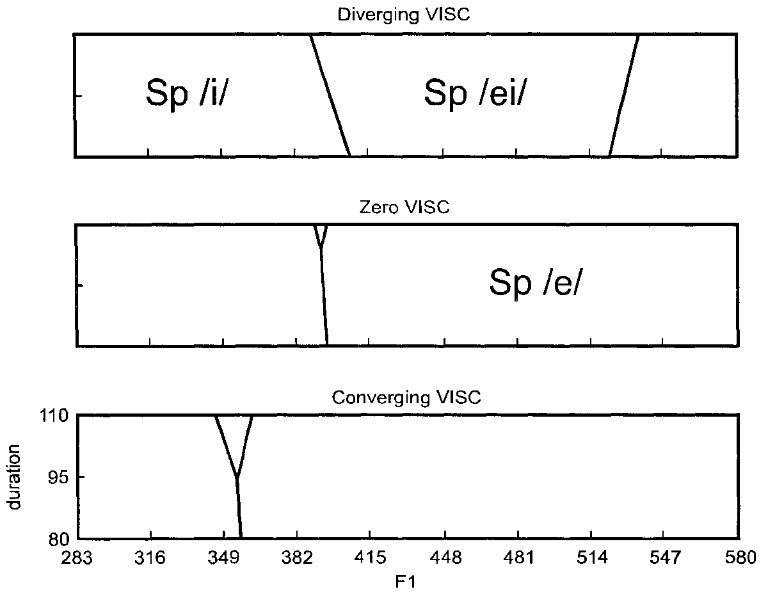


Figure 4.1 Territorial map based on classification of synthetic stimuli by the logistic regression model trained on monolingual Spanish listeners' L1 vowel identification data.

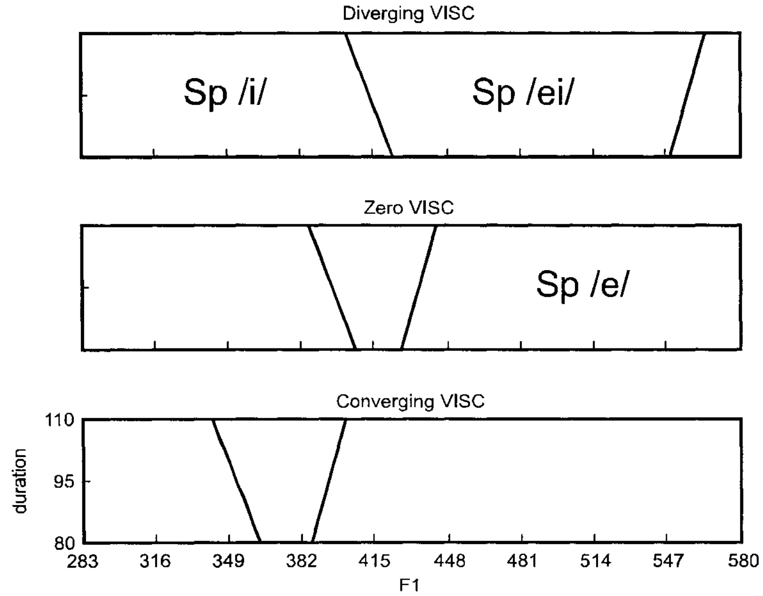


Figure 4.2 Territorial map based on classification of synthetic stimuli by the logistic regression model trained on bilingual L1-Spanish listeners' L1 vowel identification data.

4.1.2 L1-English speakers' perception

A series of nested models were fitted to the monolingual English speakers' pooled vowel identification data, the models and corresponding goodness-of-fit measures are given in Table 4.4, and quasi-likelihood *F* tests testing the improvement in goodness-of-fit are given in Table 4.5. Adding F1-tuning to the baseline model containing only the biases resulted in a substantial improvement in goodness-of-fit. Adding ΔF1-tuning to a smaller model also resulted in a substantial improvement in goodness-of-fit. Adding duration-tuning to a smaller model resulted in smaller improvements in goodness-of-fit compared to adding ΔF1-tuning, and adding duration-tuning to a model already including ΔF1-tuning did not result in a significant improvement in goodness-of-fit.[7] This indicates that although monolingual English listeners' L1 vowel identification was affected by spectral and VISC differences in the stimuli, the effect of duration was negligible. However, since L2-English listeners are expected to make use of duration, duration-tuning was included in the monolingual English perception model. The coefficient values from model 5 fitted to the pooled monolingual Spanish perception data are given in Table 4.6.

Table 4.4 Goodness-of-fit measures for logistic regression models fitted to monolingual English speakers' vowel identification data.

Model	G^2	*df*	%SAEP	%MA
1. V	12717	5127	52.70	33.3
2. $V + \text{F1}\times V$	6948	5124	27.43	73.3
3. $V + \text{F1}\times V \quad + \text{dur}\times V$	6589	5121	26.35	74.4
4. $V + \text{F1}\times V + \Delta\text{F1}\times V$	4417	5118	9.49	92.2
5. $V + \text{F1}\times V + \Delta\text{F1}\times V + \text{dur}\times V$	3994	5115	6.40	95.6

[7] However, an analysis of nested logistic regression models fitted to data pooled over monolingual and bilingual L1-English participants' L1 vowel identification data found a significant and substantial improvement in goodness-of-fit whenever duration-tuning was added.

Table 4.5 Quasi-likelihood *F* test for differences in goodness-of-fit between nested logistic regression models fitted to monolingual English speakers' vowel identification data.

Models compared	Additional term	ΔG^2	heterogeneity	*F*	*df*	*p*
2 vs 1	F1×*V*	5769	2.548	754.78	3, 5124	.000
3 vs 2	dur×*V*	359	1.659	72.12	3, 5121	.000
4 vs 2	ΔF1×*V*	2530	1.659	254.09	6, 5118	.000
5 vs 3	ΔF1×*V*	2594	1.584	259.73	6, 5115	.000
5 vs 4	dur×*V*	423	126.005	1.12	3, 5115	.340

A leave-one-participant-out cross-validation was conducted, obtaining the goodness-of-fit of each monolingual English participant's raw data to a logistic regression model based on all the other monolingual English participants' data, mean modal agreement was 83.4%. Individual bilingual L1-English speakers' raw L1 vowel identification data were compared for goodness-of-fit to the logistic regression model based on all the monolingual English speakers' data, mean modal agreement was 83.7%. Although mean modal agreements were almost identical, the bilingual's fit to the monolingual model could be skewed; therefore, potential differences between the monolingual English group and the L1-English bilingual group were qualitatively assessed via examination of territorial maps from logistic regression models based on each groups' pooled vowel identification data (the internal goodness-of-fit-measures for the model fitted to the L1-English bilinguals' data were SAEP 5.72%, MA 93.3%). No substantial differences were observed.

A territorial map for the monolingual English group is given in Figure 4.3. Although the goodness-of-fit tests on nested logistic regression models suggested that duration was negligible for the monolingual English group's perception, the territorial map indicates that stimuli with intermediate F1 were more likely to be given /ɪ/ responses if they were shorter, and /e/ responses if they were longer, and duration effects were also apparent on other vowel-pair boundaries. English /ɪ/ as the modal response was restricted almost exclusively to the converging-VISC subspace. English /ɪ/ is the predicted modal response in only a small portion of the zero-VISC subspace, but the size of the modal English /ɪ/ region would increase if the space were extrapolated to shorter vowel durations.

Table 4.6 Estimated coefficient values from logistic regression model 5 fitted to pooled monolingual English perception data.

Coefficient	Value	Standard Error	Wald χ^2	*df*	*p*
i	5.7263	0.1435	1593.08	1	.0000
ɪ	1.3306	0.1183	126.45	1	.0000
e	-1.0043	0.1071	87.87	1	.0000
ε	-6.0526	0.1723	1234.71	1	.0000
V			1764.19	3	.0000
F1×*i*	-1.1829	0.0295	1610.07	1	.0000
F1×*ɪ*	-0.1295	0.0186	48.61	1	.0000
F1×*e*	0.3326	0.0164	413.58	1	.0000
F1×*ε*	0.9798	0.0226	1877.25	1	.0000
F1×*V*			2139.32	3	.0000
ΔF1$_+$×*i*	-1.9269	0.0976	390.12	1	.0000
ΔF1$_+$×*ɪ*	0.8539	0.0717	141.69	1	.0000
ΔF1$_+$×*e*	-0.4499	0.0620	52.59	1	.0000
ΔF1$_+$×*ε*	1.5229	0.0806	357.33	1	.0000
ΔF1$_+$×*V*			733.39	3	.0000
ΔF1$_-$×*i*	1.8981	0.0965	387.25	1	.0000
ΔF1$_-$×*ɪ*	-0.4938	0.1082	20.82	1	.0000
ΔF1$_-$×*e*	0.4320	0.0644	44.97	1	.0000
ΔF1$_-$×*ε*	-1.8364	0.1045	308.95	1	.0000
ΔF1$_-$×*V*			552.84	3	.0000
dur×*i*	-0.0160	0.0151	1.13	1	.2888
dur×*ɪ*	-0.2070	0.0135	235.77	1	.0000
dur×*e*	0.1804	0.0107	286.12	1	.0000
dur×*ε*	0.0425	0.0142	9.02	1	.0027
dur×*V*			376.44	3	.0000

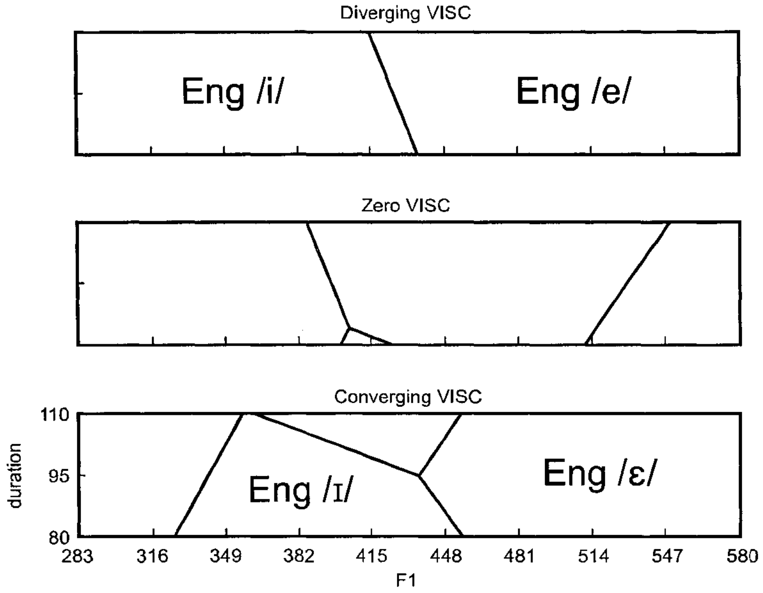


Figure 4.3 Territorial map based on classification of synthetic stimuli by the logistic regression model trained on monolingual English listeners' L1 vowel identification data.

4.2 Discussion

Logistic regression models fitted to L1-Spanish participants' pooled L1 vowel identification responses indicated that L1-Spanish listeners made little use of duration but did use initial formant values and VISC to distinguish Spanish vowels. Vowel identification shifted from /i/ to /ei/ to /e/ as F1 increased (and negatively correlated F2 decreased). Diverging-VISC perceptually distinguished the Spanish diphthong /ei/ from the monophthongs /i/ and /e/.

For monolingual Spanish listeners, /ei/ was the modal response in approximately a third of the diverging-VISC stimulus subspace, and was restricted almost exclusively to this subspace. For bilingual L1-Spanish participants /ei/ was the modal response in a larger portion of the stimulus space, occupying areas which were modally identified as /e/ by monolingual Spanish listeners, including portions of the zero- and converging-VISC subspaces. This behaviour could be the result of exposure to English /e/ which has a smaller magnitude of VISC than Spanish /ei/; note that in Section 3 the bilingual L1-Spanish speakers were also found to produce Spanish /ei/ with smaller VISC magnitude.[8] If bilingual listeners identify more vowels with smaller magnitudes of diverging VISC as /ei/, then more of the synthetic stimuli in the perception experiment with a fixed magnitude for diverging VISC (ΔF1 −99 Hz and ΔF2 +120 Hz) will also be identified as /ei/. The bilingual listeners' higher identification rates for /ei/ in the zero- and converging-VISC subspaces may also be the result of learning English: expanding Spanish /ei/ to become more English-/e/ like and forming an English /ɪ/ category may make stimuli in these regions of the stimulus space sound less Spanish /e/ like (see Flege, 1991).

Logistic regression models fitted to L1-Canadian-English participants' pooled L1 vowel identification responses indicated that L1-English listeners made substantial use of VISC to distinguish English vowels. English /e/ was the response with the highest predicted probability over the high-F1–low-F2 half of the diverging-VISC stimulus subspace, and English /ɪ/ as the modal response was restricted almost exclusively to the converging-VISC subspace. English /ɛ/ as the modal response also occurred predominantly in the converging-VISC subspace. The intermediate-F1–intermediate-F2 portion of the converging- and zero-VISC subspaces were also partially divided between /e/ and /ɪ/ on the basis of duration, longer stimuli being more likely to be identified as /e/ and shorter stimuli as /ɪ/. This indicates that converging VISC and short duration are important cues for English /ɪ/ perception. The existence of a substantial area with English /e/ as the modal predicted response in the converging-VISC subspace is unexpected since English /e/ is produced with diverging VISC.

4.3 Comparison of monolingual Spanish and monolingual English perception

4.3.1 Comparisons

A qualitative comparison of the territorial maps based on the monolingual Spanish and monolingual English groups' L1 identification data (Figures 4.1 and 4.3) reveals several similarities and differences in monolingual English and monolingual Spanish perception of the synthetic stimuli:

[8] Although since most monolinguals spoke Peninsular dialects and most bilinguals spoke American dialects, the differences between monolingual and bilingual groups in both production and perception could be due to dialect differences.

– In the diverging-VISC subspace, English has two categories as the modal response and Spanish three. The English /i/–/e/ boundary and the Spanish /i/–/ei/ boundary are at approximately the same F1 value, approximately 425 vs 400 Hz. English /e/ extends to the high-F1 edge of the diverging-VISC subspace, but Spanish has an /ei/–/e/ boundary at an F1 of around 530 Hz.

– In the zero-VISC subspace, English has four categories as the modal response and Spanish two. Spanish has an /i/–/e/ boundary at approximately the same F1 value as the English /i/–/e+ɪ/ boundary (at approximately 400 Hz). Spanish /e/ extends to the high-F1 edge of the zero-VISC subspace, but English has an /e/–/ɛ/ boundary at an F1 running from approximately 510 Hz for a vowel duration of 80 ms to approximately 550 Hz for a vowel duration of 110 ms. English /ɪ/ is the predicted modal response in only a small portion of the zero-VISC subspace.

– In the converging-VISC subspace, English has four categories as the modal response and Spanish two. English /ɪ/ occurs almost exclusively in the converging-VISC subspace, the English /i/–/ɪ/ boundary is in approximately the same location as the Spanish /i/–/e/ boundary at around 355 Hz, but, unlike the Spanish boundary, the English boundary is subject to a duration effect and runs from approximately 325 Hz at 80 ms vowel duration to 355 Hz at 110 ms vowel duration. The region of English /ɪ/ modal perception therefore has some overlap with the region of Spanish /i/ modal perception, but falls predominantly within the region of Spanish /e/ modal perception.

Table 4.7 gives a confusion matrix for the classification of each of the 90 synthetic stimuli on the basis of the logistic regression model trained on the monolingual Spanish versus the logistic regression model trained on the monolingual English listeners' L1 vowel identification data. For each L1 model, stimuli were crisply classified according to which of the vowel categories had the highest predicted probability. The confusion matrix indicates the number of stimuli which were classified as the Spanish vowel category indicated on the row and the English vowel category indicated on the column; for example, 30 stimuli were classified both as Spanish /i/ and as English /i/, and 2 stimuli were classified both as Spanish /i/ and as English /ɪ/.

Table 4.7 Confusion matrix for the classification of the 90 synthetic stimuli on the basis of logistic regression models trained on monolingual Spanish and monolingual English listeners' L1 vowel identification data. Blank cells have values of zero.

	Eng /i/	Eng /ɪ/	Eng /e/	Eng /ɛ/	sum
Sp /i/	30	2			32
Sp /ei/	3		10		13
Sp /e/		6	20	19	45
sum	33	8	30	19	90

4.3.2 Predictions

On the basis of the comparisons of the monolingual Spanish and monolingual English perception models, predictions can be made as to how L2 learners at the initial state for L2 learning will categorise the synthetic stimuli in terms of L2 categories. These predictions are based on the assumption that the listeners will assimilate the stimuli to L1 categories, and then reuse the L1 boundaries as L2 boundaries (see similar and subspace scenarios in Section 1.2.2). New boundaries will not yet have developed but the listeners may make use of category goodness to distinguish within-L1-category differences.

The following predictions are made for L1-Spanish listeners just beginning to learn English. A territorial map of the predictions is given in Figure 4.4.

– Most of the stimuli which monolingual English listeners identified as English /i/ were identified as Spanish /i/ by monolingual Spanish listeners. Beginning L2-English listeners are therefore predicted to reuse their Spanish /i/ category to identify stimuli as English /i/. This will result in slight mismatches between the L2-English /i/–/e/ and /i/–/ɪ/ boundaries (reused L1-Spanish /i/–/ei/ and /i/–/e/ boundaries) and those of L1-English listeners.

– Most of the stimuli identified as Spanish /ei/ by monolingual Spanish listeners were identified as English /e/ by monolingual English listeners. Beginning L2-English listeners are therefore predicted to reuse their Spanish /ei/ category to identify stimuli as English /e/. However, only a third of the stimuli which monolingual English listeners identified as English /e/ were identified as Spanish /ei/ by monolingual Spanish listeners. There will therefore be a large mismatch between L2-English and L1-English /e/ perception; in particular, L2-English listeners will fail to identify zero-VISC stimuli as English /e/.

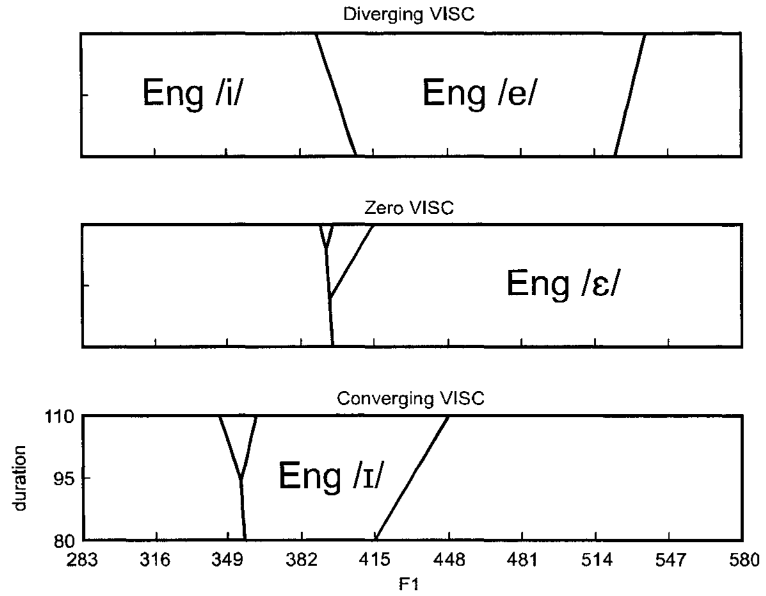


Figure 4.4 Territorial map based on the predictions of the initial state of L2-English learning for an L1-Spanish listener. Compare with L1-Spanish and L1-English territorial maps, Figures 4.1 and 4.3. The L2-English /ɪ/–/ɛ/ boundary represents a best-case scenario and is predicted to be fuzzier than the L1-English /ɪ/–/ɛ/ boundary.

– Stimuli identified as Spanish /e/ by monolingual Spanish listeners were identified as English /ɪ/, /e/, and /ɛ/ by monolingual English listeners. Beginning L2-English listeners are therefore predicted to reuse their Spanish /e/ category to identify stimuli as English /ɪ/ and /ɛ/ (but not English /e/ because Spanish /ei/ is a better match since most of the stimuli identified as Spanish /ei/ by monolingual Spanish listeners were identified as English /e/ by monolingual English listeners). Zero-VISC Stimuli which L1-English listeners would identify as English /e/ will be identified as English /ɪ/ or /ɛ/ by L2-English listeners. Since no existing L1 boundary can be reused, any L2 boundary between English /ɪ/ and /ɛ/ is expected to be relatively fuzzy. A possibility is that English /ɪ/ and /ɛ/ could be distinguished via a category-goodness-difference assimilation to Spanish /e/, in which case more L1-Spanish-/e/-like stimuli (shorter zero-VISC stimuli with F1 around 485 Hz, compare mean male production values in Table 3.1 with the range of values for male-voice synthetic stimuli identified as Spanish /e/ in Figure 4.1) may have a higher probability of being identified as one of the two English vowels, and more L1-Spanish-/e/-like stimuli (longer converging-VISC with F1 higher or lower than 485 Hz) might have a probability of being identified as the other English vowel. The L2-English /ɪ/–/ɛ/ boundary in Figure 4.4 represents a best-case scenario.

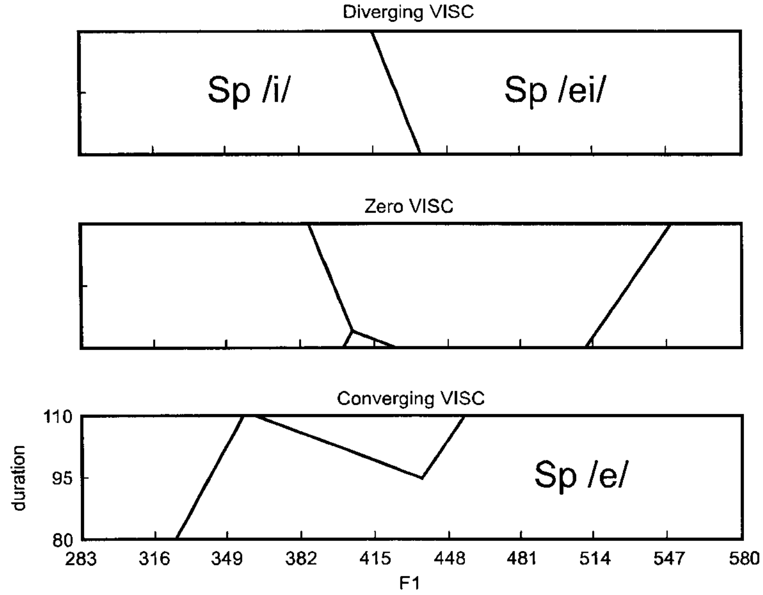


Figure 4.5 Territorial map based on the predictions of the initial state of L2-Spanish learning for an L1-English listener. Compare with L1-English and L1-Spanish territorial maps, Figures 4.3 and 4.1.

The following predictions are made for L1-English listeners just beginning to learn Spanish. A territorial map of the predictions is given in Figure 4.5.

– Most of the stimuli which monolingual Spanish listeners identified as Spanish /i/ were identified as English /i/ by monolingual English listeners. Beginning L2-Spanish listeners are therefore predicted to reuse their English /i/ category to identify stimuli as Spanish /i/. This will result in slight mismatches between the L2-Spanish /i/–/ei/ and /i/–/e/ boundaries (reused L1-English /i/–/e/ and /i/–/ɪ/ boundaries) and those of L1-Spanish listeners.

– Most of the stimuli which monolingual Spanish listeners identified as Spanish /ei/ were identified as English /e/ by monolingual English listeners. Beginning L2-Spanish listeners are therefore predicted to reuse their English /e/ category to identify stimuli as Spanish /ei/. There will therefore be a large mismatch between L2-Spanish and L1-Spanish

/ei/ perception; in particular, L2-Spanish listeners will identify zero-VISC stimuli as Spanish /ei/ rather than Spanish /e/.

– Most stimuli which monolingual English listeners identified as English /ɪ/ and /ɛ/ were identified as Spanish /e/ by monolingual Spanish listeners. Beginning L2-Spanish listeners are therefore predicted to reuse their English /ɪ/ and /ɛ/ categories to identify stimuli as Spanish /e/, giving the Spanish /e/ label to any stimuli which they perceive to be either English /ɪ/ or /ɛ/. This will result in slight mismatches between the L2-Spanish /i/–/e/ boundary (reused L1-English /i/–/ɪ/ boundary) and that of L1-Spanish listeners. There will also be a large mismatch between the L2-Spanish and L1-Spanish /ei/–/e/ boundary; in particular, L2-Spanish listeners will identify zero-VISC stimuli as Spanish /ei/ rather than as Spanish /e/.

5. Comparisons of L1 Production and Perception Models

Similar production and perception model results would be expected under the theory that L1 learners base their vowel perception on the multivariate distribution of the acoustic properties of the L1 vowels to which they are exposed, and that L1 speakers base their own vowel productions on their perception-based categories, and hence the acoustic properties of their vowel productions are representative of the acoustic properties of the vowels upon which they built their perception system.

5.1 Classification of Synthetic Stimuli by L1 Production Models

The formant values of the synthetic stimuli fell along a diagonal in the F1–F2 space which roughly corresponds to the traditional vowel height dimension for front vowels. VISC in the synthetic stimuli was restricted to movement along the same vowel height dimension.[1] The perception models were therefore built on a more restrictive acoustic space than is the case for the production models, which, being based on natural vowel productions, did not have one hundred percent correlation between F1 and F2 in initial formant values and had more variability in VISC direction and magnitude. Probing the perception models using the production data would require making assumptions as to the appropriate projection of the higher-dimension properties of the natural vowel productions onto the more restricted stimulus space examined in the perception experiment. Probing the production models using the more restricted properties of the synthetic data used in the perception experiment does not require such assumptions to be made. The following procedure was used to probe the production models using the acoustic properties of the synthetic stimuli:

– The formant values at 25 and 75% of the durations of the synthetic stimuli were

[1] The size of the synthetic stimulus space was kept small so as not to overtax participants, the number of stimuli increases exponentially as additional dimensions are added. Had additional dimensions been added so that F1 and F2, and ΔF1 and ΔF2 were not always correlated, then the natural productions could have been classified directly by the logistic regression model trained on perception data.

calculated, and converted to the log Hertz values for F1, F2, ΔF1, and ΔF2. The vowel durations of the synthetic stimuli were converted to log milliseconds.

  - The formant values at 25 and 75% of the synthetic vowel durations fell on the same F1–F2 diagonal as the synthetic stimuli's initial and final formant values (specified at inflection points 10 ms from the edges of the vowels) but included some systematic offsets along that diagonal.

- The F1, F2, ΔF1, ΔF2, and duration values for each stimulus were converted into discriminant-function-variable values using the unstandardised canonical discriminant function coefficient values derived in Section 3.
  - For the L1-Spanish model the coefficient values are given in Table 3.5.
  - For the L1-English model the coefficient values are given in Table 3.9.
- On the basis of its discriminant-function-variable values, each stimulus was classified (crisp classification) as one of the L1 vowels using the classifier trained on L1 vowel productions.
  - The quadratic classifier trained on L1-Spanish vowels was used to classify synthetic stimuli in terms of L1-Spanish vowel categories.
  - The linear classifier trained on L1-English vowels was used to classify synthetic stimuli in terms of L1-English vowel categories.

The position of the probes in the Function 1 – Function 2 space defined by the CDFA trained on L1-Spanish speakers' productions is given in Figure 5.1, and the territorial map for the classification of the probes on the basis of the CDFA is given in Figure 5.2. The production and perception models are generally very similar. Comparing Figure 5.2 with Figure 4.1 (the territorial map based on the logistic regression analysis of monolingual Spanish listeners' L1 vowel identification data), the Spanish /i/–/e/ boundaries are in approximately the same locations, but the Spanish /i/–/ei/ and /ei/–/e/ boundaries indicate a greater role for duration-tuning in the production model. The modal agreement between the two models was 90.0%, a classification confusion matrix is given in Table 5.1.

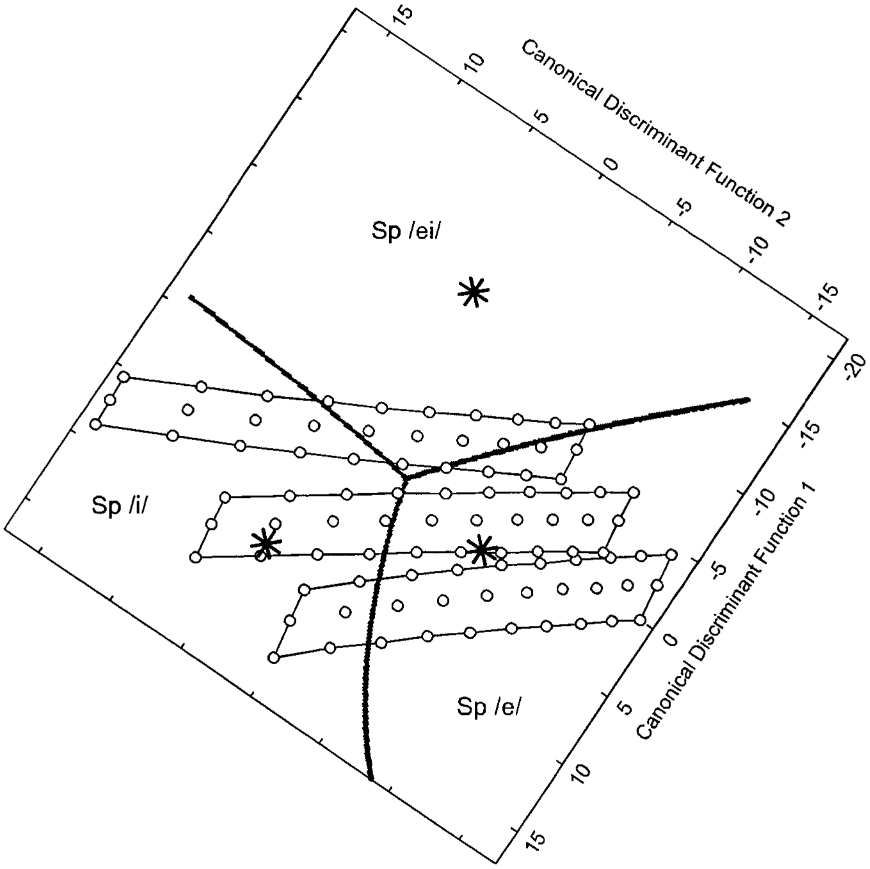


Figure 5.1 Location of synthetic stimuli in Function 1 – Function 2 space of the CDFA trained on L1-Spanish speakers' vowel production data. The axes have been rotated so that the synthetic stimuli have approximately the same orientation as in Figure 5.2. Stars represent centroids of L1-Spanish production data.

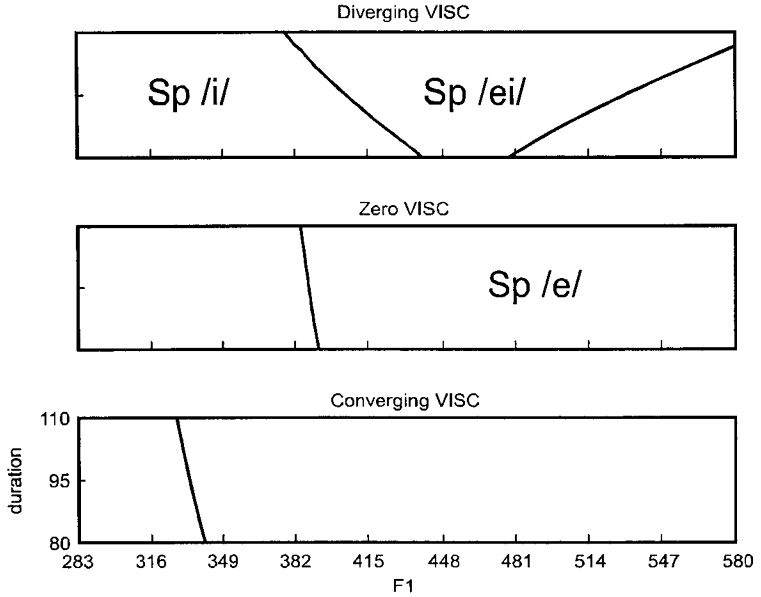


Figure 5.2 Territorial map based on classification of synthetic stimuli by the CDFA trained on L1-Spanish speakers' vowel production data. Compare with Figure 4.1.

Table 5.1 Confusion matrix for the classification of the 90 synthetic stimuli on the basis of the quadratic CDFA trained on L1-Spanish speakers' production data and the logistic regression model trained on monolingual Spanish listeners' perception data. Blank cells have values of zero.

CDFA Production Model	Logistic Regression Perception Model			
	Sp /i/	Sp /ei/	Sp /e/	sum
Sp /i/	29	1		30
Sp /ei/	1	9	2	12
Sp /e/	2	3	43	48
sum	32	13	45	90

The position of the probes in the Function 1 – Function 2 space defined by the CDFA trained on L1-English speakers' productions is given in Figure 5.3, and the territorial map for the classification of the probes on the basis of the CDFA is given in Figure 5.4. Although synthetic stimuli with different VISCs overlap in the Function 1 – Function 2 plot in Figure 4.6, they do not intersect in the Function 1 – Function 2 – Function 3 space. The production and perception models are generally similar. Comparing Figure 5.4 with Figure 4.3 (the territorial map based on the logistic regression analysis of monolingual English listeners' L1 vowel identification data), the English /i/–/e/ boundary had somewhat lower F1 values, approximately 380 Hz in the production model compared to approximately 425 Hz in the perception model in the diverging VISC subspace, and the English /i/–/ɪ/ and /e/–/ɪ/ boundaries had slightly higher F1 values in the production model. The modal agreement between the CDFA and the logistic regression model was 87.8%, a classification confusion matrix is given in Table 5.2.

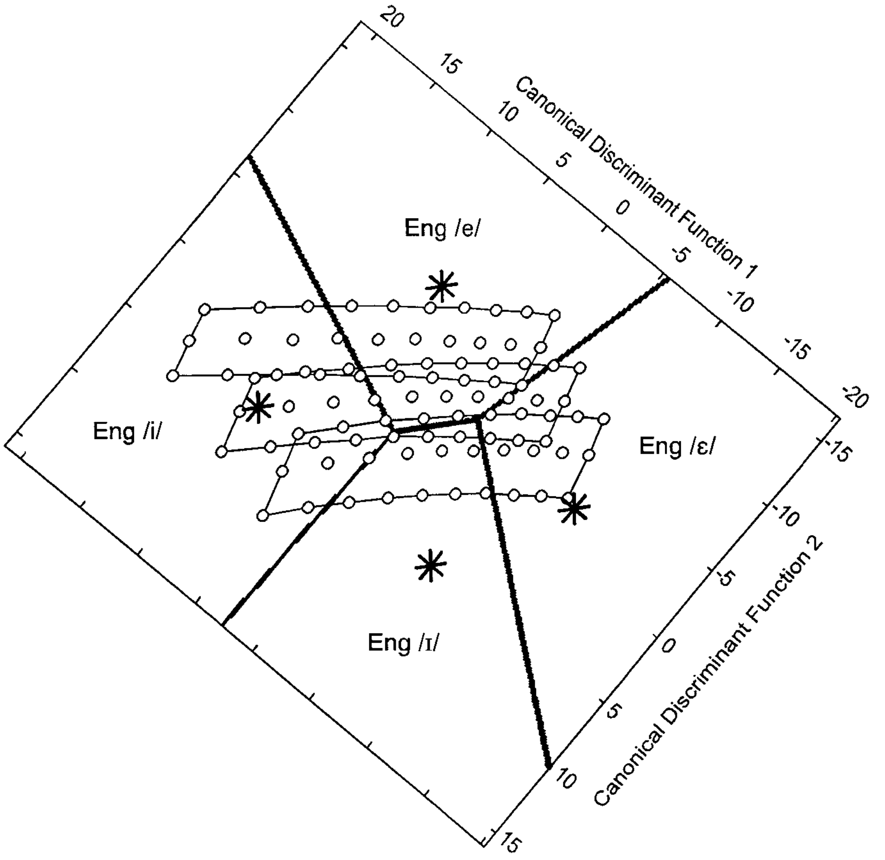


Figure 5.3 Location of synthetic stimulus values in Function 1 – Function 2 space of the CDFA trained on L1-English speakers' vowel production data. The axes have been rotated so that the synthetic stimuli have approximately the same orientation as in Figure 5.4. Stars represent centroids of L1-English production data.

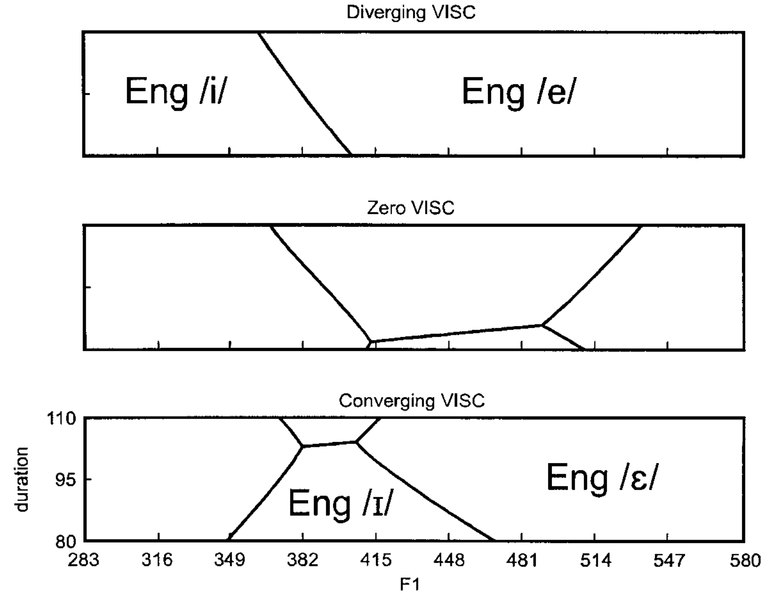


Figure 5.4 Territorial map based on classification of synthetic stimulus values by the CDFA trained on L1-English speakers' vowel production data. Compare with Figure 4.3.

Table 5.2 Confusion matrix for the classification of the 90 synthetic stimuli on the basis of the linear CDFA trained on L1-English speakers' production data and the logistic regression model trained on monolingual English listeners' perception data. Blank cells have values of zero.

CDFA Production Model	Logistic Regression Perception Model				
	Eng /i/	Eng /ɪ/	Eng /e/	Eng /ɛ/	sum
Eng /i/	28	1			29
Eng /ɪ/		7	2		9
Eng /e/	5		25		30
Eng /ɛ/			3	19	22
sum	33	8	30	19	90

Although there were some discrepancies, the production and perception models were generally similar, supporting the theory that L1 learners base their vowel perception on the multivariate distribution of the acoustic properties of the L1 vowels to which they are exposed, and that L1 speakers base their own vowel productions on their perception-based categories.

5.2 Bias in Synthetic Stimuli Relative to Natural Productions

Assuming that L1 perception is highly correlated with L1 production and that the synthetic stimulus space is an unbiassed representation of L1 production, the L1 perception results could be used to make predictions as to the assimilation of L2 vowels to L1 categories. For example, under these assumptions, L1-English listeners' perception of the synthetic stimuli which L1-Spanish listeners perceive as Spanish /i/, will be representative of L1-English listeners' perception of L1-Spanish speakers' /i/ productions; however, the assumption that the synthetic stimuli are an unbiassed representation of the properties of L1-Spanish and L1-English vowel productions is invalid.[2]

There are two types of bias in the synthetic stimuli, over-representation and under-

[2] Had the production results been available when the perception stimuli were designed, it might have been possible to avoid the bias in the synthetic stimuli. Practical considerations led to the production and perception data being collected during the same time period.

representation of the acoustic properties of natural vowel productions. Over-representation bias will be discussed first. Because the stimuli were designed to cover the acoustic properties for vowels in both languages, they include stimuli which are representative of productions in Spanish but not in English, and vice versa. For example, instances of vowels with F1 around 448 Hz and zero VISC are relatively rare for English but common for Spanish /e/, and instances of short vowels with high F1 and converging VISC are rare for any Spanish vowel but common for English /ɛ/. The factorial design also introduced stimuli which are atypical for vowels in either language, for example, instances of vowels with low F1 and diverging VISC are not common for any Spanish or English vowel category.

A procedure to remove over-representation bias was explored. Essentially, the classification results of a logistic regression model based on L1 perception data, are weighted according to the relative frequency of occurrence of natural vowels with acoustic properties in the vicinity of those of each of the synthetic stimuli:

– A logistic regression model is fitted to L1 vowel identification results, and the model's a posteriori predicted probabilities for each vowel category for each stimulus are calculated:

$APP(v,s)$

where *v* is the index of a vowel category in the set of vowel categories in the perception experiment, and *s* is the index of a stimulus in the set of synthetic stimuli in the perception experiment

– The synthetic stimuli are projected into an L1-production-trained canonical discriminant function space using the procedure described in Section 5.1, and the class-conditional probability density function values for each vowel category in the L1 production data at each stimulus point are calculated:

$PDF(CDF(s)|u)$

where *u* is the index of a vowel category in the set of vowel categories in the production experiment, and *CDF(s)* is the projection of the acoustic properties of stimulus *s* in to the canonical discriminant function space

– Each cell in the confusion matrix *CMX* is calculated as:

$CMX(u,v) = \sum_S APP(v,s) \times PDF(CDF(s)|u)$

– Each *u* row in the confusion matrix is normalised to sum to 100, simulating the percentage classification of natural stimuli falling within the synthetic stimulus space.

The APPs and PDFs can be based on perception and production experiments on the same L1, for example, an L1-Spanish-production-weighted L1-Spanish perception model, or on different L1s, for example, an L1-English-production-weighted L1-Spanish perception model.

Under-representation bias will now be discussed. Examination of PDFs projected onto the synthetic stimulus space, and the locations of the synthetic stimuli relative to the category means (centroids) in the CDFA Function 1 – Function 2 plots in Figures 5.1 and 5.3, reveal the under-representation bias, which is clearest in the cases of Spanish /ei/ and English /ɪ/: Spanish /ei/ and English /ɪ/ did not have very high PDF values in any part of the synthetic stimulus space, and the synthetic stimulus space did not cover and did not come close to covering the centroids of these vowel categories in the canonical discriminant function spaces. Figure 5.5 shows the relationship between the synthetic stimuli and the male speakers' mean production values (which are the same as the means of the normalised vowels). Spanish /ei/ was longer and had greater VISC magnitude than any of the synthetic stimuli, and English /ɪ/ was shorter and had lower F1 and F2 values than any of the synthetic stimuli.

Production-weighting the perception model would account for the over-representation bias and allow for predictions to be made regarding the perception of natural vowels in the vicinity of the synthetic stimuli, but such predictions would still be biassed because this part of the vowel space is under-representative of acoustic properties of natural productions of Spanish /i/, /ei/, /e/ and English /i/, /ɪ/, /e/, /ɛ/. Since no mechanism is available to ameliorate the under-representation bias, the perception data will not be used to make predictions as to L2 vowel perception in general. In Section 4.3.2, predictions were made as to the perception of the synthetic stimuli in terms of L2 vowel categories by L1-Spanish listeners just beginning to learn English and L1-English listeners just beginning to learn Spanish. Section 6 will compare the L1-English production model's classification of natural vowels with monolingual L1-English listeners' perception of natural vowels.

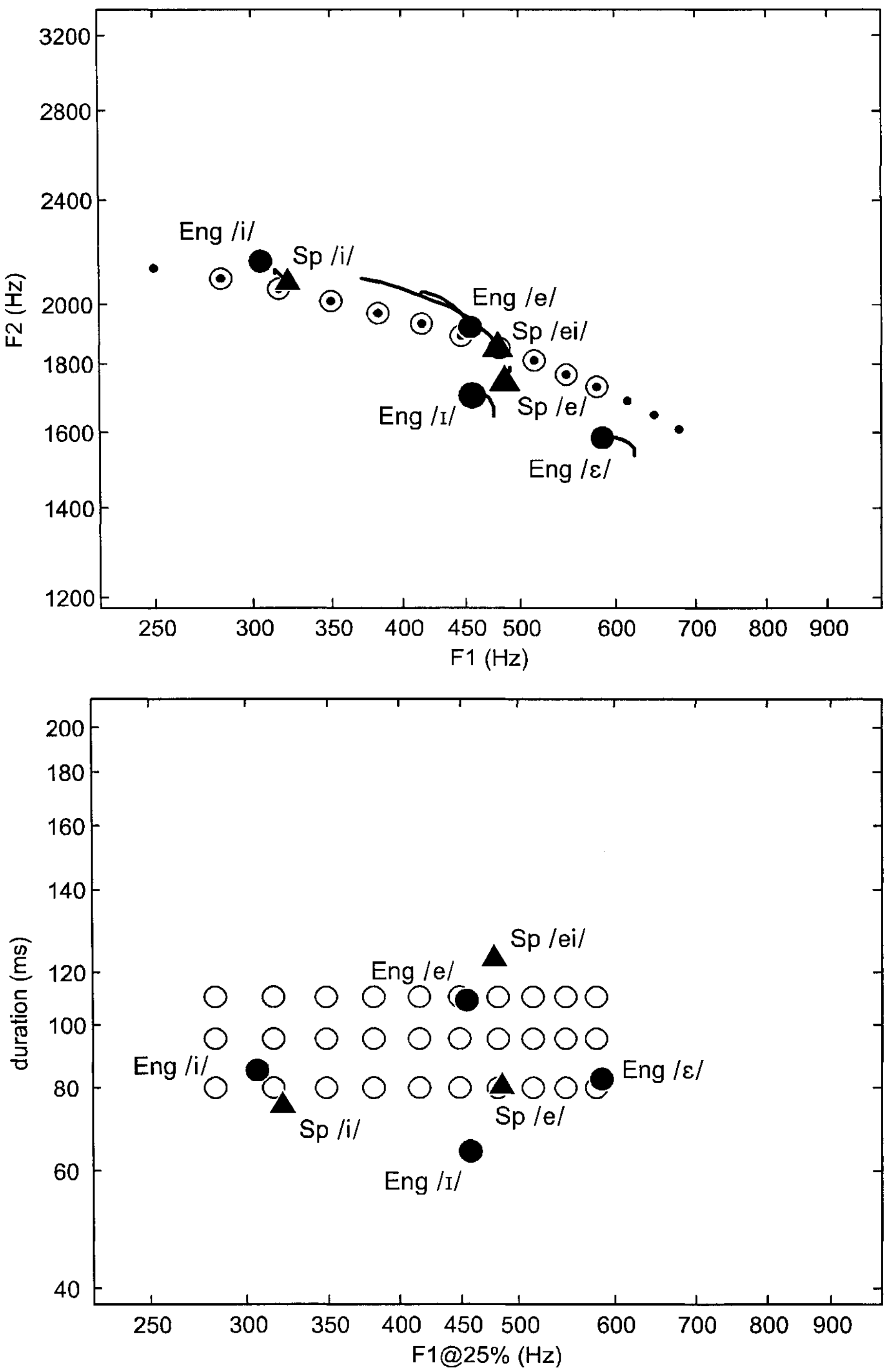


Figure 5.5 Mean acoustic properties of male speakers' L1-Spanish and L1-English vowels. Top: F1, F1, ΔF1, and ΔF2. Comet heads indicate formant values at 25% of the duration of the vowel, ends of comet tails indicate formant values at 75% of the duration of the vowel. White circles indicate the initial formant values and black dots the final formant values of the synthetic stimuli (nominally 10 ms from vowel edges). Final formant values are in the same location as the initial values, three dots to the left, and three dots to the right. Bottom: F1 at 25% of the duration of the vowel and vowel duration. White circles indicate the initial F1 values and durations of the synthetic stimuli.

6. Natural Vowel Perception

To assess whether the L1 Production models were good predictors of L1 perception, a follow-up experiment was conducted in which four of the monolingual English participants (me098, me099, me107, me118) identified natural vowel productions. They identified natural L1-English vowels produced by ten monolingual English speakers,[1] natural L1-Spanish vowels produced by ten monolingual Spanish speakers,[2] and natural L1-Spanish and L2-English vowels produced by eighteen bilingual L1-Spanish speakers.[3] Three productions of each vowel category were randomly selected from each speaker. The responses to the L1-English and L1-Spanish productions will be discussed here.

The procedures for the natural stimulus perception experiment were the same as the synthetic stimulus experiment, except that no carrier sentence was included, there was a single trial per stimulus, listeners could listen to each stimulus up to three times, and there was a 750 ms pause following a listener's response. Stimuli were presented in random order blocked by speaker. To allow the listeners to adapt to each new voice, a single stimulus was randomly selected for each speaker and played prior to the randomised block of all stimuli for that speaker; the response to this first stimulus was discarded.

6.1 L1-English Listeners' Perception of Natural L1-English Vowels

Vowel identification results for L1-English productions are presented as a confusion matrix in Table 6.1. Overall percent correct identification was 97.1%. Table 6.2 gives the confusion matrix for the classification of the same set of natural vowel productions by the

[1] five male speakers (me106, me109, me113, me115, me117) and five females speakers (me096, me099, me100, me101, me111) selected at random

[2] five male speakers (ms038, ms031, ms037, ms032, ms033) and five female speakers (ms045, ms046, ms039, ms041, ms043) selected at random

[3] eight male speakers (ms071, ms087, ms083, ms016, ms023, ms078, ms059, ms077) and ten female speakers (ms051, ms052, ms072, ms058, ms028, ms019, ms067, ms086, ms063, ms057), selected to exemplify a range of L2-English vowel production patterns

Table 6.1 Confusion matrix for four monolingual English listeners' identification of natural-vowels produced by ten monolingual English speakers (4 response to each of 30 instances of each vowel category). Results pooled over speakers and listeners, and expressed as percentages summing to 100 along each row, blank cells have values of zero.

Produced	Perceived			
	Eng /i/	Eng /ɪ/	Eng /e/	Eng /ɛ/
Eng /i/	95.8	4.2		
Eng /ɪ/		100.0		
Eng /e/		2.5	97.5	
Eng /ɛ/	1.7	2.5	0.8	95.0

Table 6.2 Confusion matrix for the classification of the of natural-vowels produced by ten monolingual English speakers (30 instances of each vowel category) by the CDFA model trained on all L1-English productions. Values are expressed as percentages summing to 100 along each row, blank cells have values of zero.

Produced	Perceived			
	Eng /i/	Eng /ɪ/	Eng /e/	Eng /ɛ/
Eng /i/	100.0			
Eng /ɪ/		93.3		6.7
Eng /e/			100.0	
Eng /ɛ/				100.0

L1-English production model (Section 3.3). The correlation between the CDFA model's a posteriori probability predictions for each vowel category for each stimulus and the listeners' pooled proportion of responses for each vowel category for each stimulus was .964.[4] Because both the model's predictions and the listeners' responses had very high correct classification rates, a high correlation between the two is to be expected. A fairer assessment of the correlation was obtained by measuring the correlation between the model and the listeners on all the 588 stimuli from the natural vowel perception experiment, including L1-Spanish

[4] The procedure for calculating this correlation coefficient is described in Nearey & Assmann (1986, appendix), the four listeners' response sets (one response per stimulus) were pooled into a single proportional response set. A residual-degrees-of-freedom correction factor was applied (Andruski & Nearey, 1992, note 14).

and L2-English vowels, in this case the correlation coefficient was .889 ($p < .05$).[5] Although the confusion matrices indicate some differences between the listeners' responses and the L1 production model, in general the CDFA model trained on L1-Speakers productions was highly correlated with the L1-English listeners' perception.

6.2 L1-English Listeners' Perception of Natural L1-Spanish Vowels

Vowel identification results for L1-Spanish productions are presented as confusion matrices in Table 6.3, identification by monolingual English listeners, and in Table 6.4, classification by the CDFA trained on L1-English productions (Section 3.3). The correlation between the CDFA model's a posteriori probability predictions for each vowel category for each stimulus and the listeners' pooled proportion of responses for each vowel category for each stimulus was .863. In general, the predictions made on the basis of the L1 production models (Section 3.5) were borne out:

– Almost all instances of Spanish /i/ were assimilated to English /i/.

   – The production model predicted an assimilation rate of 99.3% and did not predict the secondary assimilation to English /ɪ/ observed in the listeners' responses.

– Almost all instances of Spanish /ei/ were assimilated to English /e/.

   – The production model predicted a 100% rate of assimilation to English /e/, close to the 99.1% observed.

– Some instances of Spanish /e/ were assimilated to English /ɪ/, some to English /e/, and some to English /ɛ/

   – although the production model predicated assimilation to these three English categories, it did not match the relative proportions observed or even the rank order

      — observed rank order /ɪ/ > /ɛ/ > /e/,

      — production model rank order /ɪ/ > /e/ > /ɛ/

[5] The significance level was obtained via a randomisation test in which the listeners' responses to each stimulus were randomly permuted ten-thousand times, the correlation with the models' a posteriori predictions obtained each time, and a count taken of the number of times that the result exceeded the correlation calculated on the original non-permuted data. The actual count was zero.

Table 6.3 Confusion matrix for four monolingual English listeners' identification of natural-vowels produced by twenty-eight L1-Spanish speakers (4 response to each of 84 instances of each vowel category). Results pooled over speakers and listeners, and expressed as percentages summing to 100 along each row, blank cells have values of zero.

Produced	Perceived			
	Eng /i/	Eng /ɪ/	Eng /e/	Eng /ɛ/
Sp /i/	92.3	7.1	0.6	
Sp /ei/	0.3	0.6	99.1	
Sp /e/		43.8	24.7	31.5

Table 6.4 Confusion matrix for the classification by the CDFA model trained on all L1-English productions of the natural Spanish vowels produced by twenty-eight L1-Spanish speakers. Values are expressed as percentages summing to 100 along each row, blank cells have values of zero.

Produced	Perceived			
	Eng /i/	Eng /ɪ/	Eng /e/	Eng /ɛ/
Sp /i/	100.0			
Sp /ei/			100.0	
Sp /e/	1.2	58.3	29.8	10.7

7. Predictions Beyond the Initial State for L2 Learning

In Section 3 predictions were made as to the initial state for L2 learning based on statistical models of L1 production data. In Section 6, a reasonably high correlation was found between these predictions and the observed assimilation of Spanish vowels by monolingual English listeners. In Section 7, predictions will be made as to the behaviour of L2 learners beyond the initial state for L2 learning. Predictions will be made on the basis of two models: 1. a single production model based on the distribution of acoustic properties of both L1-English and L1-Spanish vowel productions (the mega-model), and 2. perception models based on L1-English and L1-Spanish listeners' L1 vowel category-goodness ratings for the synthetic stimuli. These models will provide information on the degree of overlap between L1 and L2 categories, and the likelihood that L2 vowels will be perceived as poor members of the L1 category to which they are assimilated. This is the sort of information required by Flege's SLM, in order to make a priori predictions as to how the L2 speech sounds will be learnt. The discussion below will be couched in terms of the distribution-based interpretation of the SLM presented in Section 1.2.2, the reader may find it helpful to refamiliarise themselves with Section 1.2.2 at this point.

7.1 L1 Production Mega-Model

A method for making predictions of cross-language vowel similarity was proposed by Thomson (2005). In the present context, this is a single CDFA trained on both L1-English and L1-Spanish vowel productions. The logic of this *mega-model* is that if an L1 and an L2 vowel category are very similar, then a large proportion of instances of the L2 category will be misclassified as the L1 category and vice versa, but if the L2 vowel category is very dissimilar from any L1 category, then instances of that category will be correctly classified at rates approaching 100% and very few instances of L1 vowels will be classified as that L2 category.

A single CDFA was fitted to the L1-English and L1-Spanish speakers' L1 acoustic vowel production data. Summary statistics from the derivation of the canonical discriminant functions are given in Table 7.1, unstandardised coefficients and total structure coefficients for the first three functions are given in Table 7.2, and a plot of the data transformed by the first and second functions is given in Figure 7.1. The fourth and fifth functions accounted for less than 1% of the variance and were not used for classification. A confusion matrix for vowel classification using a quadratic classifier is given in Table 7.3 (results using a linear classifier were very similar).

Table 7.1 Summary statistics from derivation of canonical discriminant functions for mega-model trained on both L1-English and L1-Spanish vowels (Wilks's Λ before the corresponding function was derived). Significance levels in χ^2 tests are unlikely to be accurate because of heterogeneity in the data due to pooling across speakers.

Function	Eigen values	Relative percentage	Canonical correlation	Wilks's Λ	χ^2	*df*	*p*
1	9.870	55.7	.953	.007	18572.427	30	.000
2	7.313	41.3	.938	.076	9641.486	20	.000
3	0.363	2.1	.516	.632	1714.628	12	.000
4	0.150	0.8	.361	.862	554.853	6	.000
5	0.009	0.1	.094	.991	33.338	2	.000

Table 7.2 First three unstandardised canonical discriminant function coefficients and total structure coefficients from the mega-modal CDFA trained on both L1-English and L1-Spanish vowels.

Original variables	Unstandardised coefficients			Total structure coefficients		
	Function 1	Function 2	Function 3	Function 1	Function 2	Function 3
Constant	13.590	-442.124	-334.363			
F1	21.758	42.156	6.744	.662	.740	-.110
ΔF1	21.397	--9.152	33.661	.686	-.595	.293
F2	-15.224	13.530	25.863	-.799	-.441	.265
ΔF2	-50.393	36.725	-37.229	-.731	.552	-.245
duration	-6.365	18.718	22.574	-.553	.742	.332

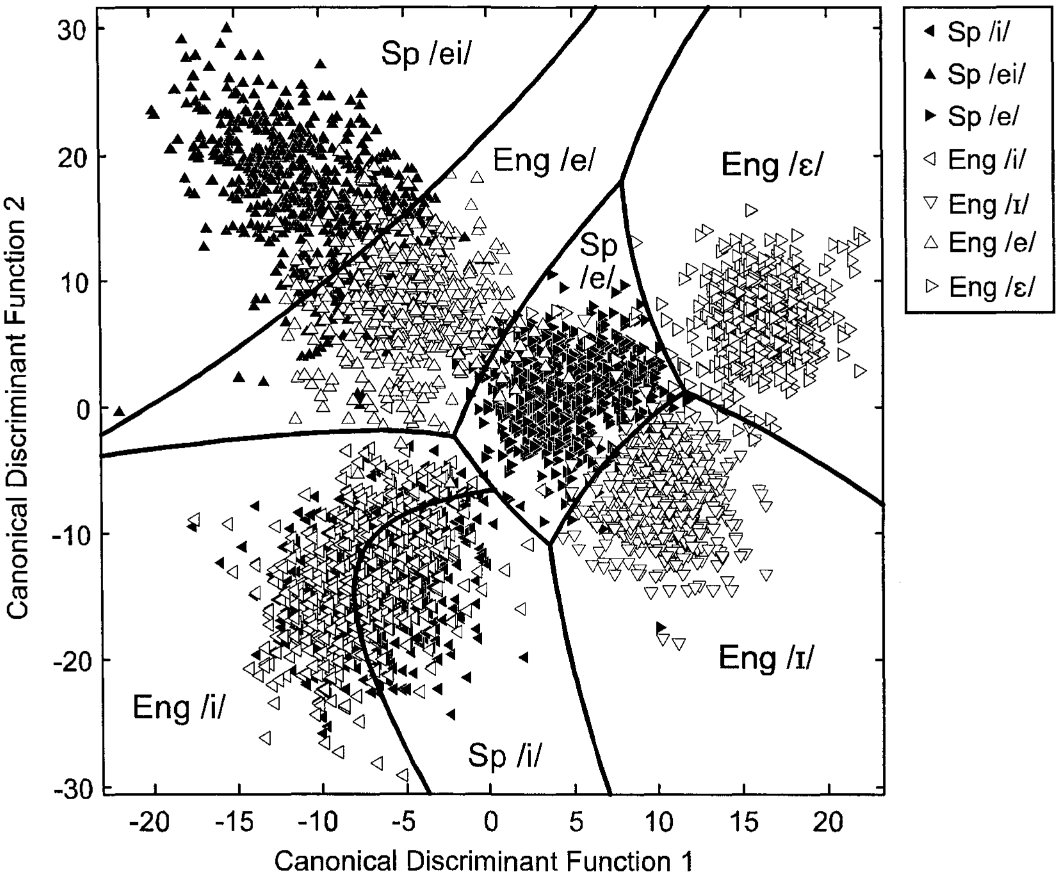


Figure 7.1 Location of L1-English and L1-Spanish speakers' vowel productions in the Function 1 – Function 2 space of a CDFA trained on both L1-English and L1-Spanish speakers' vowel production data.

Table 7.3 Confusion matrix for classification of L1-Spanish and L1-English vowels by the quadratic CDFA trained on L1-Spanish and L1-English vowels. Values are percentages summing to 100 along each row, blank cells have values of zero.

	Predicted						
Produced	Sp /i/	Sp /ei/	Sp /e/	Eng /i/	Eng /ɪ/	Eng /e/	Eng /ɛ/
Sp /i/	74.2		0.3	25.4			
Sp /ei/		84.6				15.4	
Sp /e/	0.3		91.7		4.9	3.1	
Eng /i/	22.9			76.9		0.2	
Eng /ɪ/			2.0		97.6	0.2	0.2
Eng /e/	0.4	10.2	2.4	0.4		86.6	
Eng /ɛ/			2.0		1.0		97.0

The results of the mega-model CDFA will be used to make predictions as to L2 vowel learning beyond initial assimilation to L1 vowels. The percentage of misclassifications provided by the mega-model allows for a fine-grained qualitative version of the SLM's canonical *identical*, *similar*, and *new* relationships. An L2 vowel which has a high rate of misclassification as an L1 vowel is similar to that L1 vowel, and L2 vowels which have high correct classification rates are dissimilar from any L1 vowel. The degree of similarity can be read off the mega-model confusion matrix in Table 7.3. The following predictions are speculative but would seem reasonable on the basis of a distribution-based interpretation of the SLM or on the basis of Escudero's L2LP adapted to a single phonological space (the single phonological space is inherent in the mega-model).[1]

L1-Spanish L2-English learners:

– English /i/ was misclassified as Spanish /i/ at a relatively high rate of 23% (if only the first two canonical functions were used, the misclassification rate was 41%).

– English /i/ is similar to Spanish /i/, and L1-Spanish L2-English learners are therefore predicted to form a single diaphone category covering both vowels (see SLM Hypothesis 5, Flege, 1995). L2-English learners will eventually perceive and pronounce both English /i/ and Spanish /i/ with properties intermediate between those of monolingual versions of Spanish /i/ and English /i/. Perception boundaries will shift to optimally distinguish the distribution of the diaphone (the sum of the monolingual English /i/ and monolingual Spanish /i/ distributions) from the distributions of other vowel categories. Productions will be centred around the centroid of the sum of the monolingual English /i/ and monolingual Spanish /i/ distributions (see SLM Hypothesis 7).

[1] The CDFA mega-model includes a warping of the acoustic space by the discriminant functions based on all the categories included. Another way to perform this analysis would be to apply the discriminant function warping from the L1 then built a model classifying the vowel data from both languages. Such an approach would be more in line with Vallaba & McClelland's topographical map model which models both the changes in acoustic space warping and classification changes.

- English /e/ was misclassified as Spanish /ei/ at a moderate rate of 10%.
  - The relationship of English /e/ to Spanish /ei/ is somewhere between the SLM's canonical *similar* and *new* relationships. Because of the overlap in the distributions of English /e/ to Spanish /ei/, L1-Spanish L2-English learners may initially form a single diaphone category covering both vowels, but with sufficient exposure to English may eventually be able to infer the bimodal distribution and establish a new L2 category for English /e/ with perception and production properties distinct from those of Spanish /ei/ (see SLM Hypothesis 2, Flege, 1995). If a new category is established, productions for L2-English /e/ will eventually be centred around the centroid of the monolingual version of English /e/ (see SLM Hypothesis 7).
- English /ɪ/ and /ɛ/ had correct-classification rates of 97% or greater.
  - L1-Spanish L2-English learners will almost immediately notice instances of English /ɪ/ and /ɛ/ as not good matches for any L1-Spanish vowel category, and will quickly establish new L2 categories. Since both English /ɪ/ and English /ɛ/ are far out on the tails of the distributions of all Spanish vowel categories, the L1-Spanish categories will not present a substantial impediment to the L1-Spanish L2-English learners ability to posit, on the basis of the multimodal distribution in the input, that there are two new English categories in this part of the vowel space. Once the two new L2 categories have been established, the L2 learners will also be able to exploit category labels in the form of L2 lexical information, and will adjust the boundary between the two new L2 categories for optimal English /ɪ/–/ɛ/ perception. Productions for L2-English /ɪ/ and /ɛ/ will be centred around the centroids of the monolingual versions of English /ɪ/ and /ɛ/ (see SLM Hypothesis 7).

L1-English L2-Spanish learners:

– Spanish /i/ was misclassified as English /i/ at a relatively high rate of 25% (if only the first two canonical functions were used, the misclassification rate was 40%).

  – Spanish /i/ is highly similar to English /i/, and L1-English L2-Spanish learners are therefore predicted to form a single diaphone category covering both vowels (a parallel situation to that of L1-Spanish L2-English learners above).

– Spanish /ei/ was misclassified as English /e/ at a moderate rate of 15%.

  – The relationship of Spanish /ei/ and English /e/ is somewhere between the SLM's canonical *similar* and *new* relationships. L1-English L2-Spanish learners may initially form a single diaphone category covering both vowels, but will eventually establish a new L2 category for Spanish /ei/ (a parallel situation to that of L1-Spanish L2-English learners above).

– Spanish /e/ had a correct-classification rate of 92%, with 5% misclassified as English /ɪ/ and 3% as English /e/.

  – L1-English L2-Spanish learners will almost immediately notice most instances of Spanish /e/ as not good matches for any L1-English vowel category, and will quickly establish a new L2 category between the English /i/, /e/, /ɪ/, and /ɛ/ categories. The growth of the new L2-Spanish /e/ category is constrained by the existing L1-English /i/, /e/, /ɪ/, and /ɛ/, categories. Once the new L2 category is established, learners will be able to use category labels (L1-English categories versus new Spanish category) and will therefore be able to adjust perception boundaries until they optimally distinguish L2-Spanish /e/ from the L1 vowels.

At the ultimate achievable state of L2 learning assuming a single phonological space, the L1-Spanish L2-English and L1-English L2-Spanish speakers both have a total of six vowel categories in the vowel space under consideration: Sp/i/+Eng/i/ diaphone, Sp /ei/, Eng/e/, Sp/e/, Eng/ɪ/, and Eng/ɛ/. This state can be simulated via a CDFA model trained on these six vowels, Spanish /i/ and English /i/ data being assigned the same label prior to training. The resulting canonical discriminant function space is very similar to that of the mega-model. Boundaries and centroids are plotted in Figures 7.2 and 7.3, and a territorial map for the synthetic stimulus space is given in Figure 7.4, a confusion matrix is given in Table 7.4.[2] This is a model of the ultimate achievable state of L2 learning, most L2 learners are expected to be at some point in a progression from the initial state for L2 learning to the ultimate achievable state.

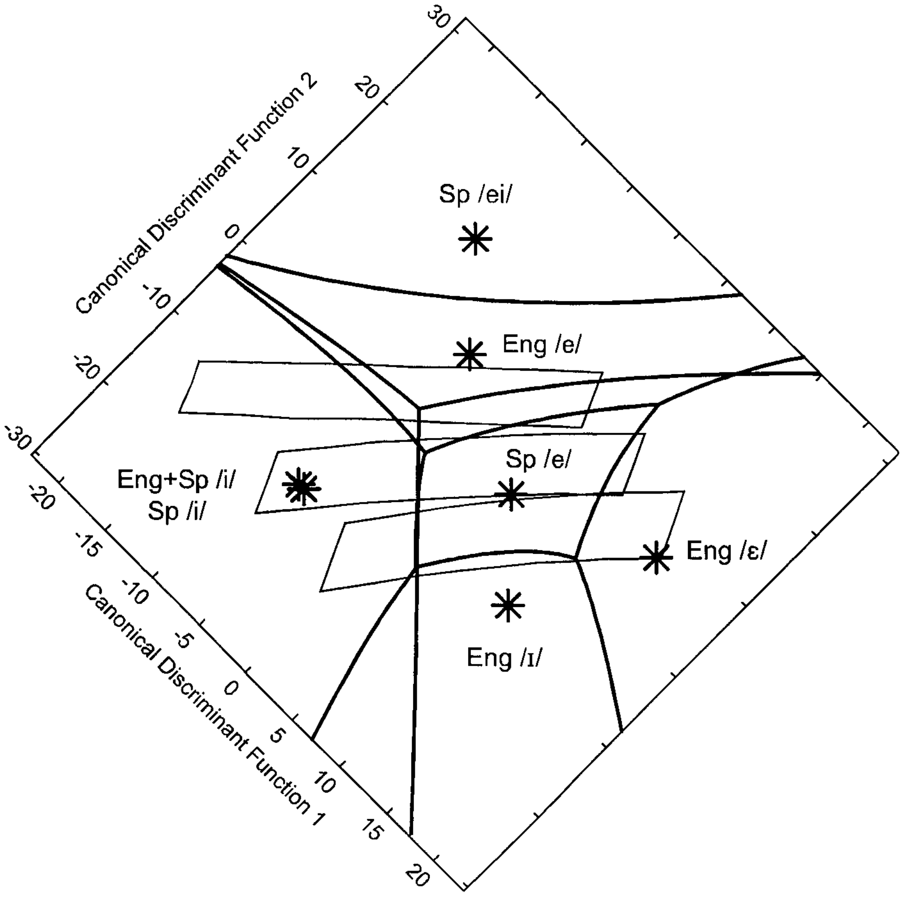


Figure 7.2 Function 1 – Function 2 space of the CDFA trained on vowel production data for the six vowels at the ultimate achievable state of L2 learning. The axes have been rotated so that the dimensions have approximately the same orientation as the synthetic stimulus space in Figure 7.4. Blue lines and stars: boundaries and centroids at L1-Spanish speakers' initial state for L2-English learning. Red lines and green and red stars: boundaries and centroids at L1-Spanish L2-English speakers' ultimate state for L2 learning. Black rhomboids: outlines of synthetic stimulus space.

[2] Differences in boundary location apparent in Figure 7.2 versus 7.3 are due to the fact that the territorial map in the synthetic stimulus space in Figure 7.3 is calculated using three discriminant functions, but only two discriminant function dimensions are plotted in Figure 7.2.

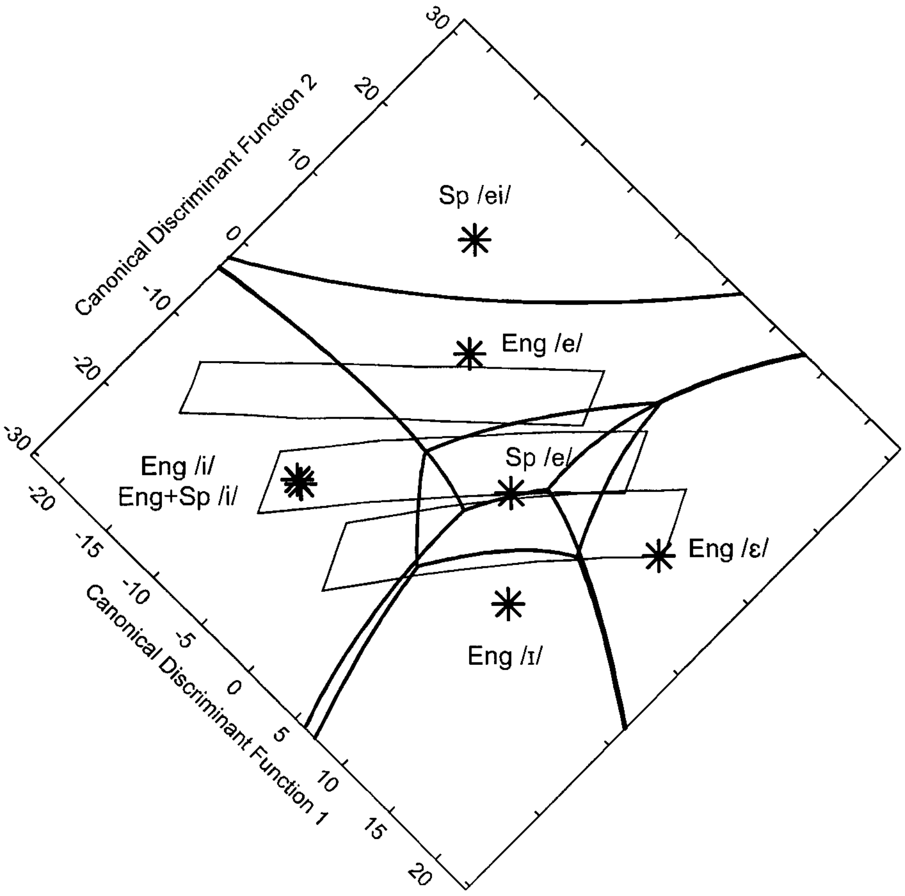


Figure 7.3 Function 1 – Function 2 space of the CDFA trained on vowel production data for the six vowels at the ultimate achievable state of L2 learning. The axes have been rotated so that the dimensions have approximately the same orientation as the synthetic stimulus space in Figure 7.4. Green lines and stars: boundaries and centroids at L1-English speakers' initial state for L2-Spanish learning. Red lines and blue and red stars: boundaries and centroids at L1-English L2-Spanish speakers' ultimate state for L2 learning. Black rhomboids: outlines of synthetic stimulus space.

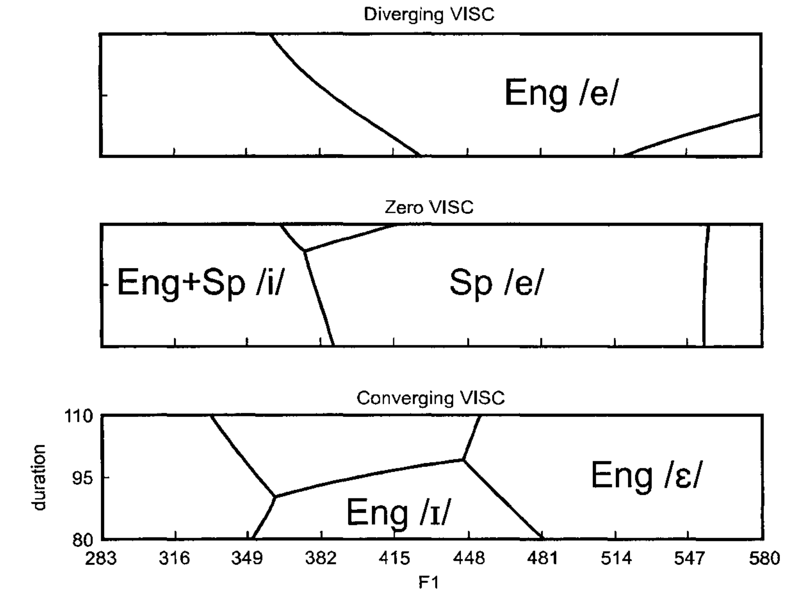


Figure 7.4 Territorial map based on classification of synthetic stimulus values by the CDFA trained on vowel production data for the six vowels at the ultimate achievable state of L2 learning.

Table 7.4 Confusion matrix for classification of the six L1 and L2 vowels at the ultimate achievable state of L2 learning. Values are percentages, blank cells have values of zero.

	Predicted								
Produced	Eng+Sp /i/	Sp /ei/	Eng /e/	Sp /e/	Eng /ɪ/	Eng /ɛ/	Conflated Sp /ei/, Eng /e/	Conflated Sp /e/, Eng /ɪ/, Eng /ɛ/	Conflated Eng /ɛ/, Sp /e/
Eng+Sp /i/	99.7			0.2		0.1			
Sp /ei/		85.3	14.7				100.0		
Eng /e/	0.6	10.8	85.6	3.0			96.4		
Sp /e/	0.3		2.4	92.3	4.9		2.4	97.3	92.3
Eng /ɪ/				2.9	96.9	0.2		100.0	
Eng /ɛ/				2.0	0.8	97.2		100.0	99.2

As well as boundaries and centroids for the ultimate state of L2 learning, Figure 7.2 also includes boundaries and centroids for the initial state of L2 learning by L1-Spanish speakers (i.e., boundaries and centroids based only on L1-Spanish production data). Continuing with predictions made in terms of a distribution-based interpretation of the SLM, to reach the ultimate achievable state of L2 learning by L1-Spanish listeners:

– The absorption of instances of English /i/ into Spanish /i/, has resulted in a diaphone category with a small shift in the location of the centroid.

– The perceptual spaces for the Spanish /ei/ and /e/ categories have been split to accommodate the new L2-English /e/, /ɪ/, and /ɛ/ categories. New L2 category versus L1 category boundaries have been established.

– The old Spanish /i/–/ei/, /ei/–/e/, and /e/–/i/, boundaries have been reused as new Eng+Sp/i/–Eng/e/, Eng/e/–Sp/e/, Sp/e/–Eng+Sp/i/, and Eng/ɪ/–Eng+Sp/i/ boundaries and have shifted to optimal locations for the distinguishing the six vowels.

As well as boundaries and centroids for the ultimate state of L2 learning, Figure 7.3 also includes boundaries and centroids for the initial state of L2 learning by L1-English speakers (i.e., boundaries and centroids based only on L1-English production data). Continuing with predictions made in terms of a distribution-based interpretation of the SLM, to reach the ultimate achievable state of L2 learning by L1-English listeners:

– The absorption of instances of Spanish /i/ into English /i/, has resulted in a diaphone category with a small shift in the location of the centroid, and a small shift in the location of the Eng+Sp/i/–Eng/ɪ/ boundary relative to the original Eng/i/–Eng/ɪ/ boundary.

– The growth of the new L2-Spanish /e/ category is constrained by the existing L1-English /i/, /e/, /ɪ/, and /ɛ/, categories.

Although listeners may initially perceptually categorise an incoming vowel as a member of one of the six L1–L2 vowels, they will ultimately have to identify it with one of the vowel categories in the language to which they are listening. A possible solution to this problem is that they may conflate vowel categories subsequent to initial categorisation; for example, in Spanish mode, vowels initially categorised as either English /e/ or Spanish /ei/ may be identified as Spanish /ei/, and in English mode they may be identified as English /e/. Likewise, in Spanish mode, vowels initially categorised as either English /ɪ/, English /ɛ/, or Spanish /e/ may be identified as Spanish /e/. In English mode, vowels initially categorised as either Spanish /e/ or English /ɛ/ may be identified as English /ɛ/. Identification rates for these conflated categorisations are given in Table 7.4. In production, bilingual speakers may attempt to maintain the distinctions between the six L1–L2 vowels (SLM postulate 2), in which case, in order to minimise the number of vowel productions which cross category boundaries, bilingual speakers would be expected to produce tighter distributions compared to monolinguals' productions of the same categories.

In Section 4.2 it was observed that bilingual L1-Spanish listeners identified some converging-VISC stimuli as Spanish /ei/ (see Figure 4.2) even though Spanish /ei/ is produced with diverging VISC, and it was proposed that this might be a result of L2-English learning. Assuming that the bilingual L1-Spanish listeners' Spanish /ei/ responses in Section 4 were a conflation of the L1-Spanish /ei/ and L2-English /e/ categories, the model of the ultimate achievable state for L2 learning lends support to this proposal. In the territorial map based on the CDFA model of the six L1–L2 vowels, a portion of the converging-VISC subspace has English /e/ as the category with the highest predicted probability.

7.2 L1 Perception Goodness Ratings

Category-goodness ratings may also be helpful in making predictions as to L2 learning beyond initial assimilation of L2 vowels to L1 vowels. Simultaneously with their identification responses to the synthetic stimuli, participants gave category-goodness ratings: The range of possible goodness ratings was 0–100, a stimulus had a goodness rating of zero for a given vowel category if that vowel category was never used as a response for that stimulus, else the goodness rating depended on how high the listeners clicked in the response rectangle (see Figure 2.2). Monolingual listeners' L1 category-goodness ratings will be examined here.

In order to produced smoothed representations of the category-goodness data, a binomial logistic regression model was independently fitted to the pooled-across-listeners goodness ratings for each vowel category (i.e., one binomial model fitted to the pooled-across-listeners goodness ratings for Spanish /i/, another binomial model fitted to the pooled-across-listeners goodness ratings for Spanish /ei/, etc.). The logistic regression models were quadratic:

$$V + \mathrm{F1}\times V + \Delta\mathrm{F1}\times V + \mathrm{dur}\times V + \mathrm{F1}^2\times V + \mathrm{dur}^2\times V + \mathrm{F1}\times\mathrm{dur}\times V$$

with ΔF1 coded as three discrete levels. Goodness-of-fit measures for the models fitted to each vowel category are given in Table 7.5.

Table 7.5 Goodness-of-fit measures for binomial quadratic logistic regression models fitted independently to each vowel category in monolingual Spanish and monolingual English listeners' category-goodness data.

Vowel	G^2	*df*	%SAEP
Sp /i/	114	82	2.14
Sp /ei/	691	82	6.18
Sp /e/	663	82	8.23
Eng /i/	193	82	2.91
Eng /ɪ/	326	82	4.17
Eng /e/	746	82	8.31
Eng /ɛ/	243	82	4.20

For descriptive purposes, the range of possible goodness ratings are divided into three parts: stimuli with goodness ratings above 66.7 may be considered good examples of the respective vowel category, stimuli between 33.3 and 66.7 acceptable examples, and stimuli below 33.3 poor examples; these boundaries fall a little above the *poor* label and a little below the *good* label on the perception experiment's response screen (see Figure 2.3). Figures 7.5 and 7.6 give L1-Spanish and L1-English territorial maps for category-goodness ratings, with 33.3, 50, and 66.7 category-goodness contours for each vowel category.

Visual inspection of Figure 7.5 indicated that:

- The best examples of Spanish /i/ had low F1 and diverging or zero VISC, stimuli with converging VISC were poorer examples of Spanish /i/. Goodness ratings also tended to be better for shorter stimuli.
- The best examples of Spanish /e/ had relatively high F1, zero or converging VISC, and tended to have short vowel duration.
- The best examples of Spanish /ei/ had relatively high F1, diverging VISC, and long duration.

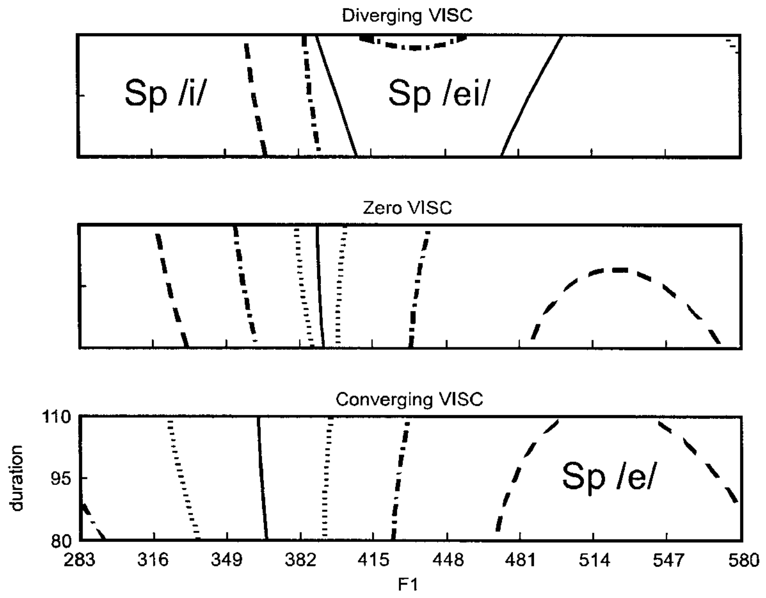


Figure 7.5 Territorial map for monolingual Spanish listeners' category-goodness ratings for each Spanish vowel category in the synthetic stimuli in the perception experiment. Thick lines are category-goodness contours for each vowel category: dotted line 33.3, dotted-dashed line 50, and dashed line 66.7.

A portion of the converging-VISC subspace with F1 from approximately 330 to 395 Hz, had low goodness ratings indicating that stimuli with these acoustic properties were not acceptable examples of any of the Spanish vowel response categories. Note that this is not simply a boundary effect, there is a similar low-goodness area around the Spanish /i/–/e/ boundary in the zero-VISC subspace, but it is much smaller than the low-goodness area in the converging-VISC subspace.[3] This area overlaps primarily with the areas of L1-English /i/ and /ɪ/ production (see Figure 5.4). Although the poor-category-goodness area is on the tails of the production distributions for both English vowels, extrapolating to shorter durations without changing the VISC pattern would extend it towards the English /ɪ/ centroid. Even if L1-Spanish learners of English initially assimilate instances of English /i/ and /ɪ/ in this area to the nearest Spanish vowel category (/i/ or /e/), because of the poor category goodness for any L1-Spanish vowel they will likely be reluctant to continue to do so for very long, and are therefore may quickly establish a new English /ɪ/ category in this part of the converging-VISC subspace.

[3] There is the possibility that they gave poor goodness ratings for any of the response categories because they heard these stimuli as closer to Spanish /ie/ which was not available as a response option.

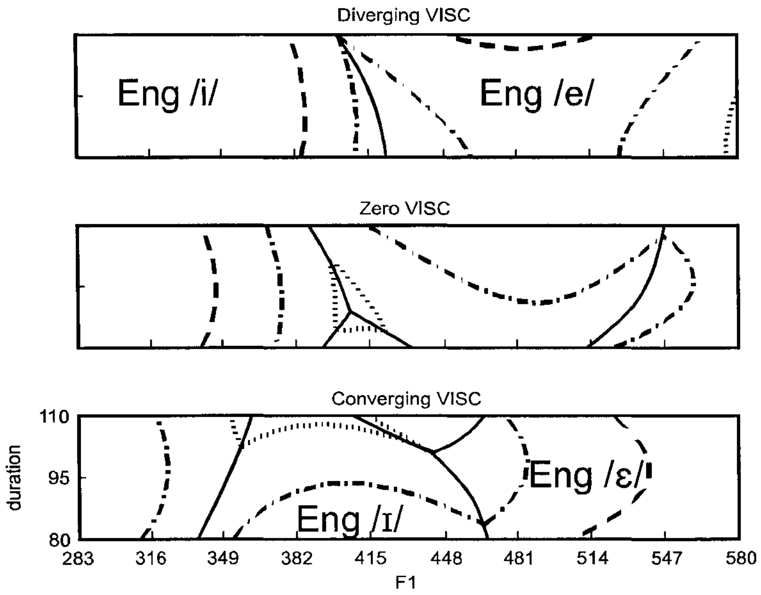


Figure 7.6 Territorial map for monolingual English listeners' mean category-goodness ratings for each English vowel category in the synthetic stimuli in the perception experiment. Thick lines are mean category-goodness contours for each vowel category: dotted line 33.3, dotted-dashed line 50, and dashed line 66.7.

Visual inspection of Figure 7.6 indicated that:

- The best examples of English /i/ had low F1 and diverging or zero VISC.
- The best examples of English /ɛ/ had high F1, converging VISC, and generally had short duration.
- The best examples of English /e/ had relatively high F1, diverging VISC, and long duration.
- The best examples of English /ɪ/ had relatively low F1, converging VISC, and short duration.

A portion of the zero-VISC subspace with short duration had acceptable goodness ratings for English /e/ and was close to the centroid for L1-Spanish /e/ production (see Figure 5.1). Since the goodness ratings were only acceptable and not good, and L1-English L2-Spanish learners will be exposed to a large number of instances of Spanish /e/ in this area, it may be relatively easy for them to establish a new L2-Spanish /e/ category.

A priori, one might expect goodness ratings in perception to be reflective of probability density functions in production. Instances of a vowel category with the most typical production values for that category might be expected to be perceptually the best examples of that vowel category. There were some broad similarities between the PDFs and the goodness ratings, but there were also some differences: Probability density functions indicated low probabilities for the occurrence of English /i/ and Spanish /i/ with diverging VISC, and low probabilities for the occurrence of Spanish /e/ with high F1 and converging VISC; however, listeners gave high goodness ratings to some stimuli with these properties. If the acoustic properties of a vowel place it on the tail of the distribution for its category, and that tail is peripheral in the sense that it is in a direction away from any other vowel categories, then there is even less competition for group membership than in the case of a vowel at the category centroid, and a hyperspace effect (Johnson, Flemming, & Wright, 1993; Frieda et al., 2000) may result so that the more peripheral vowel is perceived as an equally good or even better example of its vowel category than an example at the centroid. To illustrate, adding diverging VISC to a Spanish /i/ moves it further away from Spanish /e/, and as long as the addition of VISC does not move it substantially closer to Spanish /ei/, then on average it will be more distinct from the other Spanish vowels and may therefore be perceived as an equally good or better example of Spanish /i/ than the original version with no VISC. This leads to some potential conflicts between the production-distribution-based and goodness-rating-based predictions. For example, on the basis of goodness rating, it could be hypothesised that instances of Spanish /ei/ will be perceived as hyperspace versions of English /e/ and will therefore be treated as similar to English /e/ even though they are far out on the tail of English /e/. On the basis of production distributions, it could be hypothesised that not only will instances of Spanish /ei/ be perceived as poor members of the English /e/ category because they are far out on the tail of English /e/, but also, because they are further out on the tails of all other English categories, they are more likely to be noticed as not good members of any English category, and therefore the formation of a new L1 Spanish /ei/ category is more likely. Only empirical evidence will demonstrate which of these two hypotheses make the correct prediction.

8. L2-English Perception Results & Discussion

In L1 perception and production, listeners and speakers may reasonably be assumed to exhibit variation around some population average values for properties of speech sound categories which form part of a shared coding system for linguistic transmission and reception. If a listener or speaker were to deviate too far from these population average values, then they would no longer be able to efficiently communicate with other members of the linguistic community which shares this coding system. Results from L1 perception and production experiments may therefore be assumed to be representative of a relatively homogeneous population. In contrast, L2 learners tend to be heterogeneous, and frequently deviate from the population average values of the target language. Although there may be general similarities for L2 learners who share the same L1 and are learning a common L2, these similarities can be difficult to discern because individuals may be at different stages of L2 learning. Differences in the quantity and quality of L2 exposure, in age, in social interaction, and in intrinsic aptitude can lead to different individuals having different rates of progress in L2 learning (see Flege & Liu, 2001). In order to make sense of L2 speech learning, it is therefore necessary to sift through heterogeneous data in search of coherent patterns. Once such patterns have been identified, they may lead to more concrete hypotheses which can be tested in subsequent research. Section 8 therefore presents an account of patterns which have emerged from an extensive exploratory study of the results of the L2-English perception experiments. Hypothesised developmental paths are presented to explain how the patterns observed in the cross-sectional perception results may reflect different stages of L2-English learning. This theory is then tested for consistency with data on L1-Spanish L2-English learners length of residence in Canada. Section 9 presents an account of patterns which have emerged from an extensive exploratory study of the results of the L2-English production experiments. The hypothesised developmental paths, established on the basis of perception results, are then tested for consistency with patterns in the L2-English production results. Finally, in Section 10, the hypothesised developmental paths are tested for consistency with the results of longitudinal case studies.

8.1 Exploratory Results for L2-English Perception of Synthetic Vowels

8.1.1 Principal component analysis

English perception data were available from a total of 86 participants: 19 monolingual English speakers (see Table 2.2), 27 bilingual L1-English speakers (see Table 2.4), and 40 bilingual L1-Spanish speakers (see Table 2.3).

The full logistic regression model $V + \mathrm{F1}\times V + \Delta\mathrm{F1}\times V + \mathrm{dur}\times V$ with ΔF1 coded as three discrete levels, was fitted to each individual participant's English vowel identification data. This provided 86 sets of 15 logistic regression coefficients describing the English vowel perception of each listener.[1] As a method of searching for patterns in the results, these sets of coefficients were entered into a principal component analysis conducted on the correlation matrix (see for example, Johnson, 1998, ch. 5). A principal component analysis transforms the data so that the first principal component accounts for more of the variance than the second, which in turn accounts for more of the variance than the third, etc.. The last few principal components may simply account for noise. Rather than looking for patterns in the space defined by the 15 logistic regression coefficients, it is easier to look for patterns in the space defined by the first few principal components; the space is smaller and the noise level is reduced. The first two principal components accounted for 39.6%, and 19.7% (cumulatively 59.3%) of the variance. Figure 8.1 gives a plot of the first two principal component loading scores,[2] and Table 8.1 gives the total structure coefficients, the univariate correlations between the logistic regression coefficient values (including redundant values) and the principal component values.

[1] The 15 non-redundant coefficients consisted of three bias coefficients, three F1-tuned coefficients, three diverging-VISC-tuned coefficients, three converging-VISC-tuned coefficients, and three duration-tuned coefficients.

[2] Comparisons are not normally made between principal components because each component is normalised (the eigenvectors are normalised to have length 1), component loadings include a scaling factor related to the proportion of variance accounted for by each principal component (the eigenvectors are multiplied by the square root of the corresponding eigenvalue, see Johnson ,1998, p. 98–99).

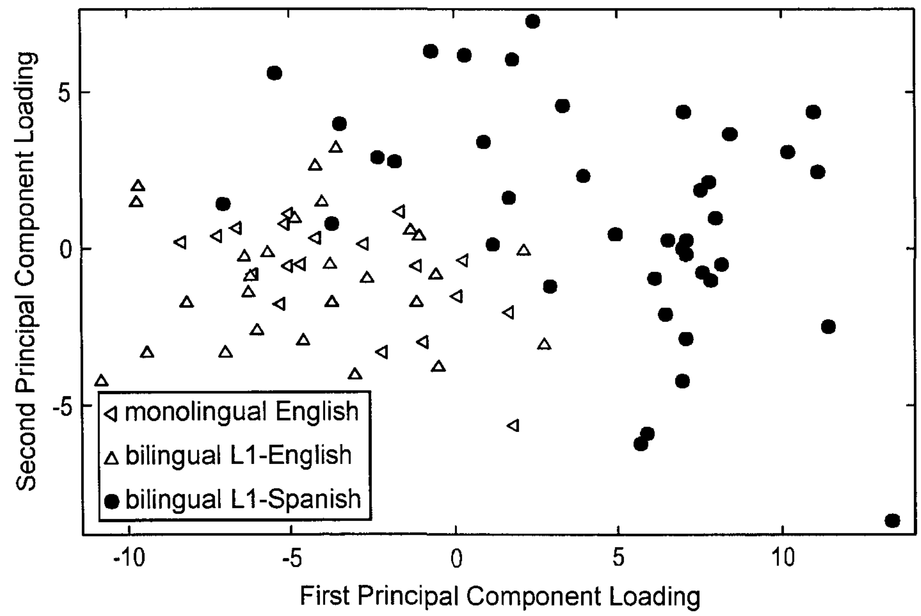


Figure 8.1 First and second principal component loading scores from the principal component analysis conducted on the sets of coefficients from the logistic regression models fitted to each individual listener's English vowel identification data.

Table 8.1 Total structure coefficients: Univariate correlations between logistic regression coefficients from models fitted to individual L1- and L2- English listeners' perception data, and the principal component (PC) transformation of the logistic regression coefficients.

Coefficient	PC1	PC2	Coefficient	PC1	PC2
i	-.825 *	-.194	ΔF1$_{-}$×*i*	.798 *	.199
ɪ	.695 *	-.323 *	ΔF1$_{-}$×*ɪ*	-.883 *	.264 *
e	.071	.811 *	ΔF1$_{-}$×*e*	.316	-.649 *
ɛ	.213	.056	ΔF1$_{-}$×*ɛ*	-.011	-.058
F1×*i*	.857 *	.196	ΔF1$_{+}$×*i*	-.725 *	-.021
F1×*ɪ*	-.838 *	.299 *	ΔF1$_{+}$×*ɪ*	.823 *	-.077
F1×*e*	-.092	-.810 *	ΔF1$_{+}$×*e*	-.021	.606 *
F1×*ɛ*	-.014	-.176	ΔF1$_{+}$×*ɛ*	-.256	-.202 *
dur×*i*	.342	-.097			
dur×*ɪ*	.385 *	.398 *			
dur×*e*	-.584 *	-.389 *			
dur×*ɛ*	-.335	-.063			

* two highest correlations within each family of coefficients

The plot of the first two principal component loading scores indicated that much of this variance was due to inter-L1-group differences: L1-Spanish speakers typically had higher values for the first, second, or both the first two component loadings. A three-dimensional plot of the first three principal component loading scores (accounting for an additional 10.9% of the variance) was also explored, but this did not lead to any additional insight.

Examination of the total structure coefficients indicated that the first principal component was highly correlated with coefficients related to English /i/ and /ɪ/ perception. The second principal component was most highly correlated with coefficients related to English /e/ perception and secondarily with coefficients related to English /ɪ/ perception. The English /i/–/ɪ/ and /ɪ/–/e/ boundaries were therefore targeted for further investigation.

8.1.2 English /i/–/ɪ/ boundary

The boundary between English /i/ and /ɪ/ was explored via examination of /i/–/ɪ/ *contrast coefficients*. Contrast coefficients and their relationship to boundary crispness are discussed in detail in Appendix 7 (see also Morrison, 2005a, 2005b). A contrast coefficient such as (*i-ɪ*)×dur is a measure of the logistic regression model's predicted rate of change from an /ɪ/ to an /i/ response as duration increases. Another contrast coefficient important for quantifying the /i/–/ɪ/ boundary is (*i-ɪ*)×F1. Because the F1 and duration properties of the stimuli were entered into the logistic regression models as just-noticeable differences, the scale for both contrast coefficients is logits per JND (see Appendix 7 for the relationship between logits and probability). These two contrast coefficients can be converted to polar coordinates to give intuitive orthogonal measures of the orientation and crispness of the /i/–/ɪ/ boundary in the F1–duration plane.

Example territorial maps demonstrating boundary angles are given in Figure 8.2. The *angle* part of the polar coordinates was calculated such that an angle of 0° represents the situation where the boundary is perpendicular to the duration dimension (horizontal in the territorial maps) with a greater predicted probability of /i/ for durations greater than the boundary duration (above the boundary line in the territorial maps), angles greater than 0° represent a boundary with a greater predicted probability of /i/ for F1 values greater than the boundary F1 value (to the right of the boundary line in the territorial maps), angles less than 0° represent a boundary with a greater predicted probability of /i/ for F1 values less than the boundary F1 value (to the left of the boundary line in the territorial maps). Angles of 0°, ±90°, and 180° will be invariant irrespective of the relative scaling of the two axes, but intermediate angles will be affected by the relative scaling. Because both the spectral and duration contrast coefficients were scaled to JND units, an angle of 45° (gradient 1) indicates that spectral and duration differences are equally perceptually important, and an angle of 22.5° (gradient 0.5) would indicate that spectral cues are half as important as duration cues, and an angle of 67.5° (gradient 2) would indicate that spectral cues are twice as important as duration cues, etc.. The territorial maps are not scaled to JNDs so plotted angles may not be exactly equal to the calculated angles.

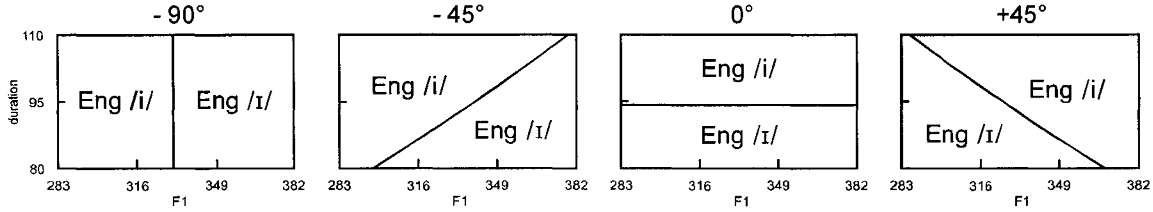


Figure 8.2 Example territorial maps illustrating different angles for the English /i/–/ɪ/ boundary.

The *magnitude* part of the polar coordinates is a measure of how fast the predicted response changes from /ɪ/ to /i/ in the direction in the F1–duration plane perpendicular to the angle of the boundary: the greater the magnitude, the faster the change from /ɪ/ to /i/, the crisper the boundary. Example probability surface plots demonstrating boundary magnitudes are given in Figure 8.3 (see Appendix 7 for an explanation of probability surface plots). Boundaries between well established categories would be expected to be relatively crisp, but boundaries which have just begun to develop would be expected to be relatively fuzzy. Use of JND units for both the spectral and duration contrast coefficients prevents the magnitude varying as a function of boundary angle. If one were to scale the dimensions in, say, Hertz and seconds, then angles near 0° or 180° would have much smaller magnitudes than angles near ±90°.[3] If the magnitude of a boundary is very small, then this indicates that the listener makes no or little differentiation between the two vowel categories, and the angle of the boundary may be irrelevant since it may simply be the result of noise in the vowel identification data (for example, Participant bs088 had an extreme /i/–/ɪ/ boundary angle of +98.5°, but this angle is not meaningful because the boundary magnitude is only 0.16).

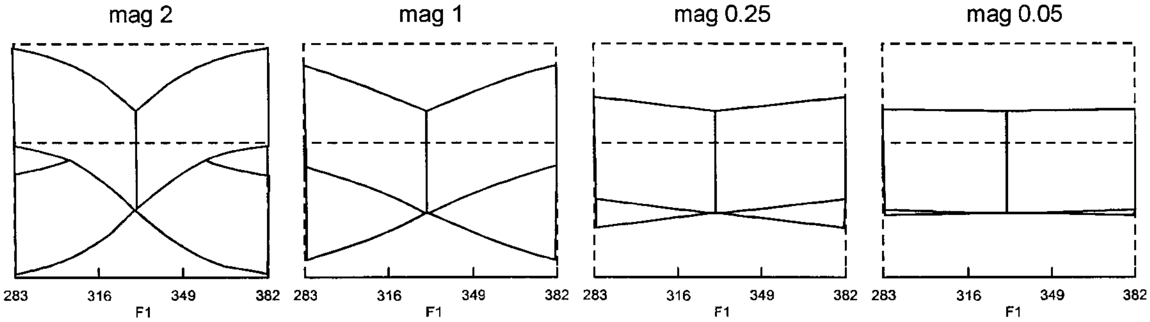


Figure 8.3 Example probability surface plots illustrating different magnitudes for the English /i/–/ɪ/ boundary. Boundary angle is fixed at -90°.

The third contrast coefficient describing the /i/–/ɪ/ boundary is the (*i*-*ɪ*)×(ΔF1$_+$)-tuned (converging-VISC-tuned) contrast coefficient, which indicates the difference in the predicted rate of /i/ versus /ɪ/ responses between the converging-VISC and the zero-VISC stimulus subspace, here measured as the change in logits for the presence versus absence of converging VISC. Since L1-English speakers' natural productions of English /i/ have zero VISC and their natural productions of English /ɪ/ have converging VISC, the diverging-VISC-tuned contrast coefficient is less relevant for describing the /i/–/ɪ/ boundary. A positive

[3] A Euclidian space will be assumed in which the perceptual importance of magnitudes are calculated as the square root of the sum of the squares of the magnitudes on each dimension. This geometry is consistent with the space in which the logistic regression models are calculated, and the reported magnitudes match those of the boundary slopes in the logistic space, which are related to the boundary slopes in the probability surface plots. Another logical possibility would be to calculate the perceptual importance of magnitudes as the sum of the absolute values of the magnitudes on each dimension (city-block distance). Relative to the former geometry, the latter would inflate magnitudes for angles away from 0°, ±90°, and 180°.

value for the converging-VISC-tuned contrast coefficient indicates a greater predicted ratio of /i/ to /ɪ/ responses in the converging-VISC subspace relative to the zero-VISC subspace. Example territorial maps demonstrating the converging-VISC contrast are given in Figure 8.4. A change in the converging-VISC value results in a shift in the location of the /i/–/ɪ/ boundary in the converging-VISC subspace relative to the zero-VISC subspace (but no change in angle or magnitude). A change in bias coefficients would result in an equal shift in the location of the boundary in both the zero- and converging-VISC subspaces, but no change in angle, magnitude, or relative location of the boundaries between the zero- and converging-VISC subspaces; each of these measures is orthogonal.

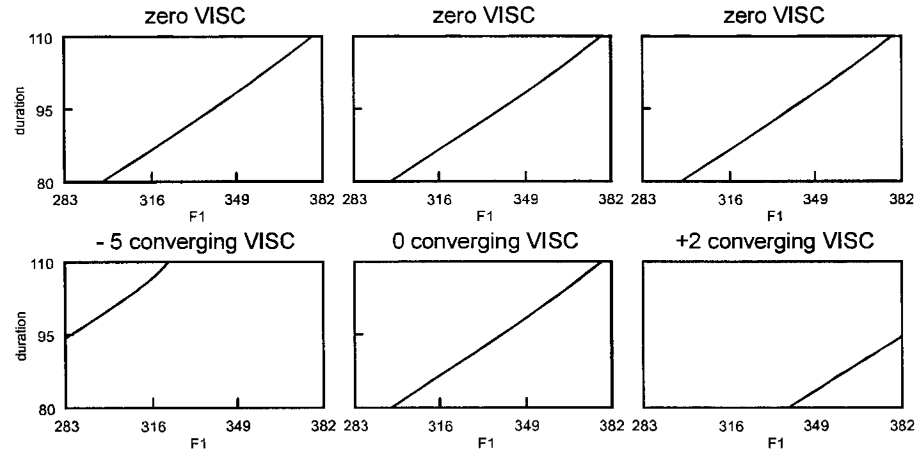


Figure 8.4 Example territorial maps illustrating different converging-VISC values for the English /i/–/ɪ/ boundary. Boundary angle is fixed at −45°.

A plot of the contrast coefficient values for the F1–duration plane angle and magnitude, and the converging-VISC contrast for the /i/–/ɪ/ boundaries of each individual is given in Figure 8.5. L1-English listeners' clustered around a mean /i/–/ɪ/ boundary angle of − 82.0°, a mean magnitude of 1.74, and mean converging-VISC contrast coefficient value of −4.45. Some L1-Spanish L2-English listeners fell within the L1-English range, but most fell into one of two groups, a group with more negative boundary angles, in particular angles exceeding −90°, and a group on a diagonal of more positive boundary angles and more positive converging-VISC contrast coefficient values. L1-Spanish listeners also tended to have smaller boundary magnitudes, mean 0.85.

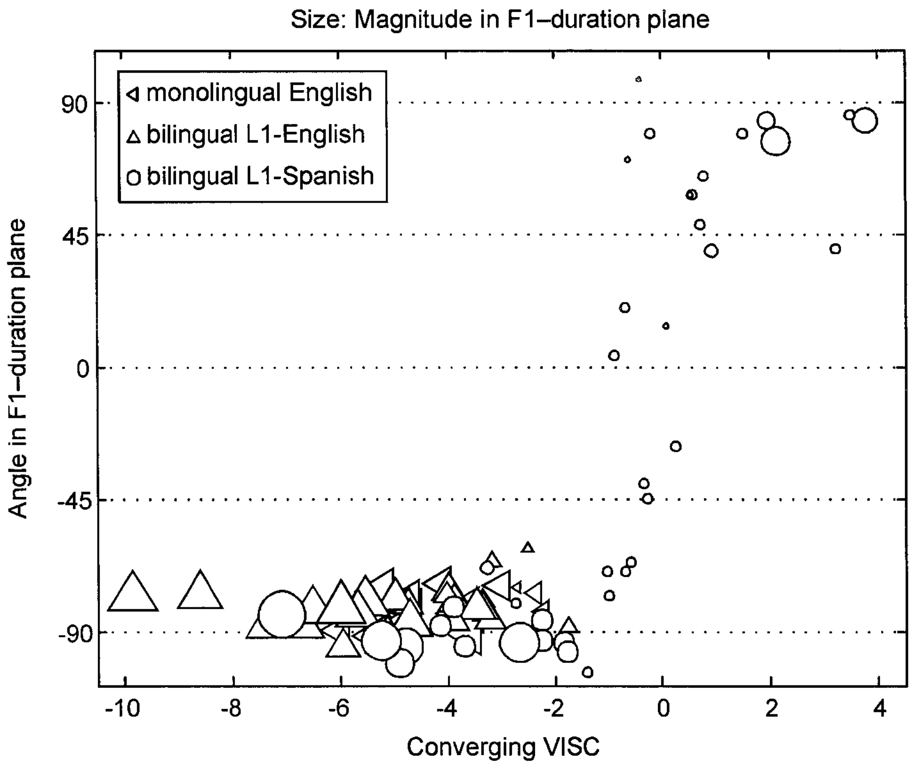


Figure 8.5 Plot of the F1–duration-plane angle and magnitude, and the converging-VISC contrast for the English /i/–/ɪ/ boundaries from logistic regression models calculated for each participant.

The relationship between the English /i/–/ɪ/ boundary perception patterns and L2-English learning will be discussed in detail in Section 8.2. Table 8.2 gives the angles and magnitudes in the F1–duration plane and the converging-VISC coefficients for the /i/–/ɪ/ boundaries based on individual logistic regression models of each bilingual L1-Spanish listener's English vowel identification data. The grouping of the participants into paths and stages will be discussed in Section 8..2.

8.1.3 English /e/–/ɪ/ boundary

The boundary between English /e/ and /ɪ/ was explored via examination of /e/–/ɪ/ contrast coefficients. The /e/–/ɪ/ boundary will be characterised using a set of coordinates parallel to the set used to characterise the /i/–/ɪ/ boundary (in the example territorial maps and probability surface maps above, substitute /e/ for /i/). Since L1-English speakers' natural productions of /e/ have diverging VISC and their natural productions of /ɪ/ have converging VISC, the VISC contrast coefficient used was calculated as the converging-VISC-tuned contrast coefficient minus the diverging-VISC-tuned contrast coefficient. A plot of the F1–duration plane angle and magnitude and the VISC contrast for the /e/–/ɪ/ boundary of each individual is given in Figure 8.6. L1-English listeners clustered around a mean /e/–/ɪ/ boundary angle of +47.4°, a mean magnitude of 0.84, and mean VISC contrast coefficient value of -3.18. Most L1-Spanish L2-English listeners had more positive boundary angles, most close to +90°, and more positive VISC values. L1-Spanish listeners also tended to have larger boundary magnitudes, mean 1.14.

Table 8.2 Angles and magnitudes in the F1–duration plane and VISC contrast values for the English /i/–/ɪ/ boundaries calculated on the basis of bilingual L1-Spanish listeners' English vowel responses. Participants arranged according to assignment to hypothesised indirect and direct developmental paths for English /i/–/ɪ/ learning, see section 8.2. Dotted boxes: Notable boundary properties outside L1-English range. Solid boxes: Notable boundary properties within L1-English range

		English /i/–/ɪ/ boundary					English /i/–/ɪ/ boundary		
Stage	Participant	Angle°	Magnitude	VISC	Stage	Participant	Angle°	Magnitude	VISC
0	bs071	+59.3	0.09	0.52	2	bs062	–66.1	0.37	–0.59
	bs114	+14.0	0.12	0.05		bs019	–69.0	0.52	–0.69
	bs088	+98.5	0.16	–0.42		bs049	–69.2	0.40	–1.06
	bs065	+71.3	0.17	–0.64		bs056	–77.4	0.58	–1.01
½	bs050	+86.1	0.56	3.48	3	bs067	–85.7	1.10	–2.31
	bs076	+84.8	1.49	3.75		bs028	–79.8	0.47	–2.77
	bs051	+84.5	0.98	1.92		bs073	–67.8	0.71	–3.28
	bs069	+80.4	0.54	1.48		bs023	–81.6	1.16	–3.92
	bs052	+80.4	0.31	–0.23		bs077	–87.6	1.18	–4.16
	bs064	+77.7	1.55	2.09		bs057	–83.7	2.73	–7.11
	bs001	+65.4	0.48	0.75	direct path	bs061	–103.6	0.55	–1.42
	bs075	+59.4	0.35	0.55		bs082	–100.6	1.63	–4.90
	bs072	+49.3	0.44	0.69		bs002	–94.9	1.28	–3.70
	bs003	+40.7	0.62	3.21		bs074	–96.7	1.08	–1.81
	bs087	+40.4	0.74	0.87		bs068	–94.7	1.93	–4.83
1	bs083	+20.5	0.36	–0.69		bs078	–93.4	1.15	–1.87
	bs081	+3.9	0.22	–0.91		bs059	–93.3	1.42	–4.93
	bs058	–26.7	0.32	0.24		bs017	–93.2	2.17	–2.72
	bs016	–39.4	0.28	–0.36		bs063	–92.8	2.15	–5.27
	bs091	–44.9	0.47	–0.30		bs086	–92.7	1.26	–2.31

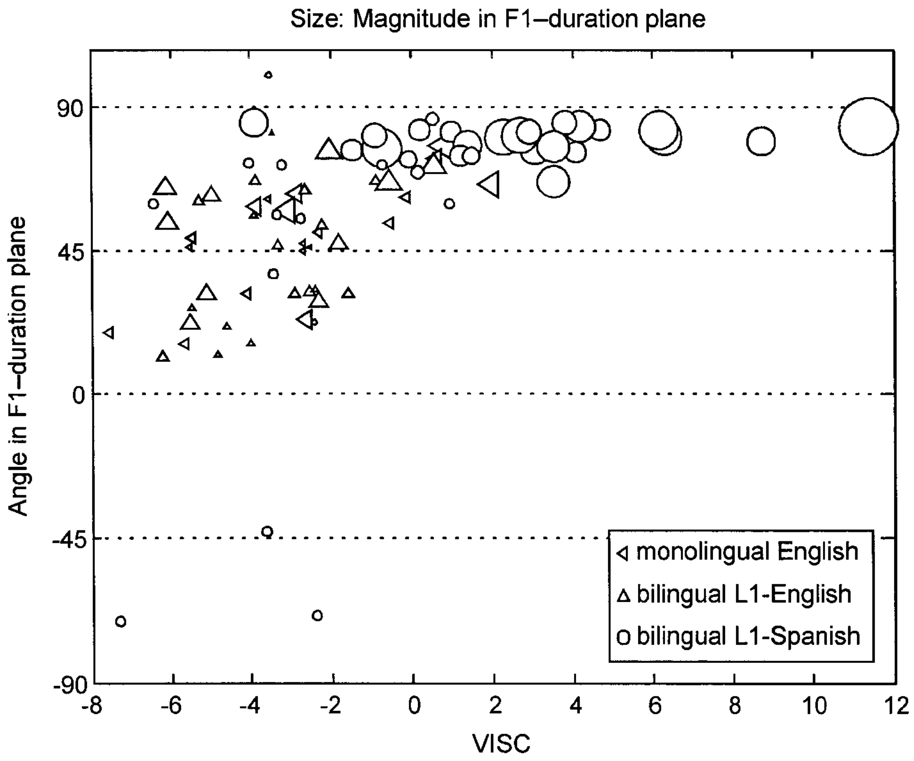


Figure 8.6 Plot of the F1–duration-plane angle and magnitude, and the VISC contrast for the English /e/–/ɪ/ boundaries from logistic regression models calculated for each participant (outlier with a large negative angle and small magnitude not plotted).

The small F1–duration plane boundary magnitudes for the L1-English listeners may be because spectral and duration properties are secondary cues for the English /e/–/ɪ/ contrast, and relatively unimportant compared to VISC cues. Another possible explanation for the small magnitudes is related to the fact that L1-English /e/ and /ɪ/ are widely separated in production. When two categories are close to each other, such as L1-English /ɛ/ and /ɪ/ (see in Figure 3.4), then a crisp stable boundary is needed to obtain a high correct-classification rate, but when there is a gap between the two categories, such as between L1-English /e/ and /ɪ/, then a high correct-classification rate can be obtained even if the boundary is relatively fuzzy and the exact location of the boundary is not as important. In the perception experiment, listeners were asked to identify synthetic stimuli which were in the gap between natural L1-English /e/ and /ɪ/ productions. This may therefore account for the small F1–duration plane boundary magnitudes, and relatively high variance in boundary angle.

The large F1–duration plane magnitudes and angle of approximately +90° for the L1-

Spanish L2-English listeners may indicate that they are reusing an existing L1-Spanish boundary. Both the L1-Spanish /i/–/e/ and /i/–/ei/ boundaries had angles close to +90°, the /i/–/e/ boundary had a positive VISC contrast, and the /i/–/ei/ boundary had a negative VISC contrast, see Figure 4.1.

Given that the English /e/–/ɪ/ boundary is rather poorly defined for L1-English listeners, we will not focus further on this contrast.

8.2 Hypothesised Developmental Paths for English /i/–/ɪ/ Learning

8.2.1 Hypothesised indirect developmental path

Visual inspection of the plot of English /i/–/ɪ/ contrast coefficients in Figure 8.5 suggested that the L1-Spanish L2-English listeners' could be divided into three major groups: those who had more negative F1–duration-plane boundary angles than L1-English listeners (in particular angles beyond −90°), those who had larger converging-VISC values and mostly more positive F1–duration-plane boundary angles than L1-English listeners, and those whose English /i/–/ɪ/ boundary properties were within the L1-English range.

The patterns for the latter two groups were consistent with stages on the hypothetical path for L1-Spanish speakers' learning of the English /i/–/ɪ/ contrast proposed by Escudero (2000) and extended by Morrison (2005b), see Section 1.3.2. Although L2 perception development is envisioned as continuous movement along a path, to simplify qualitative description, this movement is described in terms of stages. Participants were assigned to the hypothesised stages on the basis of their English /i/–/ɪ/ contrast coefficients. This was a subjective procedure which involved identifying groups of listeners who were contiguous in Figure 8.5 and who had territorial maps which matched the descriptions of each of the stages. These stage-groups are indicated in Figure 8.7 and Table 8.2. This hypothesised developmental path will be identified as the *indirect developmental path*, and is described below in relation to the English /i/–/ɪ/ perception results.

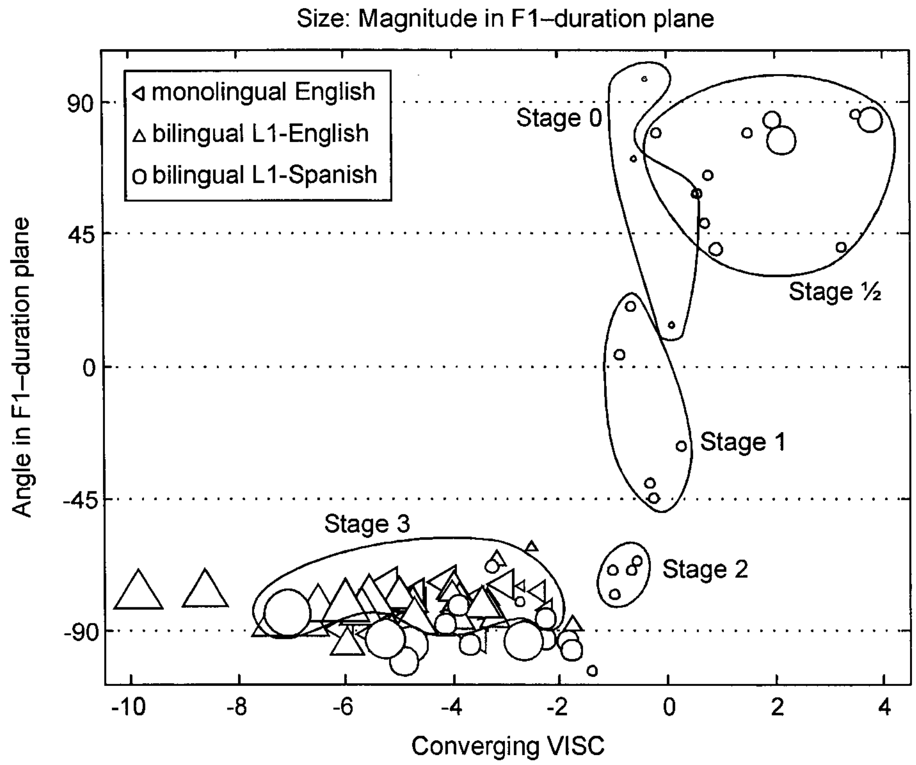


Figure 8.7 Plot of the F1–duration-plane angle and magnitude, and the converging-VISC contrast for the English /i/–/ɪ/ boundaries from logistic regression models calculated for each participant. Groupings indicate participants at different stages of the hypothesised indirect developmental path, and participants on the direct developmental path.

Hypothesised Indirect Developmental Path:

- Stage 0, general inability to perceive the English /i/–/ɪ/ contrast. The magnitudes of the English /i/–/ɪ/ boundaries are very small (see Table 8.2). The listeners may have chosen English /i/ and /ɪ/ responses at random, irrespective of stimulus properties, possibly with a bias towards one or the other.
- Stage ½, multidimensional category-goodness-difference assimilation of English /i/ and /ɪ/ to Spanish /i/. Figure 8.8 gives example territorial maps for participants assigned to this stage. The part of the stimulus space with the most Spanish-/i/-like properties (low F1, short duration, and non-converging VISC) has the highest predicted probability for English /ɪ/, and the adjacent part of the stimulus space with less Spanish-/i/-like properties (any combination of higher F1, longer duration, and converging VISC) has a higher predicted probability for English /i/. The use of

spectral properties (both initial formant values and VISC) is negatively correlated with L1-English speakers' productions of English /i/ and /ɪ/, and the use of duration is positively correlated.

Participants with this pattern of English /i/– /ɪ/ were hypothesised to be making a category-goodness-difference assimilation because the pattern is consistent with such a hypothesis and is not consistent with any pattern which could be induced by exposure to English. For example, if L1-Spanish L2-English listeners first used duration to perceptually distinguish English /i/ and /ɪ/, and then used this to bootstrap the use of spectral cues, then their use of duration and spectral cues would both be positively correlated with duration and spectral properties in L1-English production (or both negatively correlated if the L2-English listeners reversed the English /i/ and /ɪ/ labels). A perception pattern could not emerge in which duration cues are positively correlated and spectral cues negatively correlated with L1-English production properties.

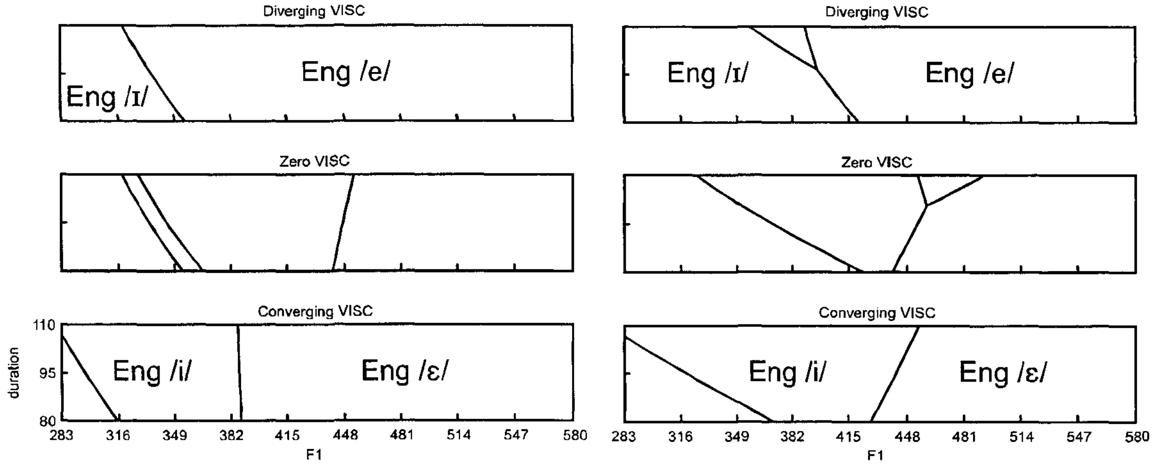


Figure 8.8 Example territorial maps for participants assigned to Stage ½ of the hypothesised indirect developmental path. Left: bs064, /i/–/ɪ/ angle +77.7°, converging VISC 2.09
Right: bs075, /i/–/ɪ/ angle +59.4°, converging VISC 0.55

– Stage 1, duration-based perception of English /i/ and /ɪ/. Figure 8.9 gives example territorial maps for participants assigned to this stage. The use of spectral properties (both initial formant values and VISC) has declined relative to the use of duration, the latter being the only cue used at Stage ½ which was positively correlated with native English production, and thus the only cue reinforced by exposure to English.

The low-F1 part of the stimulus space with longer duration has the highest predicted probability for English /i/ responses, and the low-F1 part of the stimulus space with shortest duration has the highest predicted probability for English /ɪ/ responses.

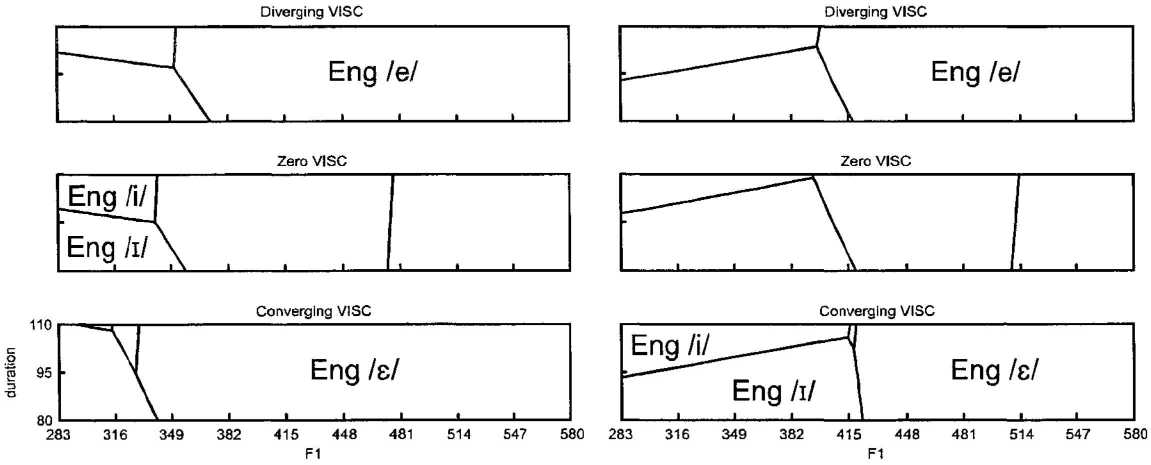


Figure 8.9 Example territorial maps for participants assigned to Stage 1 of the hypothesised indirect developmental path. Left: bs083, /i/–/ɪ/ angle +20.5°, converging VISC –0.69
Right: bs058, /i/–/ɪ/ angle –26.7°, converging VISC 0.24

– Stage 2, L1-English-like duration- and spectral-based perception of the English /i/–/ɪ/ contrast, but with weaker cue weighting. Figure 8.10 gives example territorial maps for participants assigned to this stage. The part of the stimulus space with the longest duration, lowest F1 and non-converging VISC has the highest predicted probability for English /i/, and the parts of the stimulus space with shorter duration, higher F1 and converging VISC have higher predicted probabilities for English /ɪ/. The use of F1–duration cues is within the L1-English range in terms of boundary angle (see Figure 8.7), but the boundary magnitudes are small compared to the L1-English mean. Although used in the same direction, VISC is also not as strong a cue for L1-Spanish L2-English listeners at this stage as it is for L1-English listeners. The use of spectral cues is hypothesised to have been bootstrapped off the duration cue since initial formant values and VISC are partially correlated with duration in L1-English production of English /i/ and /ɪ/ (all else being equal, English /ɪ/ is shorter than English /i/ as well as having higher-F1–lower-F2 and converging VISC compared to zero VISC).

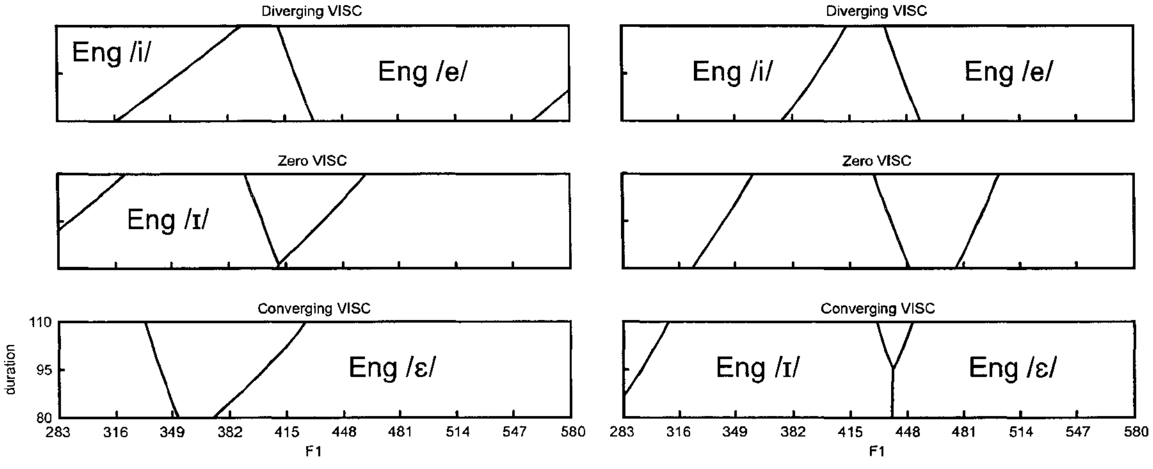


Figure 8.10 Example territorial maps for participants assigned to Stage 2 of the hypothesised indirect developmental path. Left: bs062, /i/–/ɪ/ angle –66.1°, converging VISC –0.59
Right: bs056, /i/–/ɪ/ angle –77.4°, converging VISC –1.01

The Stage 2 presented here has maintained the spirit of Escudero's (2000) original description of Stage 2 as duration and spectral cues used in the same direction as L1-English listeners but short of the weighting magnitudes used by L1-English listeners. Escudero's original formulation was based only on duration and steady-state spectral properties, and her "weighting" was the ratio of spectral to duration reliance, akin to the boundary angles presented here. In the formulation presented here L2-English listeners have L1-English-like use of duration and initial spectral properties in terms of boundary angle, but boundary magnitude is weaker than the L1-English mean, and they also have more weakly weighted use of VISC properties. Participants assigned to Stage 2 here would have been assigned to Stage 3 in the original formulation, and some of the participants assigned to Stage 1 here, those with the most negative boundary angles, would have been assigned to Stage 2 in the original formulation.

– Stage 3, L1-English-like perception of the English /i/–/ɪ/ contrast. Figure 8.11 gives example territorial maps for participants assigned to this stage. Perception of the English /i/–/ɪ/ contrast is now within the L1-English range for F1-duration boundary angle and magnitude, and VISC cue magnitude (see Figure 8.7). This is viewed as the result of additional exposure to English leading to a strengthening of the cues already in use at Stage 2.

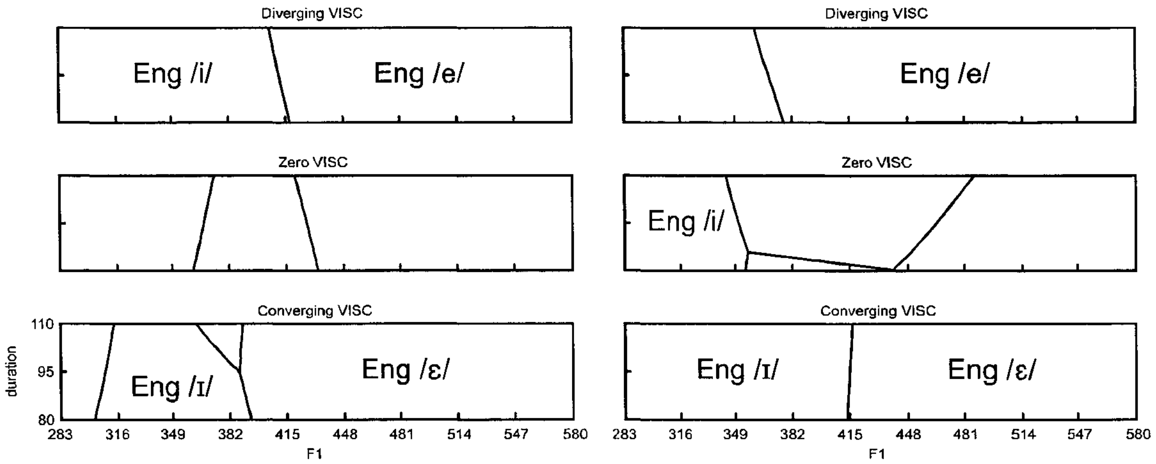


Figure 8.11 Example territorial maps for participants assigned to Stage 3 of the hypothesised indirect developmental path. Left: bs067, /i/–/ɪ/ angle –85.7°, converging VISC –2.31
Right: bs077, /i/–/ɪ/ angle –87.6°, converging VISC –4.16

8.2.2 Hypothesised direct developmental path

The other major group of L1-Spanish L2-English listeners identified in Section 8.1.1 had more negative English /i/–/ɪ/ boundary angles than L1-English listeners, in particular boundary angles beyond –90°. They did not have perceptual patterns which were consistent with any of the stages of the hypothesised indirect path for English /i/–/ɪ/ learning. Figure 8.12 gives example territorial maps for participants assigned to this group. As was the case for L1-English listeners, stimuli with low F1 and non-converging VISC had the highest predicted probability for /i/, and stimuli with higher F1 and converging VISC had a higher predicted probability for /ɪ/, but longer stimuli had a higher predicted probability for /ɪ/, unlike all (except three) of the L1-English listeners and all of the L1-Spanish L2-English listeners assigned to the hypothesised indirect developmental path.

This pattern would be consistent with a multidimensional category-goodness-difference assimilation of English /i/ and /ɪ/ to Spanish /i/, with stimuli which are more similar to Spanish /i/ (low F1, short duration, and non-converging VISC) labelled as English /i/, and stimuli which are less similar (any combination of higher F1, longer duration, and converging VISC) labelled as English /ɪ/. Use of F1 and VISC cues are positively correlated with L1-English speakers' productions of English /i/ and /ɪ/, and use of duration is negatively correlated; however, since the spectral cues are much more important for L1-English listeners than are duration cues (see Section 4.1.2), this immediately results in relatively L1-English-like perception. These L1-Spanish L2-English listeners are therefore hypothesised to be on an alternative English /i/–/ɪ/ learning path which will be identified as the *direct developmental path.*

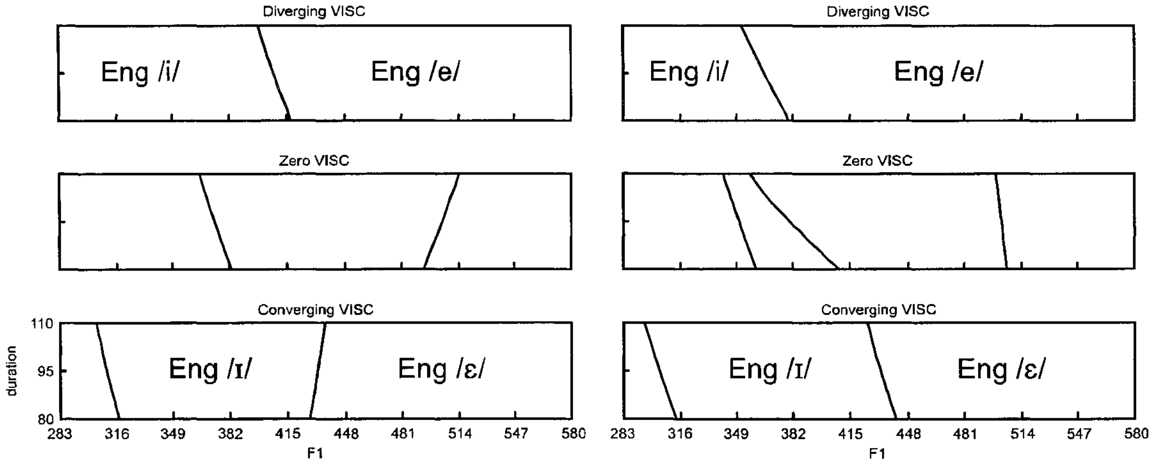


Figure 8.12 Example territorial maps for participants assigned to the hypothesised direct developmental path.
Left: bs068, /i/–/ɪ/ angle –94.7°, converging VISC –4.83
Right: bs074, /i/–/ɪ/ angle –96.7°, converging VISC –1.81

An alternative interpretation for this pattern is that these participants are actually at Stage 3 of learning and that their use of duration cues is effectively zero and within the L1-English range. A third interpretation is that the stimuli identified as English /ɪ/ are perceived as poor examples of Spanish /e/ rather than Spanish /i/ (the a priori hypothesis based on L1 production and perception models). Support for these hypotheses comes from the fact that listeners in this group had relatively large boundary magnitudes which would be consistent with the boundary having been firmly established, either because the listeners are at an advanced stage of L2-English learning, or because they are reusing their Spanish /i/–/e/ boundary. These possibilities will be discussed more when the L2-English production results are analysed.

8.2.3 Discussion

We now have two hypothesised developmental paths for English /i/–/ɪ/ learning by L1-Spanish speakers, the *indirect path* and the *direct path.* Both paths begin with multidimensional category-goodness-difference assimilation to Spanish /i/: Vowels with low F1 and short duration, and without converging VISC are more Spanish-/i/ like, and vowels with higher F1, longer duration, and converging VISC are less Spanish-/i/ like. However, the two hypothesised paths differ with respect to the association between more and less Spanish-/i/-like vowels and the English category labels /i/ and /ɪ/:

– In the direct path, the more Spanish-/i/-like stimuli are labelled as English /i/ and the less Spanish-/i/-like stimuli as English /ɪ/. Labelling is based primarily on the absence versus presence of converging VISC, converging VISC being uncharacteristic of Spanish /i/. This immediately results in relatively L1-English-like perception since converging VISC is a important cue for L1-English listeners' perception of English /ɪ/, and the L1-Spanish L2-English listeners use VISC in the same direction as L1-English listeners.

– In the indirect path, the more Spanish-/i/-like stimuli are labelled as English /ɪ/ and the less Spanish-/i/-like stimuli as English /i/. Labelling is based primarily on vowel duration, Spanish /i/ being a shorter vowel. This leads to very non-L1-English-like perception since the primary cues for L1-English listeners, spectral cues, are used in the opposite direction by L1-Spanish L2-English listeners, and only duration, at best a secondary cue for L1-English listeners, is used in the same direction.

It is assumed that L1-Spanish participants who are generally unable to distinguish English /i/ and /ɪ/ (Stage 0), may subsequently travel along either of the hypothesised developmental paths, and that L1-Spanish participants with L1-English-like perception (Stage 3) may have reached that stage via either hypothesised developmental path.

The question arises as to why, at the category-goodness-difference stage, some L1-Spanish learners of English base their use of the English /i/ and /ɪ/ labels on duration, and others on VISC differences. One possibility is that prior to their arrival in Canada the participants who used duration were exposed to English dialects which had a more pronounced duration difference or a less pronounced spectral difference between /i/ and /ɪ/. Another possibility is that, although they perceived both spectral and duration differences, explicit English language instruction, that English /i/ is long and English /ɪ/ is short, caused them to label the English vowels on the basis of duration. The orthographic factor mentioned in Section 1.3.2, English /ɪ/ maps to the letter 'i' which maps to Spanish /i/, would also explain the direction of labelling relative to good and poor examples of Spanish /i/.

Participants with greater formal classroom ESL education, and written-mode ESL education, may be more likely to follow the indirect developmental path.

8.2.4 Length of residence

If a hypothesised developmental path is in fact developmental, then one would expect there to be a relationship between length of residence (LOR) in the L2 environment and progression along the path. Because L2 learners may fossilise, early stages of development may subsume L2 learners with relatively long LORs as well as L2 learners with short LORs, but later stages of development would be expected to include only L2 learners with long LORs.

A plot of LOR against stage of development in the hypothesised indirect developmental path is given in Figure 8.13. The majority of participants fell withing the predicted triangular pattern.[4] Note that LOR was not taken into account when assigning L1-Spanish L2-English listeners to the different stages of L2-English /i/–/ɪ/ learning, the author did not refer to LOR data until after the participants had been assigned to stages; hence there was no possibility of even subconscious influence of LOR data when making the assignments.

The predicted pattern was spoilt by the results from three participants who were assigned to Stage 3 but turned out to have short LORs. Participant bs023 is not a particularly extreme outlier, he may have been a particularly gifted language-learner and may have been able to reach Stage 3 in only six months. Participant bs067 had been in Canada for only two weeks at the time of the test, but her work involved extensive interaction with L1-English speakers, and thus LOR was not a representative measure of her exposure to English. This leaves Participant bs073 as the only clear spoiler; it would seem highly unlikely that he had progressed through the earlier stages and reached Stage 3 after only one month in Canada. One participant therefore has results which are clearly inconsistent with the hypothesised stages of development actually being developmental, and a single outlier is probably not sufficient to falsify the hypothesis. In general, the LOR results therefore appear to be

[4] Within Stage 1, the three participants with the longest LORs also has the most L1-English-like /i/–/ɪ/ boundary angles.

consistent with the hypothesised indirect developmental path being, in fact, developmental.

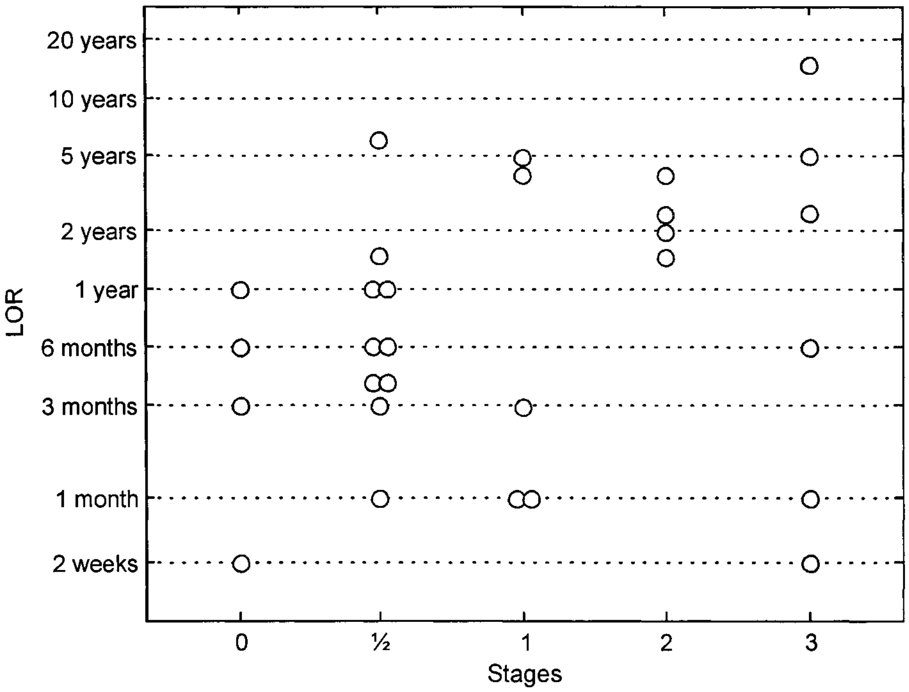


Figure 8.13 Plot of length of residence (LOR) in Canada against stage in hypothesised indirect developmental path for individual L1-Spanish L2-English participants.

Most participants assigned to the hypothesised direct developmental path had LORs of 1 year or more which would be consistent with either of the category-goodness-difference hypotheses, or the hypothesis that these participants were actually at Stage 3 of L2-English /i/–/ɪ/ learning. However, two participants had LORs of 1 month or less which would only be consistent with a category-goodness-difference hypothesis.

In Section 9, L2-English production results will be presented and discussed in relation to the hypothesised developmental paths.

9. L2-English Production Results & Discussion

9.1 L2-English Production Patterns

L2-English production data were available from 40 bilingual L1-Spanish speakers (see Table 2.3). The recordings of L2 vowel productions were acoustically analysed using the same methodology used to analyse the L1-vowel-production recordings. For each individual L1-Spanish L2-English speaker, an F1–F2 comet plot and an F1–duration scatter plot was produced showing the properties of each instance of each L1-Spanish and L2-English vowel category. Judgements of similarity or separation between vowel categories were based on visual inspection of the degree of overlap or size of gap between clouds of comets/points representing each vowel category. Focussing on English /i/–/ɪ/ production, five production patterns (A, B, C, D, E) were identified:

– *Production Pattern A*, substitution of L1-Spanish vowels for L2-English vowels: L1-Spanish L2-English speakers produced no, or negligible, separation between their English /i/ and /ɪ/ productions, and these L2 vowels were produced with properties similar to the speakers' own L1-Spanish /i/ productions. They also produced English /e/ with properties similar to Spanish /ei/, and produced English /ɛ/ with properties similar to Spanish /e/. An example of Production Pattern A is given in Figure 9.1.

Table 9.1 gives examples of classification of Production Pattern A L2-English productions by the linear CDFA model trained on L1-English speakers' vowel productions. L2-English /i/ and /ɪ/ productions were usually not differentially classified by the L1-English production model, both vowels were typically classified as English /i/. L2-English /e/ was almost always correctly classified. L2-English /ɛ/ was frequently misclassified as English /ɪ/ and less frequently as English /e/. When compared with the classification of L1-Spanish vowels by the L1-English production model, it was apparent that the classifications were consistent with these participants substituting Spanish /i/ for English /i/ and /ɪ/, expected if both English vowels were assimilated to Spanish /i/. Classifications were also consistent with substituting Spanish /ei/ for English /e/. The case for L1-Spanish L2-English speakers substituting Spanish /e/ for English /ε/ was not as strong. Some differences in production were apparent in the plots of acoustic properties, and there were also some differences in classification. Crisp classification on a small number of tokens can sometimes exaggerate differences, a posteriori probabilities are therefore also provided in Table 9.1, averaged over each category produced; L2-English /ε/ and L1-Spanish /e/ classifications for Participant bs071 appear to be quite different on the basis of crisp classification, but more similar on the basis of mean a posteriori probabilities.

Table 9.1 Examples of classification of individual bilingual L1-Spanish speakers' L2-English vowels, Production Pattern A, and L1-Spanish vowels by the L1-English production model.

Left: Crisp classification, numbers are counts, blank cells have zero counts.

Right: Mean a posteriori probabilities, blank cells have probabilities less than .0005.

bs071 60.0% correct

	Predicted			
Produced	Eng /i/	Eng /ɪ/	Eng /e/	Eng /ε/
Eng /i/	10			
Eng /ɪ/	10			
Eng /e/			10	
Eng /ε/		6		4
Sp /i/	10			
Sp /ei/			10	
Sp /e/				10

	Predicted			
Produced	Eng /i/	Eng /ɪ/	Eng /e/	Eng /ε/
Eng /i/	.995	.005		
Eng /ɪ/	.998	.001	.001	
Eng /e/			1.000	
Eng /ε/		.471	.001	.528
Sp /i/	.990	.007	.003	
Sp /ei/			1.000	
Sp /e/		.285	.004	.711

bs075 50.0% correct

	Predicted			
Produced	Eng /i/	Eng /ɪ/	Eng /e/	Eng /ε/
Eng /i/	10			
Eng /ɪ/	10			
Eng /e/			10	
Eng /ε/		9	1	
Sp /i/	10			
Sp /ei/			10	
Sp /e/		1	9	

	Predicted			
Produced	Eng /i/	Eng /ɪ/	Eng /e/	Eng /ε/
Eng /i/	1.000			
Eng /ɪ/	1.000			
Eng /e/			1.000	
Eng /ε/		.732	.105	.164
Sp /i/	1.000			
Sp /ei/			1.000	
Sp /e/		.197	.796	.007

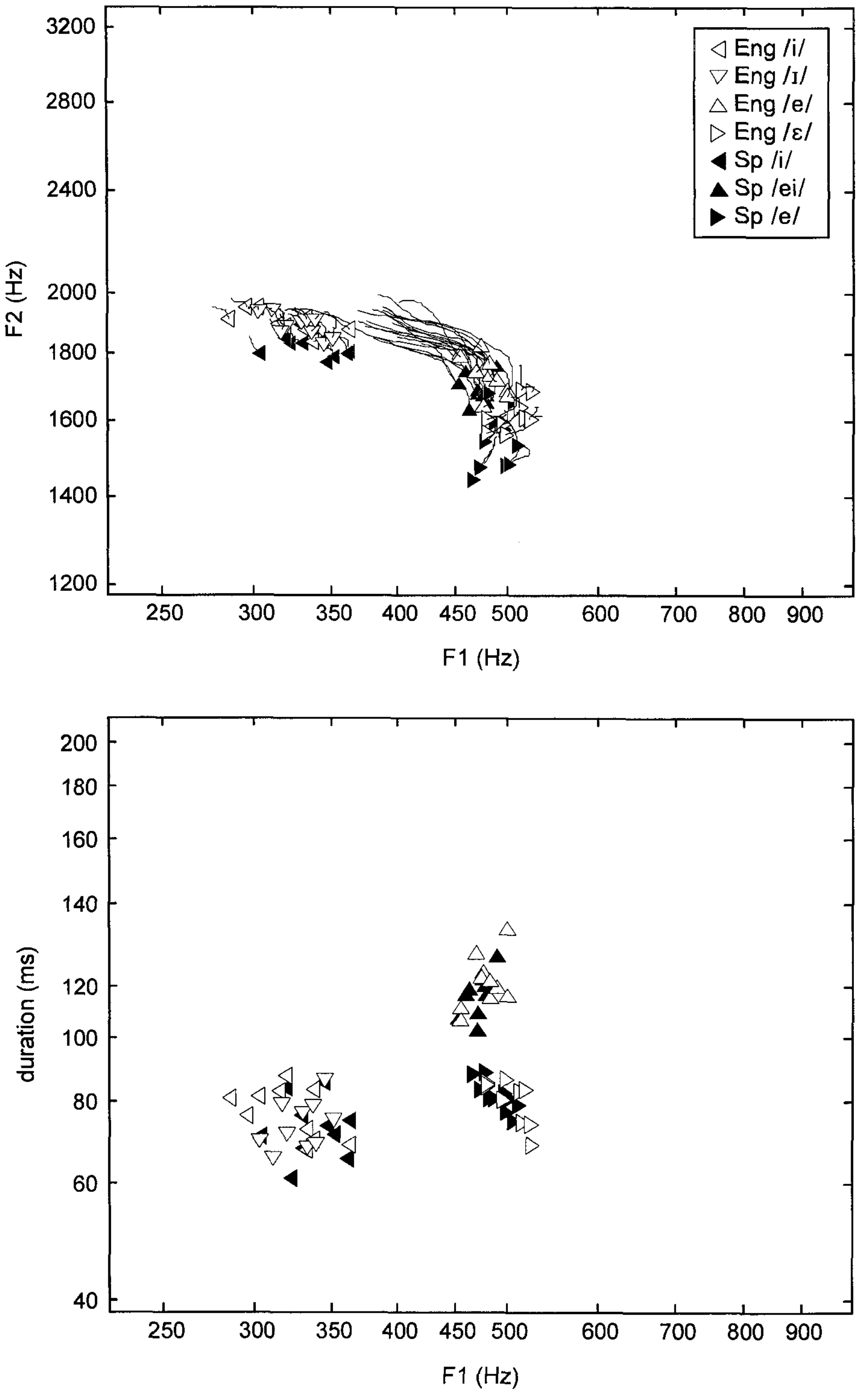


Figure 9.1 Example of Perception Pattern A, substitution of Spanish vowels for English vowels (Participant bs071).

– *Production Pattern B*, L1-Spanish L2-English speakers produced longer English /i/ than English /ɪ/, and usually English /ɪ/ was similar in duration to Spanish /i/. An example of Production Pattern B is given in Figure 9.2.

Table 9.2 gives examples of classification of Production Pattern B L2-English productions by the linear CDFA model trained on L1-English speakers' vowel productions. Differences in the duration of English /i/ and /ɪ/ productions did not result in correct classification for L2-English /ɪ/, they were almost always classified as English /i/. L2-English /i/ productions were usually classified as English /i/, but sometimes the long duration resulted in them being misclassified as English /e/. Classification results were otherwise similar to those for Production Pattern A

Table 9.2 Examples of classification of individual bilingual L1-Spanish speakers' L2-English vowels, Production Pattern B, and L1-Spanish vowels by the L1-English production model. Crisp classification, numbers are counts, blank cells have zero counts.

bs083 52.5% correct

	Predicted			
Produced	Eng /i/	Eng /ɪ/	Eng /e/	Eng /ɛ/
Eng /i/	9		1	
Eng /ɪ/	10			
Eng /e/			10	
Eng /ɛ/		7	1	2
Sp /i/	10			
Sp /ei/			10	
Sp /e/		7	3	

bs016 42.9% correct

	Predicted			
Produced	Eng /i/	Eng /ɪ/	Eng /e/	Eng /ɛ/
Eng /i/	6		4	
Eng /ɪ/	11			
Eng /e/			11	
Eng /ɛ/		5	4	1
Sp /i/	10			
Sp /ei/			10	
Sp /e/		8	2	

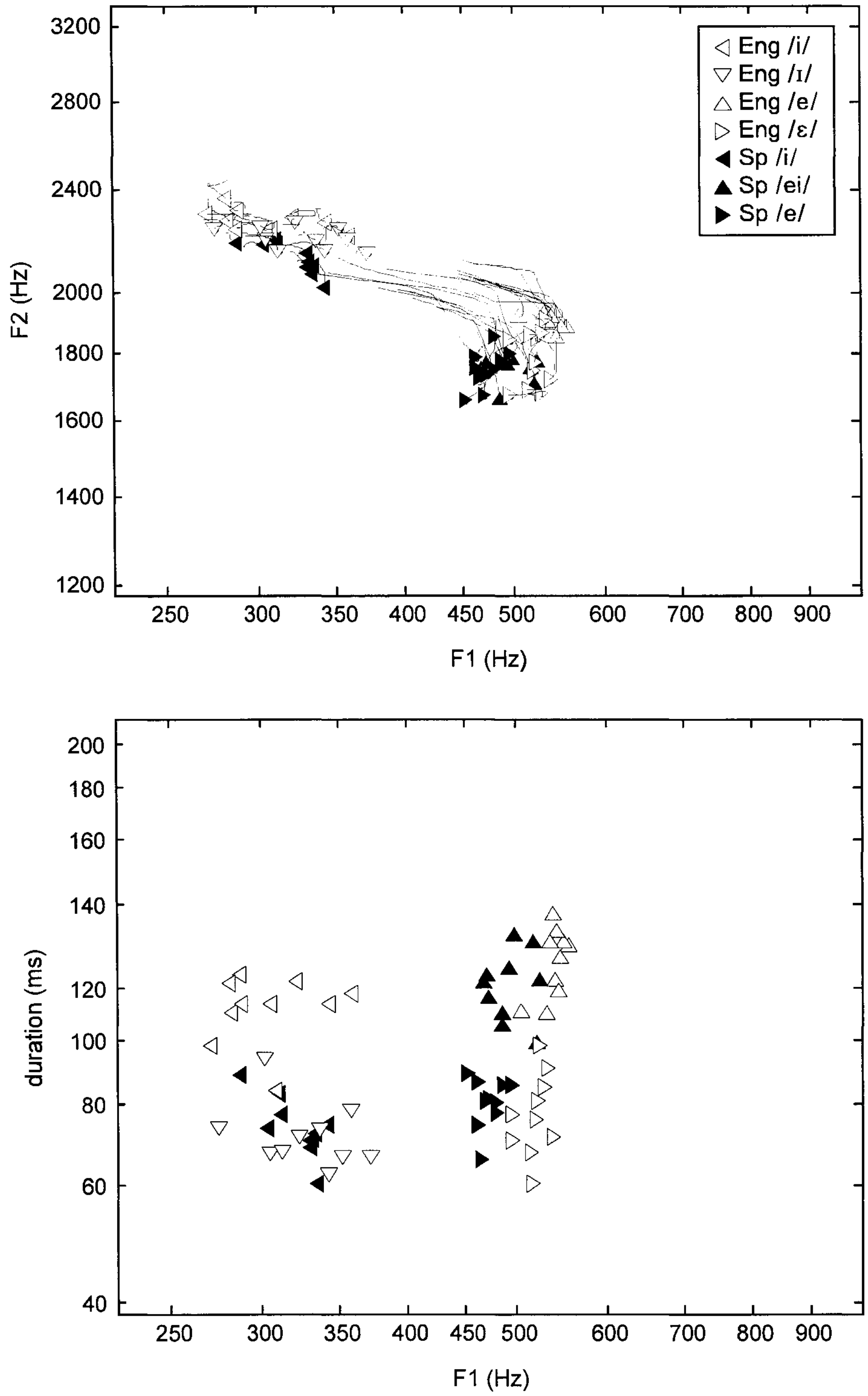


Figure 9.2 Example of Production Pattern B, English /i/ longer than English /ɪ/ (Participant bs083).

– Production Pattern C, L1-Spanish L2-English speakers produced English /ɪ/ which had higher F1 and lower F2 than English /i/, and usually English /i/ had spectral properties similar to Spanish /i/. L2-English /ɪ/ was not typically produced with converging VISC. An example of Production Pattern C is given in Figure 9.3.

Table 9.3 gives examples of classification of Production Pattern C L2-English productions by the linear CDFA model trained on L1-English speakers' vowel productions. Differences in the spectral properties of English /i/ and /ɪ/ resulted in some correct classifications for L2-English /ɪ/; however, they were mostly classified as English /i/ and occasionally as English /e/. L2-English /i/ productions were usually classified as English /i/. Classification results were otherwise similar to those for Production Pattern A

Table 9.3 Examples of classification of individual bilingual L1-Spanish speakers' L2-English vowels, Production Pattern C, and L1-Spanish vowels by the L1-English production model. Crisp classification, numbers are counts, blank cells have zero counts.

bs063 57.5% correct

	Predicted			
Produced	Eng /i/	Eng /ɪ/	Eng /e/	Eng /ɛ/
Eng /i/	10			
Eng /ɪ/	7	3		
Eng /e/			10	
Eng /ɛ/		10		
Sp /i/	10			
Sp /ei/			10	
Sp /e/		1	9	

bs086 66.7% correct

	Predicted			
Produced	Eng /i/	Eng /ɪ/	Eng /e/	Eng /ɛ/
Eng /i/	10			
Eng /ɪ/	7	2		
Eng /e/			10	
Eng /ɛ/		6		4
Sp /i/	10			
Sp /ei/			10	
Sp /e/		5	5	

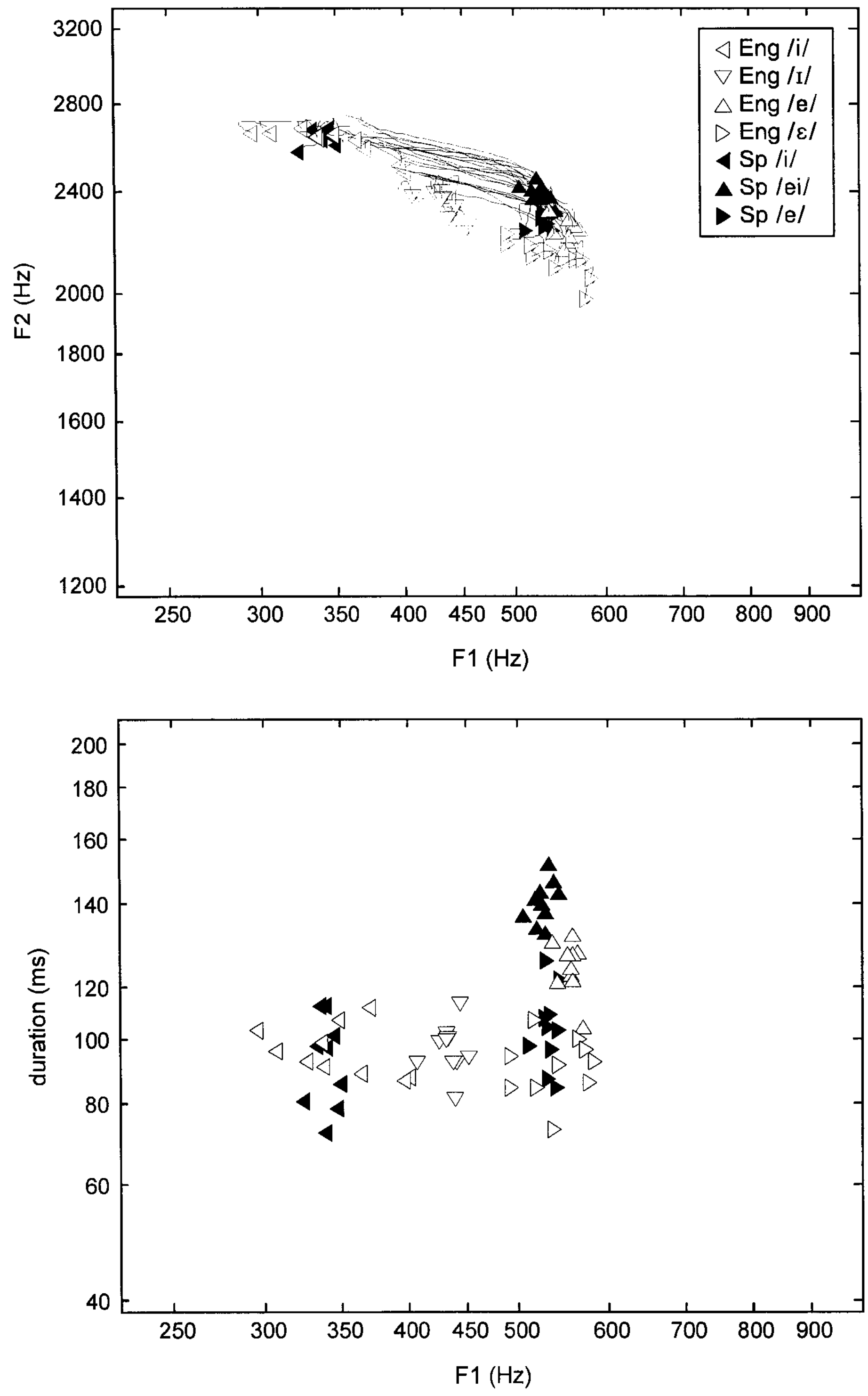


Figure 9.3 Example of Production Pattern C, English /ɪ/ produced with higher-F1–lower-F2 than English /i/ (Participant bs063).

– *Production Pattern D*, L1-Spanish L2-English speakers produced English /ɪ/ which had higher F1 and lower F2 and were shorter than their English /i/ productions. L2-English /ɪ/ productions also typically had higher-F1–lower-F2 and shorter duration than Spanish /i/. L2-English /ɪ/ was not typically produced with converging VISC. An example of Production Pattern D is given in Figure 9.4.

Table 9.4 gives examples of classification of Production Pattern D L2-English productions by the linear CDFA model trained on L1-English speakers' vowel productions. Differences in the spectral and duration properties of English /i/ and /ɪ/ resulted in high rates of correct classifications for L2-English /ɪ/.

Some borderline cases between Production Patterns B and D had small spectral differences. L2-English /ɪ/ productions had spectral and duration properties similar to Spanish /i/, and L2-English /i/ productions had slightly lower F1 and slightly higher F2 than Spanish /i/. Classification patterns were like those of Production Pattern A.

Table 9.4 Examples of classification of individual bilingual L1-Spanish speakers' L2-English vowels, Production Pattern D, and L1-Spanish vowels by the L1-English production model. Crisp classification, numbers are counts, blank cells have zero counts.

bs023 70.0% correct

	Predicted			
Produced	Eng /i/	Eng /ɪ/	Eng /e/	Eng /ɛ/
Eng /i/	10			
Eng /ɪ/	1	9		
Eng /e/		1	9	
Eng /ɛ/		10		
Sp /i/	9	1		
Sp /ei/			10	
Sp /e/		3		7

bs078 80.5% correct

	Predicted			
Produced	Eng /i/	Eng /ɪ/	Eng /e/	Eng /ɛ/
Eng /i/	10			
Eng /ɪ/		11		
Eng /e/			10	
Eng /ɛ/		8		2
Sp /i/	10			
Sp /ei/			10	
Sp /e/		6		4

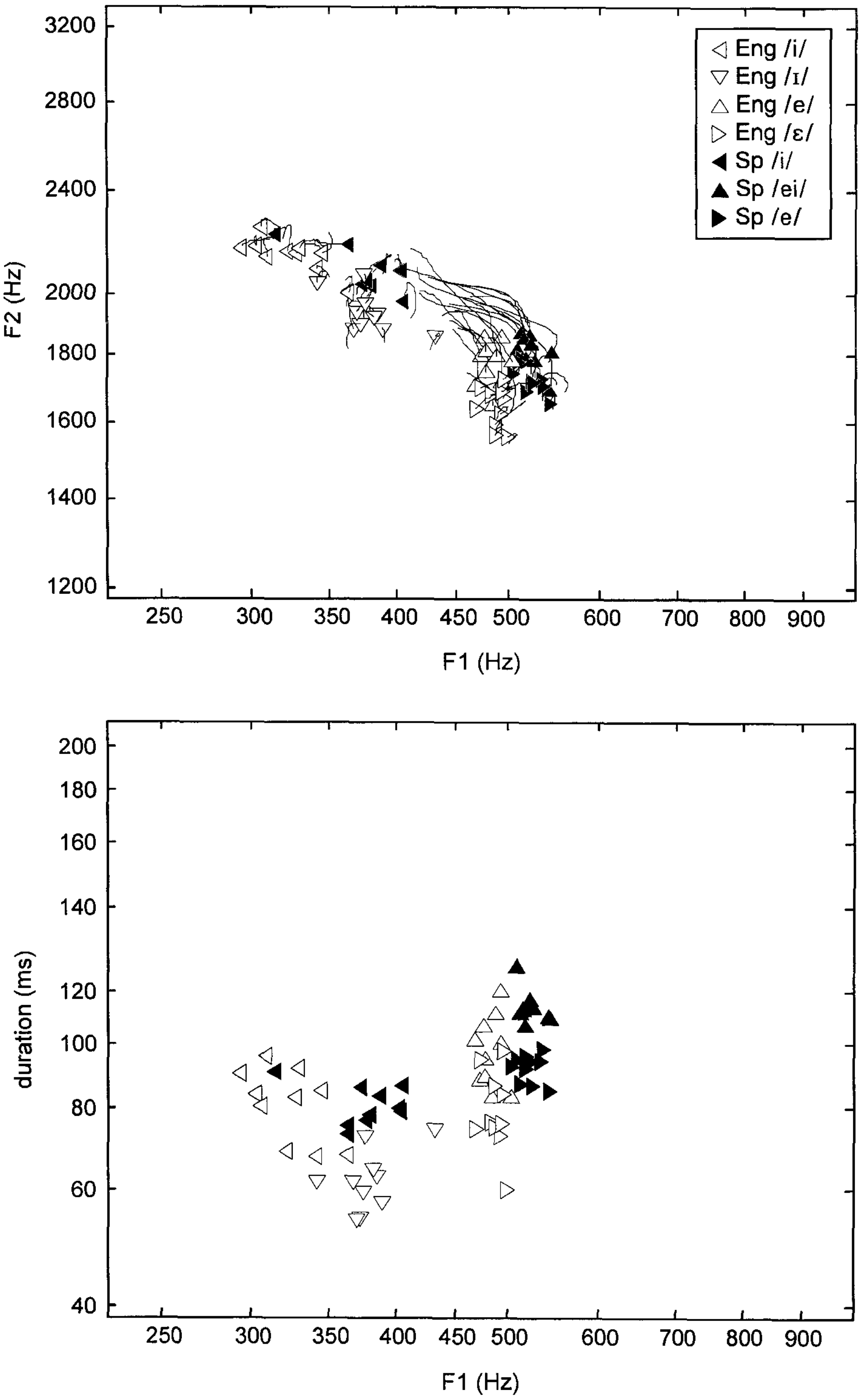


Figure 9.4 Example of Production Pattern D, English /ɪ/ produced with higher-F1–lower-F2 and shorter duration than English /i/ (Participant bs023).

– *Production Pattern E,* L1-Spanish L2-English speakers produced English /ɪ/ and English /ɛ/ with spectral and duration properties similar to their Spanish /e/ productions. L2-English /ɪ/ and /ɛ/ were not typically produced with converging VISC. This pattern is consistent with L1-Spanish participants assimilating both these English vowels to Spanish /e/ and substituting Spanish /e/ in production (although there were usually some small differences between the productions of the L1-Spanish and the two L2-English categories). This pattern is consistent with the a priori predictions made on the basis of L1 production and perception models for the initial state of L2-English learning (see Sections 3.5 and 4.3.2). An example of Production Pattern E is given in Figure 9.5.

Table 9.5 gives examples of classification of Production Pattern E L2-English productions by the linear CDFA model trained on L1-English speakers' vowel productions. Producing English /ɪ/ with spectral and duration properties similar to Spanish /e/ resulted in some correct classifications, but also some misclassifications as English /ɛ/ and /e/.

Table 9.5 Examples of classification of individual bilingual L1-Spanish speakers' L2-English vowels, Production Pattern E, and L1-Spanish vowels by the L1-English production model. Crisp classification, numbers are counts, blank cells have zero counts.

bs068 55.0% correct

	Predicted			
Produced	Eng /i/	Eng /ɪ/	Eng /e/	Eng /ɛ/
Eng /i/	10			
Eng /ɪ/		2	3	5
Eng /e/			10	
Eng /ɛ/		2	8	
Sp /i/	10			
Sp /ei/			10	
Sp /e/		8	2	

bs057 72.5% correct

	Predicted			
Produced	Eng /i/	Eng /ɪ/	Eng /e/	Eng /ɛ/
Eng /i/	10			
Eng /ɪ/		8		2
Eng /e/			10	
Eng /ɛ/		9		1
Sp /i/	10			
Sp /ei/			10	
Sp /e/		2	8	

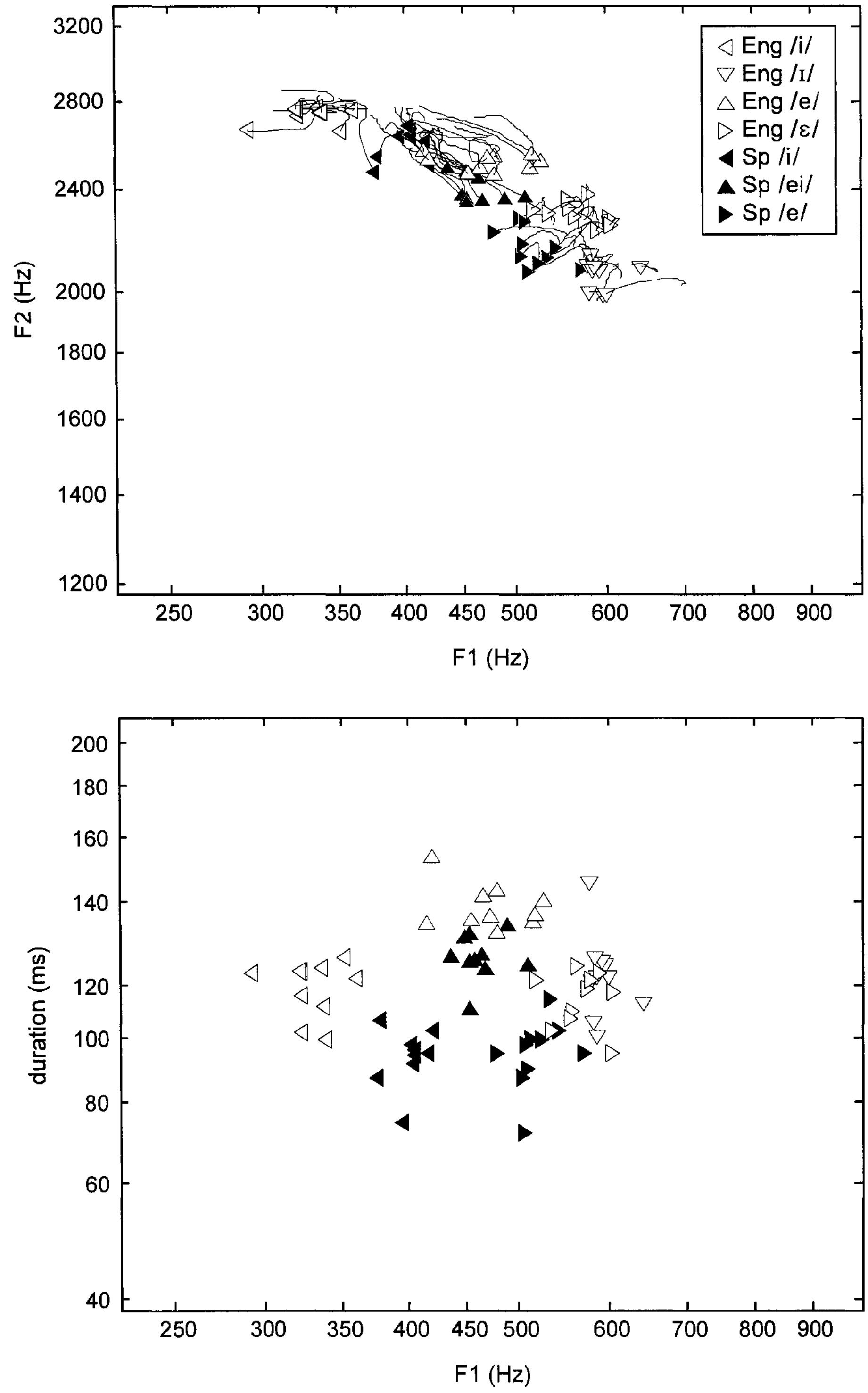


Figure 9.5 Example of Production Pattern E, English /ɪ/ produced with properties similar to Spanish /e/ (Participant bs068).

9.2 L1-English Listeners' Perception of L2-English Natural Vowel Productions

The experiment in which four of the monolingual English listeners identified natural vowel productions, Section 6, included L2-English vowel productions by bilingual L1-Spanish speakers. Each of the bilingual L1-Spanish speakers produced three instances of each English vowel category, and each of the L1-English listeners gave one response to each instance of each vowel. The L1-Spanish L2-English speakers were originally selected at random, but at least one participant represented each of the Production Patterns A, B, C, D, and E.

– Production Pattern A. The CDFA trained on L1-English speakers' productions predicted that for L1-Spanish L2-English speakers who did not produce spectral or duration differences between L2-English /i/ and /ɪ/, both these L2 vowels would be perceived as English /i/ by L1-English listeners (see Table 9.1). This prediction was borne out: A confusion matrix for L1-English listeners' perception of productions by four speakers (bs071, bs051, bs052, bs087) who produced Production Pattern A is given in Table 9.6. The L1-English listeners' perception of these bilingual L1-Spanish speakers' L2-English /i/ & /ɪ/, /e/, and /ɛ/ productions was also similar to their perception of L1-Spanish speakers' Spanish /i/, /ei/, and /e/ (see Table 5.3). These results are therefore consistent with these speakers substituting Spanish /i/ for English /i/ and /ɪ/, substituting Spanish /ei/ for English /e/, and substituting Spanish /e/ for English /ɛ/. Note, however, that the L1-English production model underestimated the percentage of L2-English /i/ and /ɪ/ productions which would be identified as English /ɪ/ by L1-English listeners.

Table 9.6 Confusion matrix for four monolingual English listeners' identification of natural-L2-English-vowel stimuli produced by two male and two female L1-Spanish L2-English speakers who produced Production Pattern A. Results pooled over speakers and listeners, and expressed as percentages (rows sum to 100, blank cells have a value of zero).

Produced	Perceived			
	Eng /i/	Eng /ɪ/	Eng /e/	Eng /ɛ/
Eng /i/	97.9	2.1		
Eng /ɪ/	95.8	4.2		
Eng /e/			100.0	
Eng /ɛ/		39.6	16.7	43.8

– Production Pattern B. The L1-English production model predicted that for L1-Spanish L2-English speakers who produced duration differences between L2-English /i/ and /ɪ/, both these L2 vowels would usually be perceived as English /i/ by L1-English listeners, but some L2-English /i/ productions would be perceived as English /e/ (see Table 9.2). A confusion matrix for L1-English listeners' perception of productions by six speakers (bs072, bs083, bs058, bs016, bs028, bs077) who produced Production Pattern B is given in Table 9.7. The prediction that most instances of L2-English /i/ and /ɪ/ would be perceived as English /i/ was borne out, but no instances of L2-English /i/ were perceived as English /e/. Comparing Tables 9.6 and 9.7, it appears that producing duration differences between L2-English /i/ and /ɪ/ did not make it easier for L1-English listeners to correctly identify L2-English /ɪ/.

Table 9.7 Confusion matrix for four monolingual English listeners' identification of natural-L2-English-vowel stimuli produced by three male and three female L1-Spanish L2-English speakers who produced Production Pattern B. Results pooled over speakers and listeners, and expressed as percentages (rows sum to 100, blank cells have a value of zero).

Produced	Perceived			
	Eng /i/	Eng /ɪ/	Eng /e/	Eng /ɛ/
Eng /i/	97.2	2.8		
Eng /ɪ/	95.8	4.2		
Eng /e/	1.4		98.6	
Eng /ɛ/		23.6	8.3	68.1

– Production Pattern C. The L1-English production model predicted that for L1-Spanish L2-English speakers who produced spectral differences between English /i/ and /ɪ/, some instances of English /ɪ/ would be correctly perceived by L1-English listeners (see Table 9.3). This prediction was borne out: A confusion matrix for productions by four speakers (bs067, bs059, bs063, bs086) who produced Production Pattern C is given in Table 9.8.

Table 9.8 Confusion matrix for four monolingual English listeners' identification of natural-L2-English-vowel stimuli produced by one male and three female L1-Spanish L2-English speakers who produced Production Pattern C. Results pooled over speakers and listeners, and expressed as percentages (rows sum to 100, blank cells have a value of zero).

Produced	Perceived			
	Eng /i/	Eng /ɪ/	Eng /e/	Eng /ɛ/
Eng /i/	100.0			
Eng /ɪ/	37.5	50.0	10.4	2.1
Eng /e/			100.0	
Eng /ɛ/		52.1	2.1	45.8

– Production Pattern D. The L1-English production model predicted that for L1-Spanish L2-English speakers who produced spectral and duration differences between English /i/ and /ɪ/, most instances of English /ɪ/ would be correctly perceived by L1-English listeners (see Table 9.4). This prediction was partially borne out: A confusion matrix for productions by three speakers (bs019, bs023, bs078) who produced Production Pattern D is given in Table 9.9. Half the L1-Spanish L2-English speakers L2-English /ɪ/ productions were correctly perceived, the same percentage as for participants who produced only spectral differences.

Table 9.9 Confusion matrix for four monolingual English listeners' identification of natural-L2-English-vowel stimuli produced by one male and three female L1-Spanish L2-English speakers who produced Production Pattern D. Results pooled over speakers and listeners, and expressed as percentages (rows sum to 100, blank cells have a value of zero).

Produced	Perceived			
	Eng /i/	Eng /ɪ/	Eng /e/	Eng /ɛ/
Eng /i/	97.2	2.8		
Eng /ɪ/	44.4	50.0	5.6	
Eng /e/			100.0	
Eng /ɛ/		22.2	22.2	55.6

– Production Pattern E. The L1-English production model predicted that for L1-Spanish L2-English speakers who produced English /ɪ/ with spectral and duration properties similar to Spanish /e/, some instances of English /ɪ/ would be correctly perceived by L1-English listeners and some would be perceived as English /ɛ/ (see Table 9.5). This prediction was borne out: A confusion matrix for productions by one speaker (bs057) who produced Production Pattern E is given in Table 9.10.

Table 9.10 Confusion matrix for four monolingual English listeners' identification of natural-L2-English-vowel stimuli produced by one female L1-Spanish L2-English speaker who produced Production Pattern E. Results pooled over speakers and listeners, and expressed as percentages (rows sum to 100, blank cells have a value of zero).

Produced	Perceived			
	Eng /i/	Eng /ɪ/	Eng /e/	Eng /ɛ/
Eng /i/	100.0			
Eng /ɪ/		50.0		50.0
Eng /e/			100.0	
Eng /ɛ/		83.3		16.7

In general, the L1-English listeners' perception of L1-Spanish L2-English speakers' natural vowel productions confirmed the conclusions made on the basis of the L1-English production model.

9.3 Comparison of Production Patterns and Hypothesised Developmental Paths

Individual L1-Spanish L2-English participants production patterns were compared with their perception patterns in order to assess whether their production patterns were consistent with their assignment to the direct or indirect developmental path and to stages on the indirect developmental path. Assignments to stages had been made purely on the basis of perception result, and assignment to Production Patterns was made purely on the basis of production results.[1] If production results are consistent with the assignments to developmental paths and stages, then they will support the developmental paths as hypothesised.

Table 9.11 provides a summary of the assignment of L1-Spanish L2-English participants to different production patterns and comparison with their assignment to hypothesised developmental paths and stages on the hypothesised indirect developmental path. In general, production patterns were consistent, or at least not inconsistent, with perception patterns. Results will be discussed in the order provided by the developmental paths and stages on the hypothesised indirect developmental path, with the exception of Production Pattern E which will be discussed last.

[1] Since assignments to production patterns were made after the assignments to perception paths and stages, knowledge of perception category could potentially have influence the assignment to perception stage in borderline cases. Borderline cases are therefore flagged in Table 9.11. Only in one case, bs058, would the decision on a borderline case have affected whether the production pattern was consistent or inconsistent with the perception pattern.

Table 9.11 Comparison of L2-English production patterns with assignment to hypothesised developmental paths and stages on the hypothesised indirect developmental path. Ticks: Production pattern is consistent with the developmental path and stage. Crosses: Production pattern is inconsistent with the developmental path and stage.

		Production Pattern							Production Pattern				
Stage	Participant	A	B	C	D	E	Stage	Participant	A	B	C	D	E
0	bs071	✓					2	bs062	✓				
	bs114	✓						bs019				✓[4]	
	bs088	✓[1]						bs049	✓				
	bs065	✓						bs056	✓				
½	bs050				✗		3	bs067			✓		
	bs076	✓						bs028		✓[5]			
	bs051	✓						bs073		✓			
	bs069					✗[2]		bs023				✓	
	bs052	✓						bs077		✓			
	bs064	✓						bs057					✗
	bs001	✓					direct path	bs061		✗			
	bs075	✓						bs082		✗			
	bs072		✓					bs002			✓[6]		
	bs003				✗			bs074	✓				
	bs087	✓						bs068					✗
1	bs083		✓					bs078				✗	
	bs081		✓[3]					bs059			✓		
	bs058		✓[4]					bs017					✗
	bs016		✓					bs063			✓		
	bs091	✓						bs086			✓[6]		

[1] L2-English productions were generally longer than L1-Spanish productions.

[2] Apparent labelling reversal, L2-English /i/ productions had Spanish-/e/-like properties.

[3] Small duration difference between L2-English /i/ and /ɪ/, borderline case between Patterns A and B.

[4] Large duration differences and small spectral difference between L2-English /i/ and /ɪ/, borderline case between Patterns B and D.

[5] L2-English /ɪ/ had lower F2 but similar F1 to L2-English /i/, borderline case between Patterns B and D.

[6] Small spectral difference between L2-English /i/ and /ɪ/, borderline case between Patterns A and C.

Hypothesised indirect developmental path:

– Stage 0. Participants assigned to this stage had negligible ability to perceptually distinguish English /i/ and /ɪ/, and also produced no or negligible difference between the two L2-English vowels – they were all assigned to Production Pattern A. Production results for all participants assigned to this stage were therefore consistent with perception results.

– Stage ½. Participants assigned to this stage were hypothesised to perceptually distinguish English /i/ and /ɪ/ via a category-goodness-difference assimilation to Spanish /i/, and to label less-Spanish-/i/-like English vowels (vowels with longer duration, higher-F1–lower-F2, or with converging VISC) as English /i/. Most participants assigned to this stage were also assigned to Production Pattern A and produced no or negligible difference between the two L2-English vowels. If production lags behind perception, then the use of spectral and duration cues in perception would not necessarily be reflected in production.[2] The production results are therefore not inconsistent with assignment of these participants to Stage ½. One participant produced duration differences between English /i/ and /ɪ/ (Production Pattern B), which was consistent with his use of duration in perception and assignment to Stage ½.

Two participants produced spectral and duration differences between English /i/ and /ɪ/ (Production Pattern D). These participants' production results were inconsistent with their perception results, perception results were consistent with Stage ½, but production results were consistent with Stage 3. Note that positing a reversal of the English /i/ and /ɪ/ labels during one of the experiments will not reconcile the results. Although the labelling of spectral cues was reversed in the perception experiment relative to the production experiment, in both the production and perception experiments English /i/ was longer, hence there was no reversal in labelling duration cues.

– Stage 1. Participants assigned to this stage used duration to perceptually distinguish English /i/ and /ɪ/. All except one participant also produced duration differences between English /i/ and /ɪ/ (Production Pattern B), and thus their production results were consistent with their perception results and assignment to Stage 1. One participant produced no or negligible difference between the two L2-English vowels (Production Pattern A), which,

[2] Production would logically be expected to lag behind perception: Production output must be based on a representation of the sound, and the representation would normally be based on perceptual input; some motor control learning may be required to make production match the perception-based representation, and the time taken to learn the motor control would result in a lag behind perception. Articulatory instruction could short-circuit this process so that production is based on an articulatory representation and perception lags behind production. Reviews of the literature (e.g., Leather, 1983, 1999; Llisterri, 1995; see also Flege, 2003) reveal a complex relationship between L2 perception and production.

allowing for a lag between perception and production, is not inconsistent with assignment to Stage 1.

– Stage 2. Participants assigned to this stage used both spectral and duration differences to perceptually distinguish English /i/ and /ɪ/, but did not make strong use of VISC differences. Only one participant produced spectral and duration differences, and their spectral separation between English /i/ and /ɪ/ was borderline (borderline between Production Patterns B and D). The remaining participants assigned to this stage produced no or negligible difference between the two L2-English vowels (Production Pattern A). Allowing for a lag between perception and production, none of the participants had production patterns which were inconsistent with their perception patterns or with assignment to Stage 2.

–Stage 3. Participants assigned to this stage made L1-English-like use of spectral and duration properties, including VISC, to perceptually distinguish English /i/ and /ɪ/. One participant produced L1-English-like spectral and duration differences between English /i/ and /ɪ/ (Production Pattern D), although no VISC difference. Three participants produced duration differences only (Production Pattern B), and one produced spectral differences only (Production Pattern C). Therefore, allowing for a lag between perception and production, all of the participants (except one with Production Pattern E discussed below) had production patterns which were not inconsistent with their perception patterns or with assignment to Stage 3. Production Pattern B (duration differences only) is consistent with a production lag for participants who have reached Stage 3 via the indirect developmental path, and Production Pattern C (spectral differences only) is consistent with a production lag for participants who have reached Stage 3 via the direct developmental path.

Direct developmental path:

Participants assigned to the hypothesised direct developmental path were hypothesised to perceptually distinguish English /i/ and /ɪ/ via a category-goodness-difference assimilation to Spanish /i/, and to label less-Spanish-/i/-like English vowels (vowels with higher-F1–lower-F2 and converging VISC, and to a lesser extend longer duration) as English /ɪ/. Four of the ten participants assigned to the hypothesised direct developmental path produced English /ɪ/ with higher F1 and lower F2 than English /i/ (Production Pattern C, two participants actually produced relatively small spectral differences and were borderline cases between Production Patterns A and C). Although they did not produce VISC differences, these participants' production patterns were otherwise consistent with their perception patterns and with assignment to the hypothesised direct developmental path. One participant produced no or negligible difference between English /i/ and /ɪ/ (Production Pattern A), which, allowing for a lag between perception and production, was not inconsistent with his perception patterns or with assignment to the hypothesised direct developmental path.

The results for the two participants with borderline A–C Production Patterns and the one participant with production Pattern A, English /i/ and /ɪ/ both produced with acoustic properties similar to Spanish /i/, would only be consistent with the hypothesised direct developmental path being the result of a category-goodness-difference-assimilation to Spanish /i/, and not the alternative of a category-goodness-difference-assimilation to Spanish /e/. None of these production patterns would necessarily be inconsistent with Stage 3, with boundary angles only just beyond –90° the use of duration cues could be regarded as effectively zero. However, it should be noted that the only participants assigned to Production Pattern C, use of spectral cues only, were participants who on the basis of perception results were assigned to the hypothesised direct developmental path and one participant assigned to Stage 3. A participant who is on the indirect developmental path, is hypothesised to use duration in the L1-English direction before using spectral properties in the L1-English direction, therefore if their production lags behind their perception, they would be expected to use either duration alone or duration and spectral properties. In contrast, a participant on the direct hypothesised developmental path is hypothesised to immediately use spectral difference in the L1-English direction, and make little use of duration cues; hence they might produce only spectral differences between English /i/ and /ɪ/, a pattern not expected for participants on the hypothesised indirect developmental path. In summary, the production and perception results from these participants were consistent with the hypothesised direct developmental path being the result of a category-goodness-difference-assimilation to Spanish /i/.

One participant produced L1-English-like spectral and duration differences between English /i/ and /ɪ/ (Production Pattern D), although no VISC difference. This production pattern was consistent with Stage 3 and not inconsistent with this participant's perception pattern: with an English /i/–/ɪ/ boundary angle of –93.2°, use of duration in perception was not significant: (*i*-*ɪ*)×dur contrast coefficient value –0.069, Wald $\chi^2(1) = 0.683, p = .409$. L1-English listeners also produce duration differences even though they do not use them in perception. The production and perception results would therefore not be inconsistent with this participant being at Stage 3, L1-English-like use of spectral and duration properties, reached via the hypothesised direct developmental path.

Two participants produced duration differences only between English /i/ and /ɪ/ (Production Pattern B). These participants' production patterns were not consistent with their perception patterns, but could be made consistent if a labelling reversal between English /i/ and /ɪ/ were posited for one of the experiments.[3] If a labelling reversal were posited for the production experiment, the results would then be consistent with assignment of these participants to the direct developmental path. If a labelling reversal were posited for the perception experiment, the results would then be consistent with reassignment of these participants to Stage ½ of the hypothesised indirect developmental path.

Production Pattern E:

This production pattern, English /i/ produced with Spanish-/i/-like properties and English /ɪ/ and /ɛ/ produced with Spanish-/e/-like properties, is consistent with the a priori predictions made for the initial state of L2-English learning on the basis of L1 production and perception models (see Sections 3.5 and 4.3.2). A total of four participants produced this pattern. One had been assigned to Stage ½ of the hypothesised indirect developmental path, her production and perception results were inconsistent with each other and could not be made consistent by positing labelling reversals. Of the other three participants, two were assigned to the hypothesised direct developmental path and one to Stage 3. None of these participants' perception results were actually inconsistent with their perception results; hence it appears that these three participants could be assigned to a third hypothesised

[3] Posited labelling reversals are common in L2 perception research. If an L2 learners has learnt substantial vocabulary before learning to distinguish an L2 contrast, then their lexicon will include vocabulary with is underspecified with respect to the contrast. After leaning to distinguish the contrast, it may take a considerable length of time to correct the lexicon, and, more importantly for the present study, to associate the correct orthographic forms with each member of the contrast.

developmental path in which English /ɪ/ is initially assimilated to Spanish /e/ rather than Spanish /i/.

In general the production results were consistent with the assignment of participants to the hypothesised developmental paths and stages made on the basis of perception results. Thirty-one (31) of the 40 participants had production patterns consistent with their assignments to paths and stages. Two of the remaining participants had production patterns which could be made consistent with perception patterns if a labelling reversal were posited for one of the experiments, and another had production and perception patterns which were not inconsistent and who, in light of the production pattern, could arguably be reassigned to another stage. Three participants had production and perception patterns which were not inconsistent with each other, and who could be reassigned to a third hypothesised developmental path in which English /ɪ/ was initially assimilated to Spanish /e/ (as predicted by the L1 production and perception models) rather than Spanish /i/, and in which English /ɪ/ and /ɛ/ were distinguished via a category-goodness-difference-assimilation to Spanish /e/.[4] Only three participants had production patterns which were completely inconsistent with their perception patterns.

[4] Another possibility is that this could be a single-category assimilation in which instance of both L2-English vowels were perceived as poor examples of the L1-Spanish /e/ category, and were relatively easily distinguished via their acoustic differences.

10. L2-English Longitudinal Results & Discussion

Strong positive evidence in support of the hypothesised developmental paths actually being developmental would come from longitudinal results if L1-Spanish learners of English were observed to progress through the hypothesised stages. This section will proceed by giving details of the participants, the procedures used to analyse the perception results, and then a description of each participant's results.

10.1 Participants

Four L1-Spanish speakers (see Table 2.3) participated in a longitudinal study in which they completed the English versions of the production and perception tests on four separate occasions, each occasions separated by a time period of approximately two months. These tests nominally took place at lengths of residence of approximately 1, 3, 5, and 7 months. Participants' L1-Spanish perception was also tested at LORs of 1 month and 7 months. Ideally, participants would previously have never lived in an English-speaking country, and would have professions which required extensive interaction with L1-English speakers. This was the case for participant bs087; however, due to practical difficulties in recruiting participants, some compromises were made in the inclusion of the other participants.

Three participants had had some previous exposure to English outside their home countries. Participant bs075 had spent three months in Boston eight years before arriving in Edmonton. Two years before arriving in Edmonton, Participant bs083 had lived in Germany for a year where he had used English as a lingua-franca. Participant bs069 had spent four months in Boston one year before arriving in Edmonton then four months in Edmonton before the first test; however, she had had limited contact with L1-English speakers prior to her involvement in the research. Three participants satisfied the profession criterion, but one did not. Participants bs075, bs083, and bs087 were graduate students, but Participant bs069 was engaged in full time child rearing.

10.2 Analysis of Perception of Synthetic Stimuli

For each point in time, a separate logistic regression model (Model 5) was fitted to each individual participant's L2-English vowel identification responses. These models were used to produce English /i/–/ɪ/ boundary angle, magnitude, and converging-VISC contrast coefficients, and territorial maps of each listener's perception at each point in time.

In order to determine whether there were significant changes across time, for each participant, a model was fitted to all of their vowel identification data including coefficients to encode time. In addition to the terms included in Model 5, the *longitudinal model*, Model 6, also included coefficients for interactions between time and each of the bias and stimulus-tuning effects.

Model 5: $V + \mathrm{F1}\times V + \Delta\mathrm{F1}_{-}\times V + \Delta\mathrm{F1}_{+}\times V + \mathrm{dur}\times V$

Model 6: $V + \mathrm{F1}\times V + \Delta\mathrm{F1}_{-}\times V + \Delta\mathrm{F1}_{+}\times V + \mathrm{dur}\times V$
$+ \mathrm{T}\times V + \mathrm{T}\times\mathrm{F1}\times V + \mathrm{T}\times\Delta\mathrm{F1}_{-}\times V + \mathrm{T}\times\Delta\mathrm{F1}_{+}\times V + \mathrm{T}\times\mathrm{dur}\times V$

Time, T, was coded as a continuous variable (given values of 0, 1, 2, 3). Model 6 was compared with Model 5 (fitted to data across all time slices) in order to determine whether the addition of the time parameters increased goodness-of-fit. Since the tests were conducted on data from a single listener, tests were conducted directly on the ΔG^2 statistic, rather than using the quasi-likelihood procedure. Assuming pure multinomial error, ΔG^2 is asymptotically distributed as a χ^2 with degrees of freedom equal to the difference in degrees of freedom between the two models.

10.3 Results

10.3.1 Participant bs075

English /i/–/ɪ/ boundary contrast coefficient values are plotted in Figure 10.1. English /i/–/ɪ/ boundary contrast coefficient values and goodness-of-fit measures for the logistic regression models fitted at each point in time are given in Table 10.1. Territorial maps for L1-Spanish and L2-English perception at each point in time are given in Figure 10.2.

There were significant changes in Participant bs075's English perception across time:

ΔG^2 $\chi^2(15) = 126, p <$ 001. Participant bs075 was at Stage ½ of the hypothesised indirect path at LOR = 1 month,[1] and at LOR = 7 months he was clearly at Stage 2. English /i/–/ɪ/ boundary contrast coefficient values would also have placed him at Stage 2 at LOR = 3 and 5 months, although his territorial maps were not typical for that stage. From LOR = 3 months onwards, he appeared to have picked on low F1 and diverging VISC as the cues for English /i/ perception, and then gradually expanded the English /i/ region at the expense of the English /ɪ/ region. Although the changes in this participant's perception do not match the details of the canonical description of the hypothesised indirect developmental path, the results do support the hypothesis that the perception pattern at Stage ½ (hypothesised to be a category-goodness-difference assimilation to Spanish /i/) is representative of an early stage of L2-English /i/–/ɪ/ perception learning, and that L2-learners can progress from this pattern to the more L1-English-like perception pattern at Stage 2.

Participant bs075 produced Production Pattern A at each point in time, and there was no apparent trend towards differentiating L2-English /i/ and /ɪ/ in production. If one allows for a lag between perception and production, the production results are not inconsistent with the perception results.

Learning English also appears to have affected Participant bs075's L1-Spanish perception: Zero- and converging-VISC stimuli which were identified as Spanish /ei/ at LOR = 1 month, perhaps because they were poor examples of either of the other two Spanish vowels, were identified as Spanish /i/ or /e/ at LOR = 7 months. This would be consistent with the hypothesis proposed in Section 7.1, that when listening in Spanish mode, bilingual listeners may initially classify an incoming vowel as an L2 category but will conflate this category with other categories in order to give a Spanish response. This bilingual listener may have initially classified some of the stimuli as L2-English /ɛ/, but conflated the L2-English /ɛ/ and L1-Spanish /e/ categories, and given Spanish /e/ as a response. Likewise, he may have initially classified some of the stimuli as L2-English /ɪ/, but conflated the L2-English /ɛ/ and Sp+Eng /i/ diaphone categories, and given Spanish /i/ as a response. Note that the English /ɪ/–/ɛ/ boundary at LOR = 7 months is in approximately the same location

[1] For sake of readability, this direct mode of expression is used, rather than more accurate phrases such as "At LOR 1 month the participant had a perceptual pattern which was consistent with assignment to Stage ½ of the hypothesised indirect developmental path."

as the Spanish /i/–/e/ boundary. Since the stimuli affected were not typical of L1-Spanish productions, this change would not have had a disastrous effect on L1 perception.

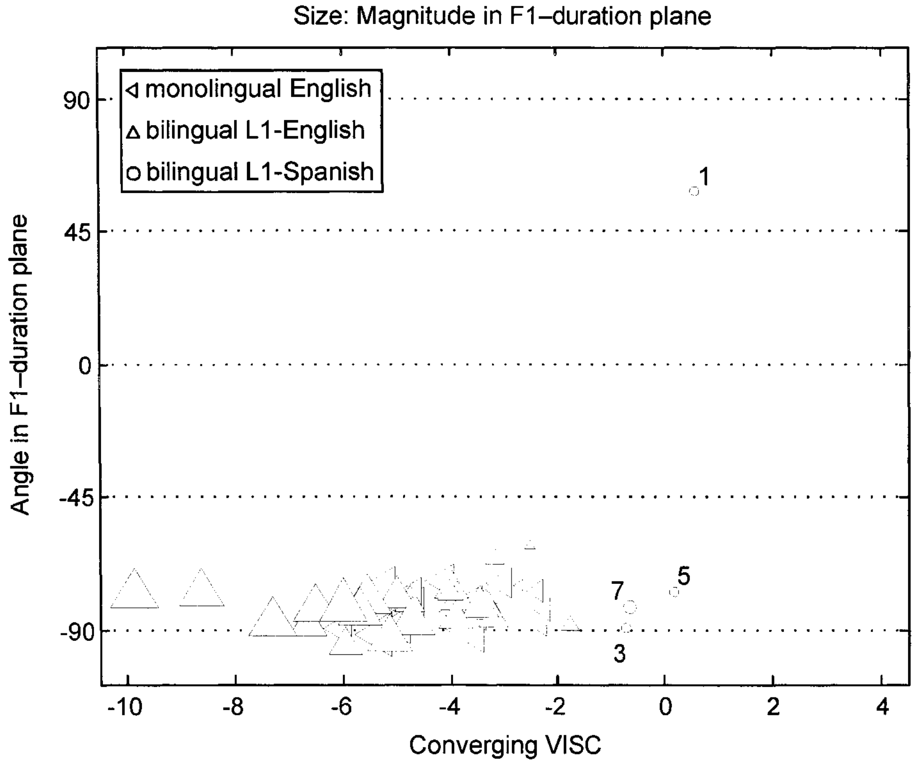


Figure 10.1 Plot of the F1–duration-plane angle and magnitude, and the converging-VISC contrast for the English /i/–/ɪ/ boundaries from logistic regression models calculated for each L1-English participant and for L1-Spanish L2-English participant bs075 at LOR = 1, 3, 5, and 7 months (numbers on plot indicate LOR in months).

Table 10.1 English /i/–/ɪ/ boundary contrast coefficient values (bias, angle and magnitude in F1–duration plane, converging-VISC contrast, diverging-VISC contrast) and goodness-of-fit measures for logistic regression models fitted to Participant bs075's English vowel identification data. Separate logistic regression models fitted at each time slice. G^2 degrees of freedom are 255.

LOR	Bias	Angle°	Mag.	con-VISC	di-VISC	G^2	%SAEP	%MA
1	–1.49	+59	0.35	0.55	–0.33	144	17.5	70.0
3	–0.06	–89	0.36	–0.73	0.37	112	15.9	85.6
5	–0.37	–77	0.26	0.15	1.64	156	17.7	81.1
7	0.96	–82	0.65	–0.67	1.87	107	12.3	88.9

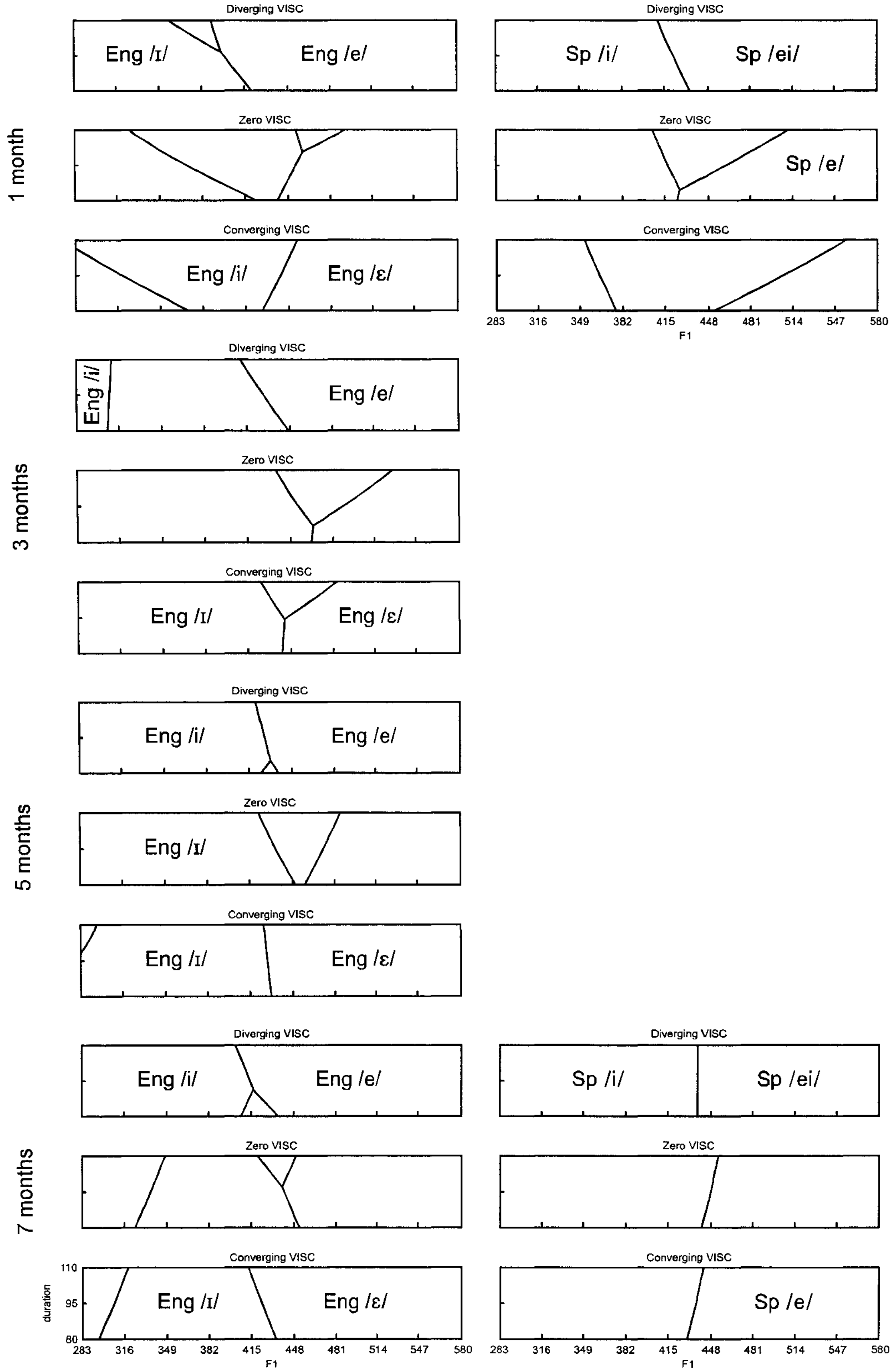


Figure 10.2 Territorial maps based on logistic regression models fitted to Participant bs075's vowel identification data at LOR = 1, 3, 5, and 7 months.

10.3.2 Participant bs083

English /i/–/ɪ/ boundary contrast coefficient values are plotted in Figure 10.3. English /i/–/ɪ/ boundary contrast coefficient values and goodness-of-fit measures for the logistic regression models fitted at each point in time are given in Table 10.2. Territorial maps for L1-Spanish and L2-English perception at each point in time are given in Figure 10.4.

There were significant changes in Participant bs083's English perception across time: $\Delta G^2\ \chi^2(15) = 36, p < 01$. Participant bs075 remained at Stage 1 of the hypothesised indirect path from LOR = 1 month to LOR = 7 months, but his general trend was to move towards more L1-English-like negative English /i/–/ɪ/ boundary angles. Although he did not move from one stage to another, the direction of movement was consistent with the hypothesised indirect developmental path.

At each point in time, Participant bs083 produced Production Pattern B, longer L2-English /i/ than L2-English /ɪ/ productions. The production results were consistent with the perception results.

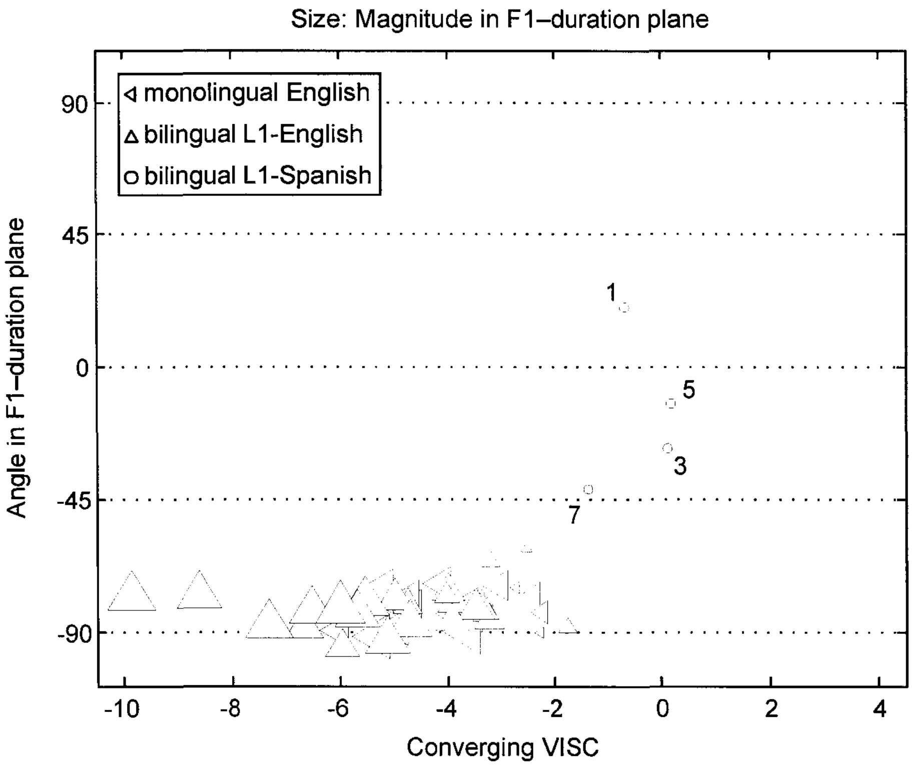


Figure 10.3 Plot of the F1–duration-plane angle and magnitude, and the converging-VISC contrast for the English /i/–/ɪ/ boundaries from logistic regression models calculated for each L1-English participant and for L1-Spanish L2-English participant bs083 at LOR = 1, 3, 5, and 7 months (numbers on plot indicate LOR in months).

Table 10.2 English /i/–/ɪ/ boundary contrast coefficient values (bias, angle and magnitude in F1–duration plane, converging-VISC contrast, diverging-VISC contrast) and goodness-of-fit measures for logistic regression models fitted to Participant bs083's English vowel identification data. Separate logistic regression models fitted at each time slice. G^2 degrees of freedom are 255.

LOR	Bias	Angle°	Mag.	con-VISC	di-VISC	G^2	%SAEP	%MA
1	−1.35	+20	0.36	−0.69	−0.17	146	14.1	90.0
3	−1.44	−28	0.56	0.09	0.11	110	10.9	88.9
5	−1.34	−13	0.50	0.16	0.43	95	9.9	91.1
7	−0.22	−41	0.44	−1.40	−0.39	108	13.6	83.3

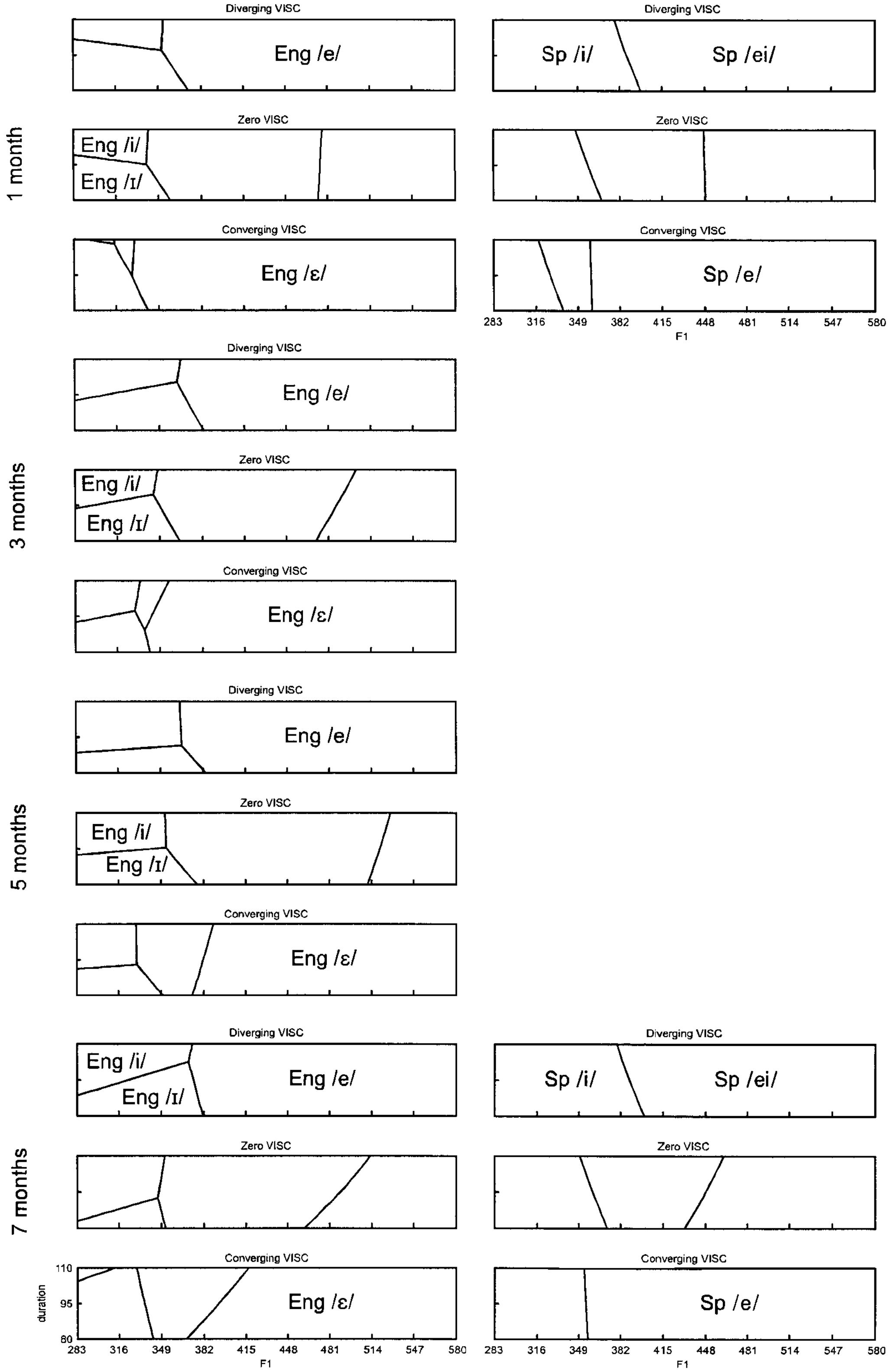


Figure 10.4 Territorial maps based on logistic regression models fitted to Participant bs083's vowel identification data at LOR = 1, 3, 5, and 7 months.

10.3.3 Participant bs087

English /i/–/ɪ/ boundary contrast coefficient values are plotted in Figure 10.5. English /i/–/ɪ/ boundary contrast coefficient values and goodness-of-fit measures for the logistic regression models fitted at each point in time are given in Table 10.3. Territorial maps for L1-Spanish and L2-English perception at each point in time are given in Figure 10.6.

There were significant changes in Participant bs087's English perception across time: ΔG^2 $\chi^2(15) = 215$, $p <$ 001. Participant bs087 remained at Stage ½ of the hypothesised indirect path from LOR = 1 month to LOR = 7 months. His general trend was to reduce his reliance on duration, and increase his reliance on spectral cues including VISC: The English /i/–/ɪ/ boundary shifted from just above –45° towards –90°, and converging-VISC contrast values increased. The magnitude of the English /i/–/ɪ/ boundary in the F1–duration plane also increased. Although and increased reliance on spectral cues and a crisper boundary is more L1-English like, because of the direction of use of the English /i/ and /ɪ/ labels, movement was actually away from L1-English norms, and opposite to the direction predicted by the hypothesised indirect developmental path.

At LOR = 1 month and 3 months Participant bs087 produced Production Pattern A, and at LOR = 5 and 7 months he produced Production Pattern B. Although the change in production pattern is consistent with the hypothesised indirect developmental path, given that, over time, Participant bs087 reduced his perceptual reliance on duration, his increased use of duration in production was unexpected.

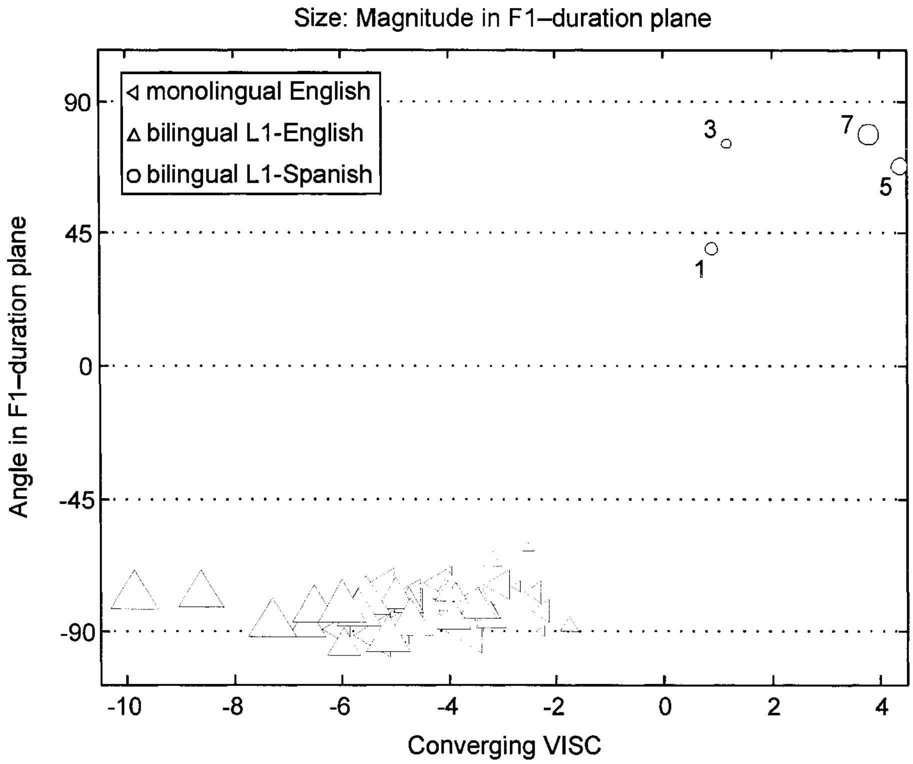


Figure 10.5 Plot of the F1–duration-plane angle and magnitude, and the converging-VISC contrast for the English /i/–/ɪ/ boundaries from logistic regression models calculated for each L1-English participant and for L1-Spanish L2-English participant bs087 at LOR = 1, 3, 5, and 7 months (numbers on plot indicate LOR in months).

Table 10.3 English /i/–/ɪ/ boundary contrast coefficient values (bias, angle and magnitude in F1–duration plane, converging-VISC contrast, diverging-VISC contrast) and goodness-of-fit measures for logistic regression models fitted to Participant bs087's English vowel identification data. Separate logistic regression models fitted at each time slice. G^2 degrees of freedom are 255.

LOR	Bias	Angle°	Mag.	con-VISC	di-VISC	G^2	%SAEP	%MA
1	–1.96	+40	0.74	0.87	–1.01	173	14.6	86.7
3	–0.77	+76	0.49	1.17	–1.19	117	10.9	90.0
5	–2.93	+69	0.91	4.37	–2.22	106	9.2	94.4
7	–3.46	+79	1.09	3.79	–1.97	152	13.4	92.2

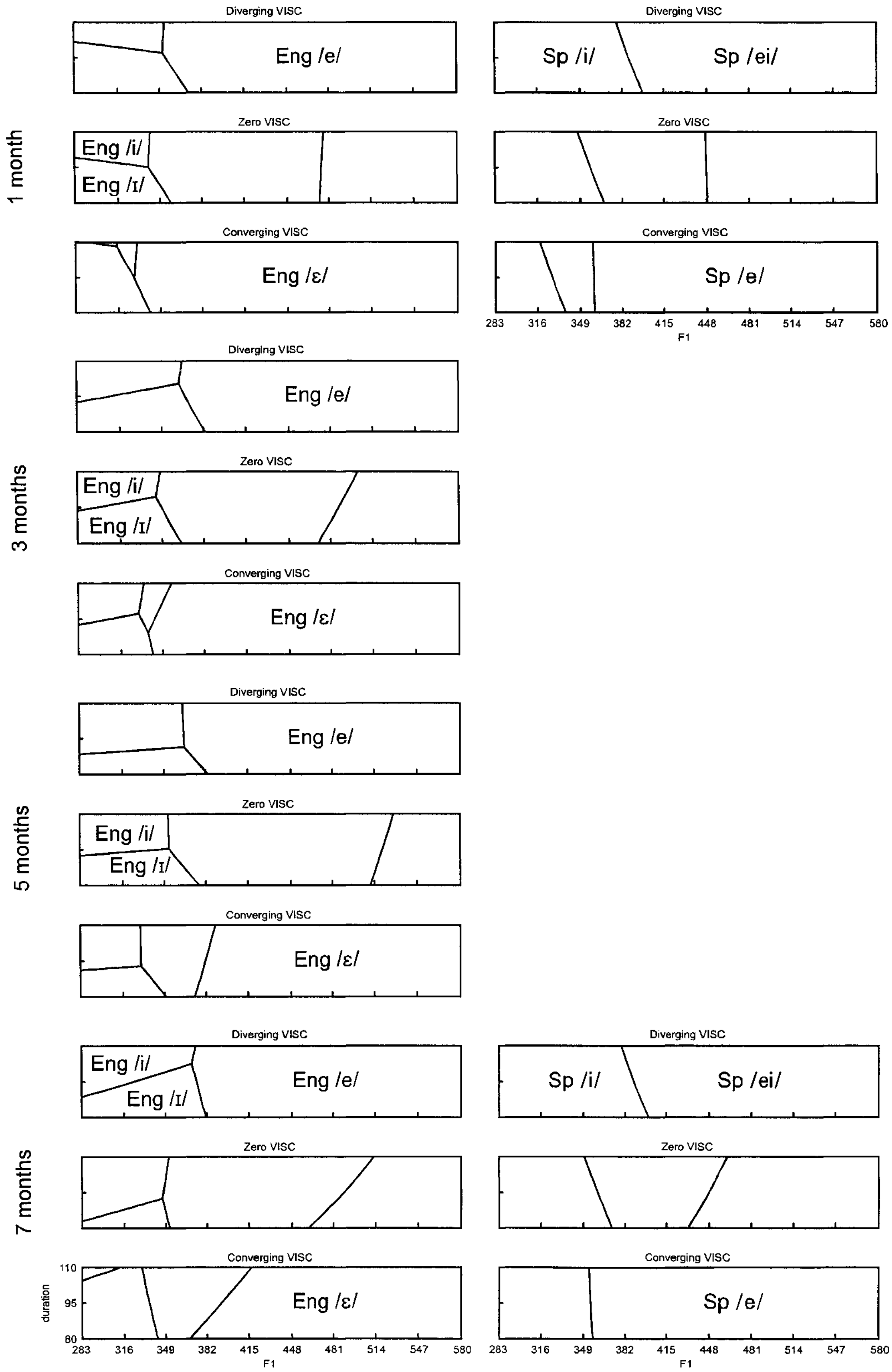


Figure 10.6 Territorial maps based on logistic regression models fitted to Participant bs087's vowel identification data at LOR = 1, 3, 5, and 7 months.

10.3.4 Participant bs069

English /i/–/ɪ/ boundary contrast coefficient values are plotted in Figure 10.7. English /i/–/ɪ/ boundary contrast coefficient values and goodness-of-fit measures for the logistic regression models fitted at each point in time are given in Table 10.4. Territorial maps for L1-Spanish and L2-English perception at each point in time are given in Figure 10.8.

There were significant changes in Participant bs069's English perception across time: $\Delta G^2\ \chi^2(15) = 72, p < 001$. Participant bs069 remained at Stage ½ of the hypothesised indirect path from LOR = 1 month to LOR = 7 months. Her perception results were similar to those of bs087 in that she had a trend towards greater use of spectral cues, but because of the direction of English /i/ and /ɪ/ labelling, the movement was away from L1-English norms.

A notable change across time for Participant bs069, was the her English /e/–/ɛ/ boundary became shallower, she relied more on duration and less on spectral cues. The same change was seen in her Spanish /ei/–/e/ boundary.

At LOR = 1 month Participant bs069 produced Production Pattern E, and at LOR = 3, 5 and 7 months she produced Production Pattern A. As noted in Section 9.3, her initial production pattern was inconsistent with her initial perception pattern.

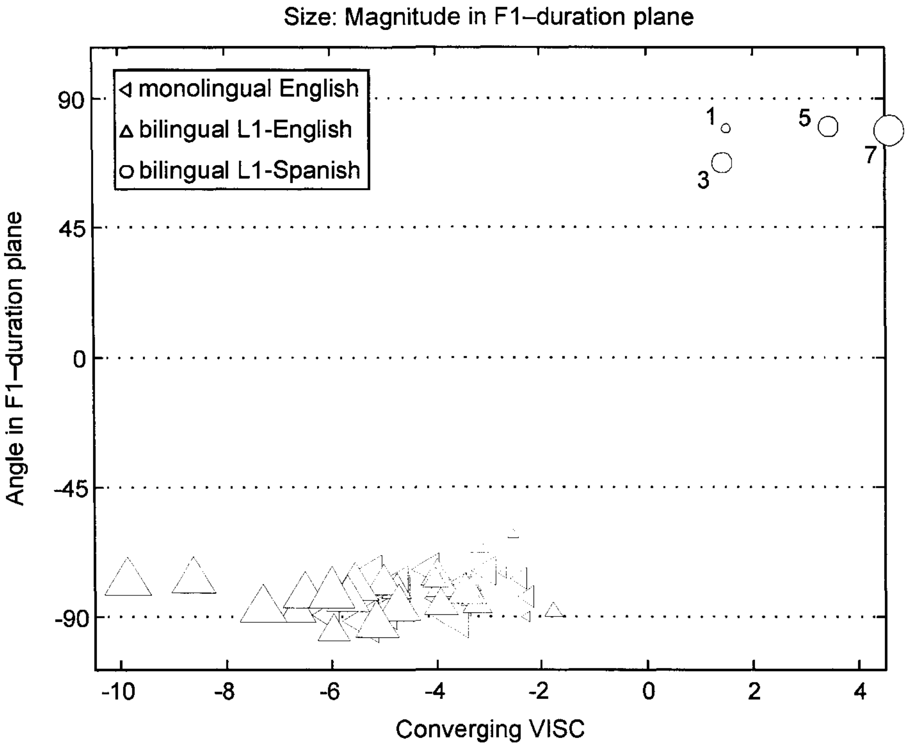


Figure 10.7 Plot of the F1–duration-plane angle and magnitude, and the converging-VISC contrast for the English /i/–/ɪ/ boundaries from logistic regression models calculated for each L1-English participant and for L1-Spanish L2-English participant bs069 at LOR = 1, 3, 5, and 7 months (numbers on plot indicate LOR in months).

Table 10.4 English /i/–/ɪ/ boundary contrast coefficient values (bias, angle and magnitude in F1–duration plane, converging-VISC contrast, diverging-VISC contrast) and goodness-of-fit measures for logistic regression models fitted to Participant bs069's English vowel identification data. Separate logistic regression models fitted at each time slice. G^2 degrees of freedom are 255.

LOR	Bias	Angle°	Mag.	con-VISC	di-VISC	G^2	%SAEP	%MA
1	−1.52	+80	0.54	1.48	−0.70	132	15.1	81.1
3	−2.50	+68	1.20	1.43	−1.68	94	13.3	86.7
5	−2.65	+81	1.10	3.44	−0.96	115	14.8	85.6
7	−3.38	+79	1.75	4.60	−0.83	107	12.6	86.7

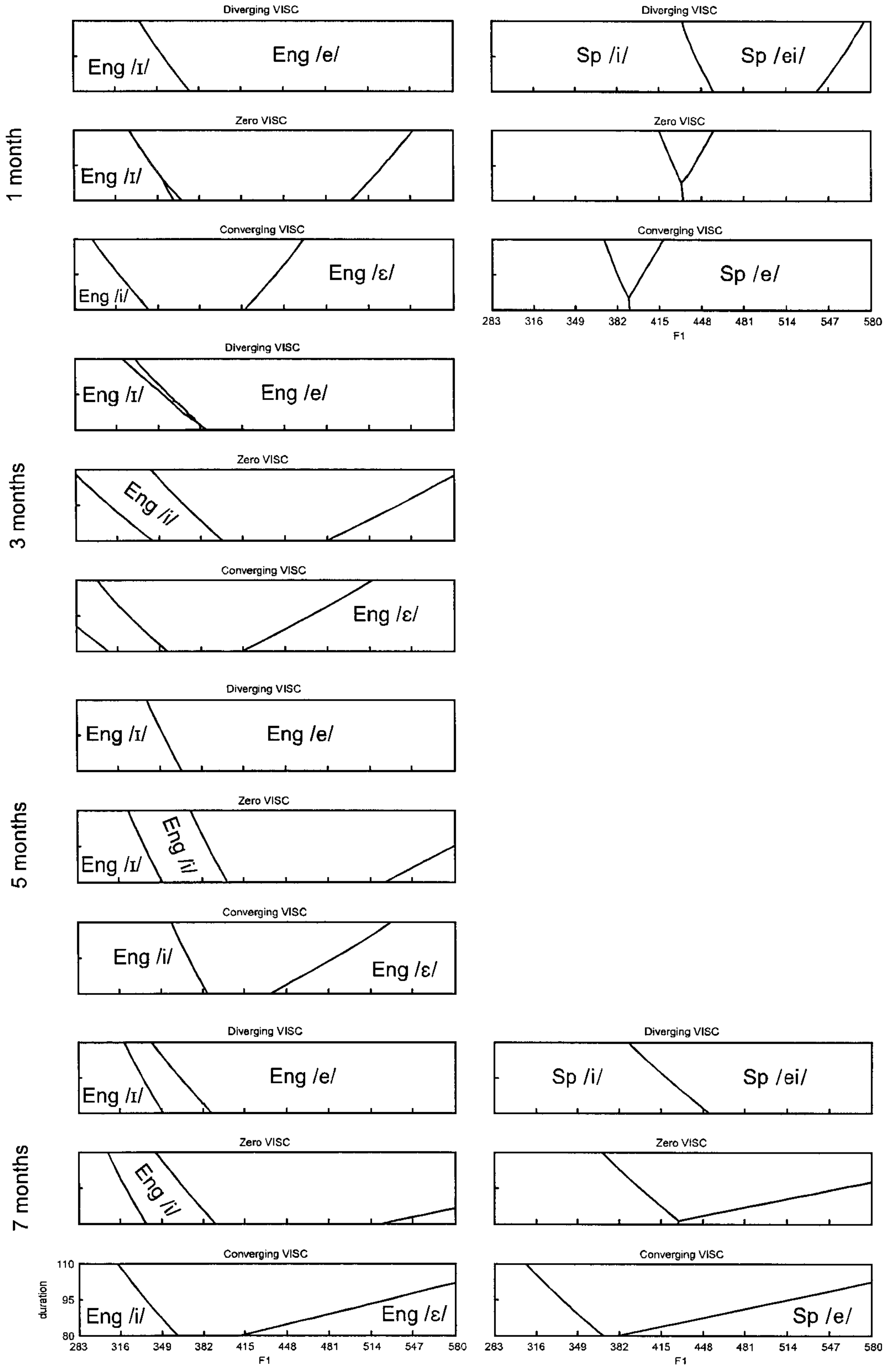


Figure 10.8 Territorial maps based on logistic regression models fitted to Participant bs069's vowel identification data at LOR = 1, 3, 5, and 7 months.

10.4 Discussion

All the L1-Spanish L2-English participants in the longitudinal case studies exhibited a trend towards making greater use of spectral cues and less use of duration cues in distinguishing the English /i/–/ɪ/ contrast. Over the course of the longitudinal study, one participant moved from a perception pattern consistent with Stage ½ of the hypothesised indirect developmental path (category-goodness-difference assimilation of English /i/ and /ɪ/ to Spanish /i/) to a perception pattern consistent with Stage 2 (use of duration and spectral properties in the same direction as L1-English listeners but with weaker boundary magnitudes). Another participant had perception patterns consistent with him remaining at Stage 1 of the hypothesised indirect developmental path (long vowels identified as English /i/ and short vowels as English /ɪ/) throughout the course of the longitudinal study, but with some movement towards L1-English norms (the boundary angle tilted from a slightly positive to a slightly negative angle, and VISC values moved towards L1-English values). This participant also produced longer English /i/ than English /ɪ/. Two participants had perception patterns consistent with them remaining at Stage ½ of the hypothesised indirect developmental path (category-goodness-difference assimilation of English /i/ and /ɪ/ to Spanish /i/) throughout the course of the longitudinal study. In conclusion, although there was some evidence in support of the hypothesised indirect developmental path, it is far from conclusive, and additional longitudinal studies will be needed to assess the adequacy of the hypothesised developmental paths.

11. L2-Spanish Perception and Production Results & Discussion

11.1 L2-Spanish Perception

Spanish perception data were available from a total of 85 participants, 18 monolingual Spanish speakers (see Table 2.1), 40 bilingual L1-Spanish speakers (see Table 2.3), and 27 bilingual L1-English speakers (see Table 2.4).

The full logistic regression model, $V + \text{F1}\times V + \Delta\text{F1}\times V + \text{dur}\times V$ with ΔF1 coded as three discrete levels (Model 5), was fitted to each individual participant's Spanish vowel identification data. The resulting logistic regression coefficients were entered into a principal component analysis conducted on the correlation matrix. The first two principal components accounted for 28.3% and 25.6%, cumulatively 54.0%, of the variance. A plot of the first two principal component loading scores is given in Figure 11.1. A three-dimensional plot of the first three principal component loading scores (accounting for an additional 13.1% of the variance) was also explored, but this did not lead to any additional insight. Although there was a tendency for L1-English L2-Spanish listeners to have greater first and second principal component scores, the majority of L1-English L2-Spanish listeners had scores which fell withing the range of L1-Spanish listeners' scores.

Territorial maps were plotted based on each individual L1-English L2-Spanish listener's L2-Spanish vowel identification data. As suggested by the principal component plots, L2-Spanish listeners' response patterns appeared to be much less variable than L2-English listeners response patterns. Since there was no strong evidence for subgroups, as had been the case in L2-English perception, a comparison was made between Spanish perception by L1-English L2-Spanish listeners as a group, and Spanish perception by the monolingual and bilingual L1-Spanish groups. A logistic regression model (Model 5) was fitted to all L1-English L2-Spanish listeners L2-Spanish vowel identification data. Goodness of fit measures were: $G^2(4850) = 6492$, SAEP = 5.68%, MA = 95.6%. The coefficient values are given in Table 11.1, and a territorial map is given in Figure 11.2.

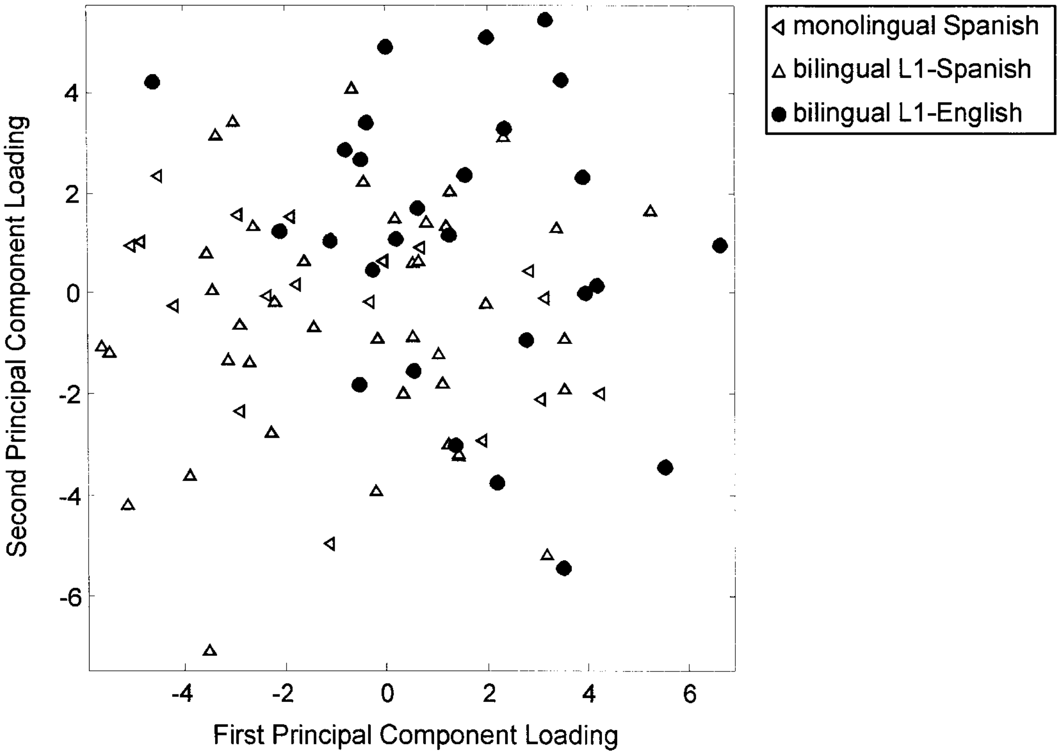


Figure 11.1 First and second principal component loading scores from the principal component analysis conducted on the sets of coefficients from the logistic regression models fitted to each individual listener's Spanish vowel identification data.

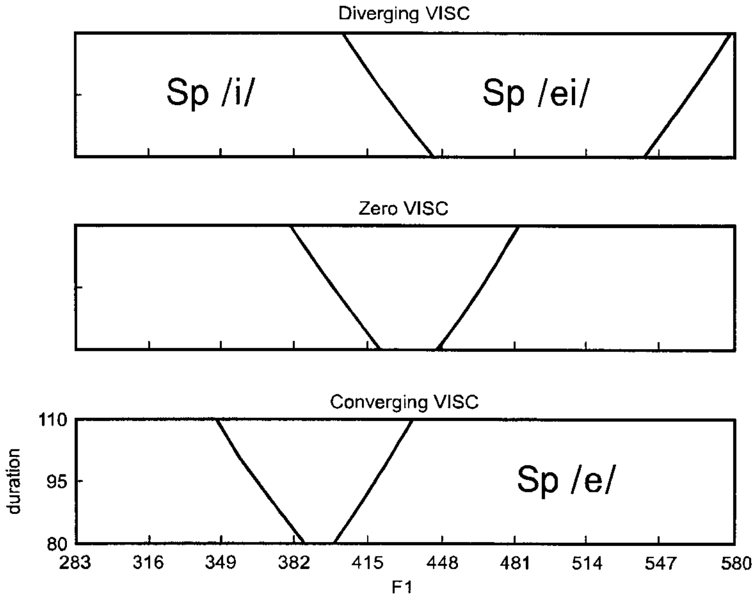


Figure 11.2 Territorial map based on classification of synthetic stimuli by the logistic regression model trained on bilingual L1-English listeners' L2-Spanish vowel identification data.

Table 11.1 Estimated coefficient values from logistic regression model 5 fitted to pooled bilingual L1-English listeners' L2-Spanish identification data.

Coefficient	Value	Standard Error	Wald χ^2	*df*	*p*
i	4.0390	0.0898	2023.71	1	.0000
ei	–0.7094	0.0626	128.53	1	.0000
e	–3.3296	0.0833	1598.84	1	.0000
V			2121.80	2	.0000
$\mathrm{F1}\times i$	–0.7996	0.0162	2450.09	1	.0000
$\mathrm{F1}\times ei$	0.1704	0.0094	325.47	1	.0000
$\mathrm{F1}\times e$	0.6292	0.0123	2609.27	1	.0000
$\mathrm{F1}\times V$			2796.99	2	.0000
$\Delta\mathrm{F1}_{+}\times i$	–1.0036	0.0593	286.40	1	.0000
$\Delta\mathrm{F1}_{+}\times ei$	0.1387	0.0413	11.26	1	.0008
$\Delta\mathrm{F1}_{+}\times e$	0.8650	0.0491	309.97	1	.0000
$\Delta\mathrm{F1}_{+}\times V$			346.48	2	.0000
$\Delta\mathrm{F1}_{-}\times i$	0.9808	0.0588	277.92	1	.0000
$\Delta\mathrm{F1}_{-}\times ei$	0.2121	0.0402	27.77	1	.0000
$\Delta\mathrm{F1}_{-}\times e$	–1.1929	0.0524	518.45	1	.0000
$\Delta\mathrm{F1}_{-}\times V$			518.61	2	.0000
$\mathrm{dur}\times i$	–0.1232	0.0099	153.77	1	.0000
$\mathrm{dur}\times ei$	0.1094	0.0069	251.30	1	.0000
$\mathrm{dur}\times e$	0.0138	0.0084	2.72	1	.0992
$\mathrm{dur}\times V$			270.33	2	.0000

The bilingual L1-English group's L2-Spanish perception was similar to that of the bilingual L1-Spanish group's L1-Spanish perception, in that, unlike monolingual Spanish listeners, they gave Spanish /ei/ as the primary response in portions of the zero- and converging-VISC subspaces (compare Figure 11.2 with Figures 4.1 and 4.2). The bilingual L1-English group's Spanish /i/–/ei/ boundary was in approximately the same location as their English /i/–/e/ boundary, but their Spanish /ei/–/e/ boundary was shifted towards lower F1 values relative to their English /e/–/ɛ/ boundary (compare Figure 11.2 with Figure 4.3, monolingual and bilingual L1-English listeners had similar L1-English perception patterns). These L2-Spanish boundary locations were consistent with the predictions made in Section 7.1 (In Figure 7.3 compare the locations of the L1-English /i/–/e/ and /e/–/ɛ/ boundaries with the bilingual Eng+Sp/i/ – Eng/e/+Sp/ei/ and Eng/e/+Sp/ei/ – Eng/ɛ/+Sp/e/ boundaries). As predicted by Flege's SLM, both bilingual L1-English and bilingual L1-Spanish listeners arrived at a perceptual pattern intermediate between those of monolingual English and monolingual Spanish listeners.

11.2 L2-Spanish Production

L2-Spanish production data were available from 26 bilingual L1-English speakers (see Table 2.4). The recordings of L2-English vowel productions were acoustically analysed using the same methodology used to analyse the L2-Spanish-vowel-production recordings.

Table 11.2 gives information on the classification of individual L1-English L2-Spanish speakers' vowels by the CDFA model trained on L1-Spanish speakers' L1-Spanish productions. Counts of classification errors represent the number of times a vowel intended as one category by the speaker was classified as another category by the CDFA model. For example, an /ei/→/e/ classification-error count of 9 indicates that 9 of the speaker's vowel productions which were intended as Spanish /ei/, were classified as Spanish /e/ by the CDFA model (10 instances of each vowel category were usually available from each speaker). The most common misclassification error was intended as Spanish /ei/ classified as Spanish /e/. Assuming that the CDFA model is highly correlated with L1-Spanish listeners perception, this error is predicted to cause the greatest confusion for L1-Spanish listeners.

Table 11.2 Classification of bilingual L1-English speakers' Spanish vowel productions by the CDFA model trained on L1-Spanish speakers' vowel productions. For counts of classification errors, the vowel symbol to the left of the arrow was the speaker's intended vowel, and the vowel symbol to the right was the CDFA model's classification of that vowel (10 instances of each vowel were usually available from each participant). Solid boxes: L2-Spanish /ei/ produced with greater VISC than L1-English /e/.
Dotted boxes: L2-Spanish /ei/ produced with less VISC than L1-English /e/.

Participant	LOR	% Correctly Classified	Counts of Classification Errors				
			/i/→/e/	/ei/→/i/	/ei/→/e/	/e/→/ei/	/e/→/i/
be009	0	36.7	9		9	1	
be021	1	50.0	7	4	4		
be014	0	65.5			10		
be020	1	65.5			10		
be012	0	66.7			10		
be024	0	66.7			10		
be089	2	66.7			10		
be079	0.33	70.0				9	
be084	0	76.7	7				
be053	0	80.0			6		
be005	2	83.3				4	
be004	0.5	86.7			3		
be010	0	90.0			2		
be119	2	93.3			2		
be090	3	93.3	2				
be027	0	96.7				1	
be055	0	96.7					1
be080	0	96.7			1		
be085	0	100.0					
be029	0	100.0					
be093	0	100.0					
be094	1	100.0					
be011	1.5	100.0					
be022	2	100.0					
be006	2	100.0					
be054	2	100.0					

Eleven (11) of the 26 bilingual L1-English speakers (be004, be005, be006, be022, be029, be054, be055, be079, be085, be093, be094) produced Spanish /ei/ which were longer and had greater VISC magnitude than their English /e/ productions. Three (3) other participants also produced large magnitude VISC.[1] This result is consistent with the prediction made in Section 4.3.2 on the basis of the L1 production models that L1-English L2-Spanish speakers would perceive Spanish /ei/ as somewhat similar to English /e/, but would notice that the Spanish vowel had exaggerated VISC and learn to produce Spanish /ei/ with greater VISC than English /e/. Producing long VISC for Spanish /ei/ was correlated with LOR, of all the participants who produced large magnitude VISC for Spanish /ei/ only three had not lived in a Spanish speaking country. This result was consistent with the prediction made in Section 7.1 that L1-English learners of Spanish would initially form a Spanish /ei/ – English /e/ diaphone, but would eventually develop a new L2-Spanish /ei/ category. In Table 11.2, the classification results for these participants are highlighted in green. Another three (3) participants (be010, be027, be084) had production patterns consistent with substitution of English /e/ for Spanish /ei/.

Nine (9) participants (be009, be012, be014, be020, be021, be024, be053, be080, be089), produced L2-Spanish /ei/ with less VISC and shorter duration or higher F1 than their L1-English /e/ productions. This result is not consistent with the predicted formation of a Spanish /ei/ – English /e/ diaphone. In Table 11.2, the classification results for these participants are highlighted in pink. Only three of these participants had lived in a Spanish-speaking country. If these participants had assimilated a large proportion of instances of Spanish /e/ to English /e/ (note that the phoneme frequency of Spanish /e/ is much greater than Spanish /ei/), then this might account for their relatively Spanish-/e/-like productions.

Turning to other vowels: Eleven (11) of the 26 bilingual L1-English speakers (be004, be005, be014, be022, be053, be055, be079, be085, be089, be093, be094) produced Spanish /i/ which were identical or very similar to their English /i/ productions. This result is consistent with the prediction made in Section 4.3.2 on the basis of the L1 production models that Spanish /i/ would be perceived as identical or highly similar to English /i/, and that

[1] Participant be119 produced greater VISC for most instances of Spanish /ei/, but not greater duration. Participant be011 produced little difference between English /e/ and Spanish /ei/, but her English /e/ were long and had a large VISC magnitude. Participant be090 appears to have produced /ai/ rather than /ei/.

English /i/ would therefore be substituted for Spanish /i/ in production. Twelve (12) speakers (be006, be010, be011, be012, be020, be024, be027, be029, be054, be080, be090, be119) produced shorter Spanish /i/ than English /i/ and no or relatively small spectral differences. These participants appear to have noticed the duration difference and learnt to produce a shorter vowel for Spanish /i/.

Nineteen (19) of the 26 bilingual L1-English speakers (be006, be010, be011, be014, be021, be024, be027, be029, be053, be054, be055, be080, be084, be085, be089, be093, be094, be119) produced Spanish /e/ which were intermediate between their English /ɛ/, /ɪ/, and /e/ productions. This result is consistent with the prediction made in Section 7 on the basis of the L1 mega-model and category-goodness data, that L1-English L2-Spanish speakers would perceive Spanish /e/ as a new vowel and develop L1-Spanish-like production. The first 16 participants listed above produced Spanish /e/ in a position close to the L1-Spanish norm (see Figure 3.1), but arguably with deflection away from English /ɪ/.

Three (3) bilingual L1-English speakers (be010, be022, be053) produced Spanish /e/ which were consistent with assimilation to, and substitution of, English /ɛ/. Three (3) bilingual L1-English speakers (be0042, be0050, be079) produced Spanish /e/ which were consistent with assimilation to, and substitution of, English /e/.

The production results were generally consistent with the a priori predictions made in Sections 4.3.2 and 7 on the basis of L1 production and perception data; the notable discrepancy being that a third of the L1-English learners of Spanish had L2-Spanish /ei/ productions which had less VISC than their L1-English /e/ productions, which was not consistent with the predicted formation of a Spanish /ei/ – English /e/ diaphone.

12. Summary

12.1 Focus and Rationale

The core aim of the present study was to obtain a better understanding of L1-Spanish speakers' learning of the English /i/–/ɪ/ contrast. L1-Spanish learners of English typically have difficulty perceiving and pronouncing this English vowel contrast, and are aware of this as a problem which interferes with their ability to communicate with L1-English speakers. The present study expanded on earlier research in several ways:

– Earlier studies typically investigated L2 vowel perception or L2 vowel production, the present study investigated both in order to test theories developed on the basis of perception results by assessing whether they were compatible with production results.

– Earlier studies, typically focussed on two acoustic dimensions, duration and steady-state spectral properties. In many North American dialects of English, including General Canadian English which was the dialect investigated in the present study, /ɪ/ is produced with converging diphthongisation / vowel inherent spectral change (VISC), and VISC has been found to be an important factor in vowel perception. The present study therefore analysed VISC in acoustic vowel production data, and included VISC as a dimension in the synthetic-speech continuum used in perception experiments.

– Earlier studies typically had two test combinations, L1-English listeners' perception of the English vowels, and L1-Spanish L2-English listeners' perception of the English vowels; however, in order to understand whether L2 perception results are due to modification of the perception system as a product of L2 learning, or whether they are due to transfer of L1 perception, it is necessary to know how L1-Spanish L2-English listeners would perceive the same stimuli in terms of their L1-Spanish categories. In production, in order to determine whether L1-Spanish L2-English speakers substitute L1-Spanish vowels for L2-English vowels, or whether they have learnt new pronunciations as a product of L2 learning, it is necessary to understand how L1-Spanish L2-English speakers pronounce L1-Spanish vowels. The present study therefore investigated L1-Spanish L2-English speakers' perception and production of Spanish as well as English vowels. In order to understand L2 speech learning, it is necessary to know the initial state for L2 learning. L1-Spanish L2-

English speakers' perception and production of Spanish may be affected by learning English, the present study therefore also investigated monolingual Spanish speakers' perception and production of Spanish vowels.

– Earlier synthetic-speech studies typically focussed on the two-way contrast between English /i/ and /ɪ/; however, these vowels may be confused with other adjacent English vowel categories as well as with each other, and may be assimilated to more than one Spanish vowel category. The present study therefore investigated the non-low front vowels of English and Spanish: English /i/, /ɪ/, /e/, /ɛ/, and Spanish /i/, /ei/, /e/.

– Numerous studies have investigated L1-Spanish speakers' perception or production of English vowels, but L1-English learners of Spanish often have pronunciation problems, and few studies have investigated L1-English speakers perception or production of Spanish vowels. The present study also investigated monolingual English speakers' English vowel perception and production, and L1-English L2-Spanish speakers' English and Spanish vowel perception and production.

12.2 L1 Production and Perception

Initial analyses of the acoustic properties of L1-Spanish and L1-English vowel productions (Section 3.1, see Figure 3.1 for a graphical summary) indicated that Spanish /i/ and /e/ had negligible VISC, and Spanish /ei/ had substantial diverging VISC, with initial and final formant values relatively close to Spanish /e/ and /i/ respectively, Spanish /ei/ was also substantially longer than Spanish /i/ and /e/. Canadian English /i/ was produced as a monophthong, English /ɪ/ and /ɛ/ had converging VISC, and English /e/ had diverging VISC, but initial and final formant values were not close to any other English vowels. English /ɪ/ was shorter than English /i/ and /ɛ/, and English /e/ was longer. English /ɪ/ had higher F1 and lower F2 values than were expected on the basis of earlier studies on Canadian English vowel production. This was most likely due to differences in consonant context effects, although the possibility of a diachronic change was also investigated.

Spanish /i/ was close to English /i/ in terms of spectral properties, but intermediate between English /i/ and English /ɪ/ in terms of duration. Spanish /e/ was almost identical to English /ɪ/ in terms of initial formant values, but English /ɪ/ was shorter and had converging rather than zero VISC. Spanish /ei/ was similar to English /e/ in terms of initial formant values, but Spanish /ei/ had substantially greater VISC magnitude and was substantially longer than English /e/.

In order to predict how instances of English vowels would be assimilated to Spanish vowel categories and vice versa, canonical discriminant function analysis (CDFA) models were trained on L1-Spanish and L1-English vowel production data, and then the L1-Spanish model was used to classify English vowel productions and vice versa (Sections 3.2–3.4). The models were trained on vowel duration, and initial and final F1 and F2 values (dual-target parameterisation of VISC). The CDFA production models classified incoming vowels on the basis of the statistical distribution of the acoustic properties of the vowel productions in the L1 vowel categories on which they were trained. The general principal is therefore compatible with current theories of L1 speech learning which posit that listeners establish and refine L1 speech sound categories on the basis of the statistical distribution of acoustic properties in the language to which they are exposed (see Section 1.2.1). The CDFA production models were highly successful at correctly classifying vowels from the training language. The L1-English model was also highly correlated with monolingual English listeners' perception of natural productions of English and Spanish vowels (Section 6, monolingual Spanish listeners' perception of natural vowel productions were not available). The CDFA models were therefore used to make predictions as to listeners' perception of L2 vowels.

The L1-Spanish CDFA production model classified almost all instances of English /i/ as Spanish /i/, and almost all instances of English /ɪ/ as Spanish /e/. At the initial state for L2-English learning L1-Spanish listeners were therefore predicted to assimilate most instances of English /i/ to Spanish /i/, and most instances of English /ɪ/ to Spanish /e/, and, in production, substitute Spanish /i/ for English /i/ and Spanish /e/ for English /ɪ/.

The L1-English CDFA production model and the monolingual English listeners classified most natural Spanish /i/ productions as English /i/, the listeners also classified a few as English /ɪ/. Almost all instances of Spanish /ei/ were classified as English /e/. Instances of Spanish /e/ were classified as a mixture of English /ɪ/, /e/, and /ɛ/, but both the CDFA model and the listeners agreed on classifying more instances of Spanish /e/ as English /ɪ/ than as any other English vowel category. That monolingual English listeners assimilate a large proportion of Spanish /e/ productions to English /ɪ/, lends additional support to the hypothesis that monolingual Spanish listeners will assimilate most instances of English /ɪ/ to Spanish /e/.

Monolingual Spanish and English listeners also identified synthetic vowels in terms of their L1 categories (Section 4), and the vowel identification data were modelled using logistic regression (LR). The Spanish /i/–/e/ boundary predicted by the L1-Spanish LR perception model was relatively close to the English /i/–/ɪ/ boundary predicted by the L1-English LR perception model. Almost all the synthetic stimuli classified as English /i/ by the L1-English model, were classified as Spanish /i/ by the L1-Spanish model, and vice versa. Most of the synthetic stimuli classified as English /ɪ/ by the L1-English model, were classified as Spanish /e/ by the L1-Spanish model, but a couple were classified as Spanish /i/. L1-Spanish listeners just beginning to learn English were therefore predicted to reuse their L1-Spanish /i/–/e/ boundary as an L2-English /i/–/ɪ/ boundary, and give English /i/ labels to vowels they perceived as Spanish /i/, and English /ɪ/ labels to vowels they perceived as Spanish /e/.

L1 production and perception data were also used to make predictions beyond the initial state of L2 learning (Section 7). Predictions were based on Escudero's L2LP and a distribution-based interpretation of Flege's SLM. A single mega-model CDFA was fitted to both L1-Spanish and L1-English production data. Spanish /i/ and English /i/ were misclassified as each other at high rates. This indicated that Spanish /i/ and English /i/ were similar and were therefore expected to form a diaphone category. The distributional properties of this L2 diaphone category were predicted to become a mixture of the distributions of the acoustic properties of L1-Spanish /i/ and L1-English /i/. English /ɪ/ and /ɛ/ were correctly identified at a rate of almost 100%. This indicated that English /ɪ/ and /ɛ/ were far out on the tails of the distributions of any other vowel category. L1-Spanish learners of English were therefore predicted to infer the bimodal distribution of English /ɪ/ and /ɛ/ productions within their L1-Spanish /e/ perception area and form new L2 categories for English /ɪ/ and /ɛ/. The boundary between the L1-Spanish /i/ and /e/ categories was predicted to be reused as the boundary between the Sp+Eng/i/ diaphone and the new L2-English /ɪ/ category, and to shift towards the optimal location for distinguishing these two bilingual-set vowels.

12.3 L2 Production and Perception

Only three of the 40 L1-Spanish L2-English participants had perception and production results which were consistent with the predictions made on the basis of the L1 production and perception models (Section 9.3). Most of the remainder had perception and production patterns which were consistent with hypothesised developmental paths which posited that instances of English /ɪ/ were initially perceived as poor examples of Spanish /i/, rather than poor examples of Spanish /e/ (Sections 8 and 9).

Perception and production results for 25 of the 40 L1-Spanish L2-English participants were consistent with the *hypothesised indirect developmental path* for English /i/–/ɪ/ learning. In the hypothesised indirect developmental path, L1-Spanish listeners are at first unable to distinguish English /i/ and /ɪ/, then distinguish English /i/ and /ɪ/ via a category-goodness-difference assimilation to Spanish /i/, with more Spanish-/i/-like English vowels (low F1, zero VISC, and short duration) labelled as English /ɪ/, and less Spanish-/i/-like English vowels (higher F1, converging VISC, and longer duration) labelled as English /i/. Duration is the only perceptual cue whose use is positively correlated with L1-English speakers' English /i/ and /ɪ/ productions, so, with increased exposure to English, the L1-Spanish listeners shift towards using duration cues to distinguish the two English vowels. In L1-English speakers' English /i/ and /ɪ/ productions duration is only partially correlated with spectral cues, there is greater overlap between the two vowels in terms of the duration distributions than in terms of the spectral distributions, and vowel duration in L1-English is used to signal other contrasts such as post-vocalic obstruent voicing. Attempting to distinguish English /i/ and /ɪ/ on the basis of duration cues alone will therefore only lead to partial success; however, duration cues are partially correlated with spectral cues (all else being equal, English /i/ has lower F1, zero VISC, and longer duration compared to English /ɪ/ which has higher F1, converging VISC, and shorter duration), therefore L1-Spanish learners of English can use duration as a bootstrap for learning the appropriate spectral cues for English /i/ and /ɪ/. This path can eventually lead to an L1-English like /i/–/ɪ/ boundary.

In the cross-sectional component of the study, L1-Spanish L2-English participants had lived in Canada for different lengths of time. Length of residence (LOR) was generally found to be consistent with the assignments of participants to earlier and later stages of the hypothesised indirect learning path, these assignments having originally been made purely on the basis of perception patterns without consideration of LOR (Section 8.2.4). In longitudinal case studies (Section 10), some limited evidence was found in support of the changes in perception patterns hypothesised by the indirect developmental path being due to increased exposure to English; however, overall, the longitudinal results were not conclusive and additional longitudinal studies are needed.

Perception and production results for six of the L1-Spanish L2-English participants were consistent with the *hypothesised direct developmental path* for English /i/–/ɪ/ learning by L1-Spanish learners of English. In the hypothesised direct path, L1-Spanish listeners are at first unable to distinguish English /i/ and /ɪ/, then distinguish English /i/ and /ɪ/ via a category-goodness-difference assimilation to Spanish /i/ with more Spanish-/i/-like English vowels (low F1, zero VISC, and short duration) labelled as English /i/ and less Spanish-/i/-like English vowels (higher F1, converging VISC, and longer duration) labelled as English /ɪ/. Use of duration cues is negatively correlated, but use of spectral cues is positively correlated with L1-English speakers' English /i/ and /ɪ/ productions. Since spectral cues are the most important perceptual cues for L1-English listeners, this immediately leads to relatively L1-English-like perception of English /i/ and /ɪ/.

The a priori predictions made on the basis of the L1 production and perception models were that English /ɪ/ would initially be assimilated to Spanish /e/, and that the Spanish /e/ category would be split to form new L2-English /ɪ/ and /ɛ/ categories. Both the hypothesised direct and indirect paths for English /i/–/ɪ/ learning predicted that most instances of English /i/ and /ɪ/ would be assimilated to Spanish /i/, and that the Spanish /i/ category would be split into L2-English /i/ and /ɪ/ according to the degree of category goodness for Spanish /i/. The discrepancy between the predictions based on the L1-Spanish production and perception models and the L2 perception and production results which gave rise to the developmental hypotheses, may be due to non-perceptual factors such as orthography.[1] Reading and writing is typically a major component of English language

[1] A possible phonetic-perception explanation is that some important acoustic cues may not have been included in the perception and production models. For instance, perhaps different vowel categories differ in terms of intrinsic fundamental frequency (f_0, see Whalen & Levitt, 1995), and this affects listeners perception. It would seem unlikely, however, that the relatively small f_0 differences would have such a large effect in the perception of English /ɪ/ as to make it sound more like a Spanish /i/ than a Spanish /e/.

instruction at high school and university levels, whereas pronunciation is seldom given priority. As pointed out in Sections 1.3.2 and 8.2.3, English /ɪ/ corresponds to 'i' in English orthography, and in Spanish orthography 'i' corresponds to Spanish /i/; in addition, English /ɛ/ corresponds to 'e' in English orthography, and in Spanish orthography 'e' corresponds to Spanish /e/. Knowledge of Spanish and English orthography may therefore cause educated L1-Spanish learners of English to associate English /ɪ/ with Spanish /i/ and to associate English /ɛ/ with Spanish /e/ to the exclusion of English /ɪ/. Further, the focus on duration and the labelling of more-Spanish-/i/-like English vowel as English /ɪ/ and less-Spanish-/i/-like English vowel as English /i/ in the hypothesised indirect path, could have orthographic origins: Orthographic 'i' corresponds to Spanish /i/ and English /ɪ/, and in addition L1-Spanish learners of English may interpret the English orthographic double letter 'ee' (in real words such as *sheep* and the nonsense stimulus word *BEEPA*) as representing a long vowel (Escudero, 2000, §4.1.2; see also Flege, Bohn, & Jang, 1997; Morrison, 2005b). Also as suggested in Section 1.3.1 and 8.2.3, classroom instruction which inadvertently or misguidedly teaches students that English /i/ is long and English /ɪ/ short, may be the explanation for why most L1-Spanish L2-English learners focus on duration rather than spectral cues. It could be that L1-Spanish speakers who follow the hypothesised direct developmental path, rather than the hypothesised indirect developmental path, are those who are least influenced by formal ESL education.

The latter hypothesis could be tested: Monolingual Spanish speakers could be presented with non-speech analogues with a multidimensional bimodal distribution of spectral and duration properties similar to that of Canadian-English /i/ and /ɪ/ (so as to avoid the influence of any prior English instruction, the stimuli should not be recognisable as English /i/ and /ɪ/). The task would be to divide the stimuli into two groups without feedback or lexical cues as to group identity. Participants would be told that the sounds originally belonged to two groups and that their task was to discover the grouping. One group of participants would be given neutral group labels such as *orange* and *lemon*, and another group would be given misleading group labels such as *long* and *short* (other groups could be given spectrally informative labels such as *high* and *low*, or *rising* and *falling*). If the neutral group optimally categorised the stimuli on the basis of spectral properties, and the misled group suboptimally categorised the stimuli on the basis of duration, then this would support the hypothesis that L1-Spanish learners of English who choose the indirect path for learning English /i/ and /ɪ/ do so because of misleading instruction. Subsequent experiments could be conducted to assess the effectiveness of training on the English /i/–/ɪ/ contrast in which participants' attention is drawn to spectral cues and away from duration cues.

L2-Spanish production and perception data were generally found to be consistent with the a priori predictions made on the basis of L1 production and perception models. L2-Spanish learning was apparently easier for L1-English speakers than was L2-English learning for L1-Spanish speakers: Almost all L2-Spanish participants had perception results within the L1-Spanish range, and produced L2-Spanish vowels which received high correct-classification rates when classified by the CDFA model trained on L1-Spanish production data. A notable discrepancy was that approximately a third of L1-English L2-Spanish speakers produced L2-Spanish /ei/ with less VISC than their L1-English /e/, rather than equal or more VISC predicted if a Spanish /ei/ – English /e/ diaphone were formed. As predicted by Flege's SLM, both bilingual L1-English and bilingual L1-Spanish listeners arrived at a perceptual pattern intermediate between those of monolingual English and monolingual Spanish listeners.

The failure of the a priori predictions for L2-English made on the basis of Escudero's L2LP model and the distributional interpretation of Flege's SLM does not, at first sight, bode well for either model. However, it may be possible to adapt these models by adding a bias effect for the orthographic/educational factor (or other non phonetic factors). A major difference between the assumptions of the two models is the issue of whether L2 learners have a single phonological space for both languages or whether they have separate phonological grammars for each language. As mentioned in Section 1.2.2, Vallabha & McClelland's neural network model could be configured to work under either of these two assumptions. Neural network models could therefore be used to test which of the two assumptions results in a learning pattern which is most similar to the perceptual learning pattern observed/hypothesised in the present study for human L1-Spanish L2-English learners. Each version of the model (single space versus two grammars) would be initially trained on L1-Spanish production data. This would model L1-Spanish learning, culminating in models of a mature L1-Spanish perception system. The models would then be further trained on L1-English production data to model L2-English perception learning. It should be relatively easy to introduce a bias node into the neural net to model the hypothesised orthographic/educational bias which leads to the association of English /ɪ/ with Spanish /i/ rather than Spanish /e/. In postdoctoral research, I plan to use data from the present study (doctoral research) to develop train and test such models of L1 and L2 speech perception learning. In contrast to the static statistical models presented in the present study, which provided snapshots of participants' perception and production which were related to hypothesised stages of L1 and L2 learning, the proposed models will be dynamic quantitative models mimicking the actual L1 and L2 learning process.

Bibliography

Adank, P. van Hout, R., & Smits R. (2004). An acoustic description of the vowels of Northern and Southern Standard Dutch. *Journal of the Acoustical Society of America, 116,* 1729–1738.

Alvar, M. (Ed.). (1996a). *Manual de dialectología hispánica: El español de América* [Manual of Hispanic dialectology: The Spanish of America]. Barcelona, Spain: Ariel.

Alvar, M. (Ed.). (1996b). *Manual de dialectología hispánica: El español de España* [Manual of Hispanic dialectology: The Spanish of Spain]. Barcelona, Spain: Ariel.

Álvarez González, J. A. (1980). *Vocalismo español y vocalismo inglés* [Spanish and English vowels]. Unpublished Doctoral dissertation, Universidad Computense de Madrid.

Andruski, J. E., & Nearey, T. M. (1992). On the sufficiency of compound target specification of isolated vowels in /bVb/ syllables. *Journal of the Acoustical Society of America, 91*, 390–410.

Assmann, P. F., & Katz, W. F. (2000). Time-varying spectral change in the vowels of children and adults. *Journal of the Acoustical Society of America, 108*, 1856–1866.

Assmann, P. F., & Katz, W. F. (2005). Synthesis fidelity and time-varying spectral change in vowels. *Journal of the Acoustical Society of America, 117*, 886–895.

Assmann, P. F., Nearey, T. M., & Hogan, J. T. (1982). Vowel identification: Orthographic, perceptual, and acoustic aspects. *Journal of the Acoustical Society of America, 71*, 975–989.

Avis, W. S. (1973). The English language in Canada. In T. A. Sebeok (Ed.), *Current Trends in Linguistics* (Vol. 10, pp. 20–74). The Hague: Mouton.

Avis, W. S. (1975). The phonemic segments of an Edmonton idiolect. In J. K. Chambers (Ed.), *Canadian English: Origins and structures* (pp. 118-128). Toronto: Methuen.

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

Best, C. T. (1994). The emergence of native-language phonological influences in infants: A Perceptual Assimilation Model. In J. Goodman, & H. C. Nusbaum (Eds.), *The development of speech perception: The transition from speech sounds to spoken words* (pp. 167–224). Cambridge, MA: MIT Press.

Best, C. T. (1995a). 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). Timonium, MD: York Press.

Best, C. T. (1995b). Learning to perceive the sound pattern of English. In C. Rovee-Collier, & L. Lipsitt (Eds.), *Advances in infancy research* (pp. 217–304). Hillsdale, NJ: Ablex.

Bladon, A. (1985). Diphthongs: A case study of dynamic articulatory processing. *Speech Communication, 4*, 145–154.

Blankenship, B. R. (1991). *Vowel perception in a second language*. Unpublished master's thesis, University of California Los Angeles.

Boberg, C. (2005). The Canadian shift in Montreal. *Language Variation and Change, 17*, 133–154.

Boersma, P. (1998). Functional Phonology. Doctoral dissertation, University of Amsterdam, The Netherlands. The Hague, The Netherlands: Holland Academic Graphics.

Boersma, P. & Escudero, P. (2004). *Learning to perceive a smaller L2 vowel inventory: An Optimality Theory account* (Rutgers Optimality Archive 684).

Boersma, P. & Escudero, P. (2005). Measuring relative cue weighting: A reply to Morrison. *Studies in Second Language Acquisition, 27*, 607–617.

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 research* (pp. 279–304). Timonium, MD: York Press.

Bohn, O.-S., & Flege, J. E. (1993). Perceptual switching in Spanish/English bilinguals. *Journal of Phonetics, 21*, 267–290.

Bond, Z. S. (1978). The effects of varying glide duration on diphthong identification. *Language & Speech, 21*, 253–278.

Bond, Z. S. (1982). "Experiments with synthetic diphthongs," *Journal of Phonetics, 10*, 259–264.

Borzone de Manrique, A.M. (1979). Acoustic analysis of Spanish diphthongs. *Phonetica, 36*, 194–206.

Bradlow, A. R. (1993). *Language-specific and universal aspects of vowel production and perception: A cross-linguistic study of vowel inventories* (Doctoral dissertation, Cornell University). Ithaca, NY: DMLL.

Bradlow, A. R. (1995) . A comparative Study of English and Spanish vowels. *Journal of the Acoustical Society of America, 97*, 1916–1924.

Brennan, E. M., & Brennan, J. S. (1981). Measurement of accent and attitude toward Mexican American speech. *Journal of Psychological Research, 10*, 487–501.

Cebrian, J. (2002). *Phonetic similarity, syllabification and phonotactic constraints in the acquisition of a second language contrast* (Toronto Working Papers in Linguistics Dissertation Series). Toronto, ON: Department of Linguistics, University of Toronto.

Cebrian, J. (2003). Input and experience in the perception of an L2 temporal and spectral contrast. In M. J. Solé, D. Recansens, & J Romero (Eds.), *Proceedings of the 15th International Congress of Phonetic Sciences: Barcelona 2003* (pp. 2297–2300). Adelaide, South Australia: Causal Productions.

Cervera, T., Miralles, J. L., & González-Álvarez, J. (2001). Acoustic analysis of Spanish vowels produced by laryngectomized subjects. *Journal of Speech, Language, and Hearing Research, 44*, 988–996.

Chambers, J. K. (1991). Canada. In J. Cheshire (Ed.), *English around the world: Sociolinguistic perspectives* (pp. 89-107). Cambridge, UK: Cambridge University Press.

Clarke, S., Elms, F., & Youssef, A. (1995). The third dialect of English: some Canadian evidence. *Language Variation and Change, 7*, 209–228.

Contreras Oller, C. E. (1997) Características acústicas de las vocales altas anteriores del inglés: Estudio contrastivo de la producción de hablantes nativos del inglés y aprendices de inglés como lengua extranjera en el Departamento Idiomas - ULA [Acoustic characteristics of English high front vowels: Contrastive study of the production of native English speakers and English-as-a-foreign-language learners in the Language Department, University of the Andes]. *Estudios de Fonética Experimental, 13*, 127–152.

Dobrovolsky, M. (1996). Phonetics: The sounds of language. In W. O'Grady & M. Dobrovolsky (Eds.), *Contemporary linguistic analysis: An introduction* (3rd ed., pp. 15–57). Mississauga, ON: Copp Clark.

Escudero, P. (2000). *Developmental patterns in the adult L2 acquisition of new contrasts: The acoustic cue weighting in the perception of Scottish tense/lax vowels in Spanish speakers*. Unpublished master's thesis, University of Edinburgh, Edinburgh, Scotland.

Escudero, P. (2005). Linguistic perception and second language acquisition: Explaining the attainment of optimal phonological categorization. Doctoral dissertation, University of Utrecht, The Netherlands. Utrecht, The Netherlands: LOT.

Escudero P., & Boersma, P. (2002). The subset problem in L2 perceptual development: Multiple-category assimilation by Dutch learners of Spanish. In B. Skarabela, S. Fish, & A. H.-J. Do (Eds.), *Proceedings of the 26th Annual Boston University Conference on Language Development* (pp. 208–219). Somerville MA: Cascadilla.

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

Esling, J. H., & Warkentyne, H. J. (1993). Retracting of /æ/ in Vancouver English. In S. Clarke (Ed.), *Varieties of English Around the World, Vol. G11, Focus on Canada* (pp. 229–246). Philadephia, PA: John Benjamins.

Evans, B. G., & Iverson, P. (2004). Vowel normalization for accent: An investigation of best exemplar locations in northern and southern British English sentences. *Journal of the Acoustical Society of America, 115*, 352–361.

Fernández Planas, A. M. (1993). Estudio del campo de dispersión de las vocales castellanas [Study of the distribution of Castillian vowels]. *Estudios de Fonética Experimental, 5*, 129–162.

Flege, J. E. (1988). The production and perception of foreign language speech sounds. In H. Winitz (Ed.), *Human communication and its disorders: A review* (pp 224–401). Norwood, NJ: Ablex Publishing.

Flege, J. E. (1991). The interlingual identification of Spanish and English vowels: Orthographic evidence. *Quarterly Journal of Experimental Psychology, 43*, 701–731.

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). Timonium, MD: 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., & Hammond, R. M. (1982). Mimicry of non-distinctive phonetic differences between language varieties. *Studies in Second Language Acquisition, 5*, 1–17.

Flege, J. E. & Liu, S. (2001). The effect of experience on adults' acquisition of a second language. *Studies in Second Language Acquisition, 23*, 527–552.

Flege, J. E., & MacKay, I. R. A. (2004). Perceiving vowels in a second language. *Studies in Second Language Acquisition, 26*, 1–34.

Flege, J. E., Bohn, O.-S., & Jang, S. (1997). Effects of experience on non-native speakers' production and perception of English vowels. *Journal of Phonetics, 25*, 437-470.

Flege, J. E., Munro, M. J., & Fox, R. A. (1994). Auditory and categorical effects on cross-language vowel perception. *Journal of The Acoustical Society of America, 95*, 3623–3641.

Flege, J.E. (2003). Assessing constraints on second-language segmental production and perception. In N. Schiller & Antje Meyer (Eds.), *Phonetics and Phonology in Language Comprehension and Production: Differences and Similarities* (pp. 319–355). Berlin: Mouton de Gruyter.

Fox, R. (1983). Perceptual structure of monophthongs and diphthongs in English. *Language & Speech, 26*, 21–49.

Fox, R. (1989). Dynamic information in identification and discrimination of vowels. *Phonetica, 46*, 97–116

Fox, R. A., Flege, J. E., & Munro, M. J. (1995). The perception of English and Spanish vowels by native English and Spanish listeners: A multidimensional scaling analysis. *Journal of The Acoustical Society of America, 97*, 2540–2551.

Freida, E. M., Walley, A. C., Flege, J. E., & Sloane, M. E. (2000). Adult's perception and production of the English vowel /i/. *Journal of Speech, Language, and Hearing Research, 43*, 129–143.

García de las Bayonas, M. (2004) *The acquisition of vowels in Spanish and English as a second language*. Unpublished doctoral dissertation, Indiana University, Bloomington.

Gay, T. (1968). Effects of speaking rate on diphthong formant movements. *Journal of the Acoustical Society of America, 44*, 1570–1573.

Gay, T. (1970). A perceptual study of American English diphthongs. *Language & Speech, 13*, 65–88.

Gnanadesikan, R. (1977). *Methods for Statistical Data Analysis of Multivariate Observations*. New York: Wiley.

Godínez, M. Jr. (1978). A comparative study of some romance vowels. *UCLA Working Papers in Phonetics, 41*, 3–19.

Gottfried, M., Miller, J. D., & Meyer, D. J. (1993). Three approaches to the classification of American English diphthongs. *Journal of Phonetics, 21*, 205–229.

Grieser, D., & Kuhl, P. K. (1989). Categorization of speech by infants: Support for speech-sound prototypes. *Developmental Psychology, 25*, 577–588.

Guirao, M.., & Borzone de Manrique, A. M. (1975). Identification of Argentine Spanish vowels. *Journal of Psycholinguistic Research, 4*, 17–25.

Gumpertz, M. L., & Pantula, S. G. (1989). A simple approach to inference in random coefficient models. *American Statistician, 43*, 203–210.

Hagiwara, R. (2005). Revisiting the Canadian English vowel space. *Journal of the Acoustical Society of America, 117*, 2461.

Hagiwara, R. (in press). Vowel production in Winnipeg. *Canadian Journal of Linguistics*.

Hammond, R. M. (1986). Error analysis and the natural approach to teaching languages. *Lenguas Modernas, 13*, 129-139.

Hargus, S. F., and Kewley-Port, D. (2002). Vowel intelligibility in clear and conversational speech for normal-hearing and hearing-impaired listeners. *Journal of the Acoustical Society of America, 112*, 259.

Harrington, J., and Cassidy, S. (1994). Dynamic and target theories of vowel classification: Evidence from monophthongs and diphthongs in Australian English. *Language & Speech, 37*, 357–373.

Hastie, T., Tibshirani, R. & Friedman, J. (2001). *The elements of statistical learning: Data mining, inference, and prediction.* New York: Springer.

Hillenbrand, J. M., & Nearey, T. N. (1999). Identification of resynthesized /hVd/ syllables: Effects of formant contour. *Journal of the Acoustical Society of America, 105*, 3509–3523.

Hillenbrand, J. M., Clark, M. J., & Nearey, T. N. (2001). Effect of consonant environment on vowel formant patterns. *Journal of the Acoustical Society of America, 109*, 748–763.

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

Højen, A. (2005). Vowel discrimination in early bilinguals: How Plastic? In V. Hazan & P. Iverson (Eds.) *Proceedings of the ISCA workshop on Plasticity in Speech Perception* (pp.120–123). London, UK: UCL Centre for Human Communication.

Højen, A., & Flege, J. E. (in press). Early learners' discrimination of second-language vowels. *Journal of the Acoustical Society of America.*

Holbrook, A., and Fairbanks, G. (1962). Diphthong formants and their movements. *Journal of Speech and Hearing Research, 5*, 38–58.

Hosmer, D. W., & Lemeshow, S. (2000). *Applied logistic regression* (2nd Ed.). New York: John Wiley & Sons.

Hualde, J. I., & Prieto M. (2002). On the diphthong/hiatus contrast in Spanish: Some experimental results. *Linguistics, 40*, 217–234.

Huang, C. B. (1991). An acoustic and perceptual study of vowel formant trajectories in American English. *RLE Technical Report 563*, Research Laboratory of Electronics, MIT, Cambridge, MA.

Huang, C. B. (1992). Modelling human vowel identification using aspects of format trajectory and context. In Y. Tohkura, E. Vatikiotis-Bateson, & Y. Sagisaka (eds.), *Speech Perception, Production and Linguistic Structure* (pp. 43–61). Tokyo: Ohmsha / Amsterdam: IOS.

Imai, S., Flege, J. E., Wayland, R. (2002). Perception of cross-language vowel differences: A longitudinal study of native Spanish learners of English. *Journal of the Acoustical Society of America, 111*, 2364.

Ingram, J. C. L., & Park, S.-G. (1997). Cross-language vowel perception and production by Japanese and Korean learners of English. *Journal of Phonetics, 25*, 343–370.

Iverson, P., & Kuhl, P. K. (2000). Perceptual magnet and phoneme boundary effects in speech perception: Do they arise from a common mechanism? *Perception and Psychophysics, 62*, 874–886.

Jenkins, J. J., Strange, W., and Miranda, S. (1994). Vowel identification in mixed-speaker silent-center syllables. *Journal of the Acoustical Society of America, 95*, 1030–1043.

Johnson, D. E. (1998). *Applied multivariate methods for data analysis*. Pacific Grove, CA: Duxbury.

Johnson, K., Flemming, E., & Wright, R. (1993). The hyperspace effect: Phonetic targets are hyperarticulated. *Language, 69*, 505–528.

Kawahara, H., Masuda-Katsuse, I., & de Cheveigné, A. (1998). Restructuring speech respresentations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds. *Speech Communication, 27*, 187–207.

Kewley-Port, D. (2001). Vowel formant discrimination II: Effects of stimulus uncertainty, consonantal context, and training. *Journal of the Acoustical Society of America, 110*, 2141–2155.

Kewley-Port, D., Akahane-Yamada, R., & Aikawa, K. (1996). Intelligibility and acoustic correlates of Japanese accented English vowels. *Proceedings of ICSLP 96,* Philadelphia, PA. 450–453.

Klatt, D. H. (1980). Software for a cascade/parallel formant synthesizer. *Journal of the Acoustical Society of America, 82*, 737–793.

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

Klatt, D. H., & Whalen, D. H. (1985). KLSYN Version 1.45 [software]. [Available: http://homepages.wmich.edu/~hillenbr/]

Klecka, W. R. (1980). *Discriminant analysis*. Beverly Hills, CA: Sage.

Kuhl, P. K. (2000). A new view of language acquisition. *Proceedings of the National Academy of Science, 97*, 11850–11857.

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

Kuhl, P. K., & Iverson, P. (1995). Linguistic experience and the "perceptual magnet effect". In W. Strange (Ed.), *Speech perception and linguistic experience: Issues in cross-language research* (pp 121–154). Timonium, MD: York Press.

Leather, J. (1983). Second-language pronunciation learning and teaching. *Language Teaching, 16*, 198–219.

Leather, J. (1999). Second-language speech research: an introduction. In J. Leather (Ed.), *Phonological issues in language learning* (pp. 1-58). Oxford: Blackwell.

Lehiste, I., & Peterson, G. E. (1961). Transitions, glides, and diphthongs. *Journal of the Acoustical Society of America, 33*, 268–277.

León Valdés, H. (1998). Determinación del campo de dispersión auditiva de las vocales de la serie anterior del español de Chile [Determining the auditory distribution of the front vowel series in Chilean Spanish]. *Revista de Lingüística Teórica y Aplicada, 36*, 113–126

Llisterri, J. (1995). Relationships between speech production and speech perception in a second language. In K. Elenius & P. Branderud (Eds.), *Proceedings of the 13th International Congress of Phonetic Sciences: Stockholm 1995* (Vol. 4, pp. 92–99). Stockholm, Sweden: KTH.

Maddox, W. T, Molis, M. R., & Diehl, R. L. (2002) Generalizing a neuropsychological model of visual categorization to auditory categorization of vowels. *Perception & Psychophysics, 64*, 584–597.

Madrid Servín, E. A., & Marín Rodríguez, M. A. (2001). Estructura formántica de las vocales del español de la ciudad de México [Formant structure of the vowels of Mexico-City Spanish]. In E. Z. Herrera (Ed.), Temas de fonética instrumental (pp. 39–58). México, DF: El Colegio de México.

Markel, J. D., & Gray, A. H. (1976). *Linear prediction of speech.* Berlin: Springer-Verlag.

Martínez Celdrán, E. (1995). En torno a las vocales del español: Análisis y reconocimiento [Concerning the vowels of Spanish: Analysis and recognition]. *Estudios de Fonética Experimental, 7*, 195–218.

Martínez Celdrán, E., Fernádez Planas, A. M., & Carrera Sabaté, J. (2003). Castilian Spanish. *Journal of the International Phonetic Association, 33*, 255–259.

Martínez Melgar, A. (1990). El vocalismo del andaluz oriental [The Eastern Andalucian vowel system]. *Estudios de Fonética Experimental, 6*, 12–64.

MathWorks. (2001 & 2004). Matlab Versions 6 & 7 [software]. Natick, MA: The MathWorks.

Maye, J., & Weiss, D. (2003). Statistical cues facilitate infants' discrimination of difficult phonetic contrasts. In B. Beachley et al. (Eds), *Proceedings of the 27th Annual Boston University Conference on Language Development* (pp. 508–518). Somerville, MA: Cascadilla.

Maye, J., Werker, J. F., & Gerken, L. A. (2002). Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition, 82*, B101–B111.

McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models.* London: Chapman and Hall.

Menard, S. (2002). *Applied logistic regression analysis.* Thousand Oaks, CA: Sage.

Miller, J. D. (1989). Auditory-perceptual interpretation of the vowel. *Journal of the Acoustical Society of America, 85*, 2114–2134.

Møller Glasbrenner, M. (2005). *Vowel identification by monolingual and bilingual listeners: Use of spectral change and duration cues.* Unpublished master's thesis, University of South Florida.

Morrison, G. S. (2002a). *Effects of L1 duration experience on Japanese and Spanish listeners' perception of English high front vowels*. Unpublished master's thesis, Simon Fraser University, Burnaby, British Columbia, Canada.

Morrison, G. S. (2002b). Japanese listeners' use of duration cues in the identification of English high front vowels. In J. Larson & M. Paster (Eds.), *Proceedings of the 28th annual meeting of the Berkeley Linguistics Society* (pp. 189–200). Berkeley, CA: Berkeley Linguistics Society.

Morrison, G. S. (2002c). Spanish listeners' use of vowel spectral properties as cues to post-vocalic consonant voicing in English. In *Collected Papers of the First Pan-American/Iberian Meeting on Acoustics* [CD-ROM]. Mexico, DF: Mexican Institute of Acoustics.

Morrison, G. S. (2003). Perception and production of Spanish vowels by English speakers. In M. J. Solé, D. Recansens, & J Romero (Eds.), *Proceedings of the 15th International Congress of Phonetic Sciences: Barcelona 2003* (pp. 1533–1536). Adelaide, South Australia: Causal Productions.

Morrison, G. S. (2004). An acoustic and statistical analysis of Spanish mid-vowel allophones. *Estudios de Fonética Experimental, 13*, 11–37.

Morrison, G. S. (2005a). An appropriate metric for cue weighting in L2 speech perception: Response to Escudero & Boersma (2004). *Studies in Second Language Acquisition, 27*, 597–606.

Morrison, G. S. (2005b). *Development of L2 vowel perception and production: L1-Spanish speakers and the acquisition of the English /i/–/ɪ/ contrast*. Manuscript submitted for publication.

Morrison, G. S. (2006). Methodological issues in L2 perception research, and vowel spectral cues in Spanish listeners' perception of word-final /t/ and /d/ in Spanish. In M. Díaz-Campos (Ed.), *Selected Proceedings of the 2nd Conference on Laboratory Approaches to Spanish Phonetics and Phonology* (pp. 35–47). Somerville, MA: Cascadilla Proceedings Project. [Available online: http://www.lingref.com document #1324]

Morrison, G. S., & Nearey, T. M. (2005, October). *Testing theories of vowel inherent spectral change*. Poster presented at the 150th Meeting of the Acoustical Society of America, Minneapolis, Minnesota, USA. [Abstract published: *Journal of the Acoustical Society of America, 118*, 1932]

Morrison, G. S., & Nearey, T. M. (In press). A cross-language vowel normalisation procedure. *Canadian Acoustics, 34(3).*

Nábělek, A. K., Czyzewski, Z., & Crowley, H. (1993). Vowel boundaries for steady-state and linear formant trajectories. *Journal of the Acoustical Society of America, 94*, 675–687.

Nábělek, A. K., Czyzewski, Z., & Crowley, H. (1994). Cues for perception of the diphthong /aɪ/ in either noise or reverberation. Part I. Duration of the transition. *Journal of the Acoustical Society of America, 95*, 2681–2693.

Nábělek, A. K., Czyzewski, Z., & Krishan, L. A. (1992). The influence of talker differences on vowel identification by normal-hearing and hearing-impaired listeners. *Journal of the Acoustical Society of America, 92*, 1228–1246.

Navarro Tomás, T. (1918). *Manual de pronunciación española* [Spanish pronunciation manual]. Madrid, Spain: CSIC [1965, 12th ed.].

Nearey, T. M. (1978). Phonetic features systems for vowels (Doctoral dissertation, University of Connecticut). Bloomington, IN: Indiana University Linguistics Club.

Nearey, T. M. (1989). Static, dynamic, and relational properties in vowel perception. *Journal of the Acoustical Society of America, 85*, 2088–2113.

Nearey, T. M. (1990). The segment as a unit of speech perception. *Journal of Phonetics, 18*, 347–373.

Nearey, T. M. (1995). Evidence for the perceptual relevance of vowel-inherent spectral change for front vowels in Canadian English. In K. Elenius & P. Branderud (Eds.), *Proceedings of the 13th International Congress of Phonetic Sciences: Stockholm 1995* (pp. 678–681). Stockholm, Sweden: KTH.

Nearey, T. M. (1997). Speech perception as pattern recognition. *Journal of the Acoustical Society of America, 101*, 3241–3254.

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

Nearey, T. M., & Assmann, P. F. (In press). Probabalistic 'sliding template' models for vowel normalization. In M. J. Solé, P. S. Beddor, & M. Ohala (Eds.), *Experimental approaches to phonology*. Oxford: Oxford University Press.

Nearey, T. M., Assmann, P. F., & Hillenbrand, J. M. (2002). *Evaluation of a strategy for automatic formant tracking*. Poster presented at the First Pan-American/Iberian Meeting on Acoustics, Cancún, Quintana Roo, Mexico.

Nearey, T. M., & Hogan, J. T. (1986). Phonological contrast in experimental phonetics: Relating distributions of production data to perceptual curves. In J. J. Ohala & J. Jaeger (Eds.), *Experimental phonology*, (pp. 141–161). New York: Academic Press.

Neel, A. T. (2004). Formant detail needed for vowel identification. *Acoustic Research Letters Online, 5*, 125–131.

Noteboom, S. G, & Doodeman, G. J. N. (1980). Production and perception of vowel length in spoken sentences. *Journal of the Acoustical Society of America, 67*, 276–287.

Pampel, F. C. (2000). *Logistic regression: A primer*. Thousand Oaks, CA: Sage.

Parker, D. M., & Diehl, R. L. (1984). Identifying vowels in CVC syllables: Effects of inserting silence and noise. *Perception & Psychophysics, 36*, 369–380.

Pols, L. C. W. (1977). Spectral analysis and identification of Dutch vowels in monosyllabic words. PhD dissertation, University of Amsterdam.

Quilis, A., & Esgueva, M. (1983). Realización de los fonemas vocálicos españoles en posición fonética normal [Realization of Spanish vowel phonemes in normal phonetic position]. In *Estudios de Fonética I, Colectanea Phonetica VII* (pp. 159–252). Madrid: Consejo Superior de Investigaciones Científicas.

Skelton, R. (1969). The pattern of Spanish vowel sounds. *International Review of Applied Linguistics, 7*, 231–237.

Smits, R., Sereno, J., & Jongman, A. (2005). *Categorization of sounds*. Manuscript submitted for publication.

SPSS Inc. (2002). SPSS Version 11.5 [software]. Chicago, IL: SPSS Inc.

Strange, W., Jenkins, J. J., and Johnson, T. L. (1983). Dynamic specification of coarticulated vowels. *Journal of the Acoustical Society of America, 74*, 695–705.

Studebaker, G. R. (1985). A "rationalized" arcsine transform. *Journal of Speech and Hearing Research, 28*, 455–462.

Summers, W. V. (1987). Effects of stress and final-consonant voicing on vowel production: Articulatory and acoustic analyses. *Journal of the Acoustical Society of America, 82*, 847–863.

Tatsuoka, M. M. (1970). *Discriminant analysis: The study of group differences*. Champaign, IL: Institute for Personality and Ability Testing.

Thomson, R. (2005). *A pattern recognition approach to English L2 vowel learning.* Unpublished manuscript, University of Alberta.

Traunmüller, H. (1990). Analytical expressions for the tonotopic sensory scale. *Journal of the Acoustical Society of America, 88*, 97–100.

Vallabha, G. K., & McClelland, J.L. (2005). *Learning new speech categories in adulthood: Costs and benefits of Hebbian attractors in topographic maps*. Manuscript submitted for publication.

Vaquero de Ramírez, M. (1996). *El español de América I: Pronunciación* [The Spanish of America I: Pronunciation]. Madrid, Spain: Arco.

Vaquero de Ramírez, M., & Guerra de la Fuente, L. (1992). Fonemas vocálicos de Puerto Rico [Puerto Rican vowel phonemes]. *Revista de Filología Española, 72*, 555–582.

Wang, X., & Munro, M. J. (1999). The perception of English tense-lax vowel pairs by native Mandarin speakers: The effect of training on attention to temporal and spectral cues. In J. J. Ohala et al. (Eds.), *Proceedings of the 14th International Congress of Phonetic Sciences: San Francisco 1999* (pp. 125–128). Berkeley, CA: University of California Berkeley.

Wang, X., & Munro, M. J. (2004). Computer-based training for learning English vowel contrasts. *System, 32*, 539–552.

Watson, C., and Harrington, J. (1999). Acoustic evidence of dynamic formant trajectories in Australian English vowels. *Journal of the Acoustical Society of America, 106*, 458–468.

Wells, J. C. (1982). *Accents of English 3: Beyond the British Isles*. Cambridge, UK: Cambridge University Press.

Werker, J. F., & Curtin, S. (2005). PRIMIR: A developmental model of speech processing. *Language Learning and Development, 1*, 197–234.

Whalen, D. H., & Levitt, A. G. (1995). The universality of intrinsic F0 of vowels. *Journal of Phonetics, 23*, 349–366.

Wise, C. M. (1964). Acoustic structure of English diphthongs and semivowels *vis-a-vis* their phonetic symbolization. *Proceedings of the5th International Congress on Phonetic Sciences* (Munster, 1964), pp. 589–593.

Zahorian, S., & Jagharghi, A. (1993). Spectral-shape features versus formants as acoustic correlates for vowels. *Journal of the Acoustical Society of America, 94*, 1966–1982.

Zahorian, S. A., & Jagharghi, A. J. (1991). Speaker normalisation of static and dynamic vowel spectral features, *Journal of the Acoustical Society of America, 90*, 67–75.

Appendix 1.
Vowel Inherent Spectral Change

The classical description of a diphthong includes an initial steady state, a glide, and a final steady state (Lehiste & Peterson, 1961); however, there is usually no second steady state (see Holbrook & Fairbanks, 1962), the first steady state may disappear at fast speaking rates (Gay, 1968), and diphthongs can be synthesised using only a glide (Gay, 1970). The English vowel system traditionally comprises true diphthongs, e.g., /aɪ, aʊ, ɔɪ/, so called phonetic diphthongs, /e, o/ [eɪ, oʊ], and nominal monophthongs, e.g., /i, ɪ, ɛ, æ/. However, several studies have observed that most nominal monophthongs are in fact diphthongised in a number of North American dialects (for example, Alberta: Assmann, Nearey, & Hogan, 1982; Nearey & Assmann, 1986; Andruski & Nearey, 1992. Ohio: Fox, 1983. General American: Nábělek, Czyzewski, & Krishnan, 1992. Michigan: Hillenbrand et al., 1995; Hillenbrand & Nearey, 1999; Hillenbrand, Clark, & Nearey, 2001. Texas: Assmann & Katz, 2000. Indiana: Hargus Ferguson & Kewley-Port, 2002. Manitoba: Hagiwara, 2005). Figure 1.1 illustrates the extent of *vowel-inherent spectral change* (VISC) from the beginning to the end of productions of phonetic diphthongs and nominal monophthongs measured by Nearey & Assmann (1986). VISC has been found to play an important role in speech perception: Listeners' vowel identifications change when they are presented with stimuli that have typical formant trajectories versus flat formant trajectories versus reversed formant trajectories (Nearey & Assmann, 1986; Nearey, 1995; Hillenbrand & Nearey, 1999; Assmann & Katz, 2000, 2005). For example, when formant trajectories are reversed, /e/ stimuli are identified as /ɪ/, and /ɪ/ stimuli as /e/ (Nearey & Assmann, 1986). Listeners also give higher goodness ratings to synthetic versions of nominal monophthongs that include VISC (Nearey, 1995). And when pattern recognition models are provided with information about formant trajectories in nominal monophthongs and phonetic diphthongs, as compared to formant measurements from a single point, higher correct classification rates are obtained, and there is higher correlation with listeners' perception patterns (see below).

Three hypotheses have been advanced as to the perceptually relevant aspects of VISC (see Nearey & Assmann, 1986; and Gottfried, Miller, & Meyer, 1993).[1] All three hypotheses agree that the initial formant frequencies are perceptually relevant to vowel identification (for supporting evidence see Gay, 1970; Bladon, 1985; Nábělek Czyzewski, & Crowley, 1993; and Nearey, 1995), but disagree on what additional cues are relevant in VISC.

- The *dual-target* hypothesis states that the relevant perceptual cues are the formant values at the end of the vowel.
- The *target plus slope* hypothesis states that the relevant perceptual cue is the velocity of formant change, i.e., whether the change is positive or negative and the rate of change in time.
- The *target plus direction* hypothesis states that the only relevant factor is the direction of formant movement in an F1–F2 (or similar) space.

Contra the dual-target hypothesis, Gay (1968) found substantial speaking-rate dependent differences in final formant values and more consistency in slope (see also Borzone de Marique, 1979, for slope consistency in Spanish, and Pols, 1977, for direction consistency in Dutch); however, it could be argued that listeners are able to compensate for target undershoot, and that substantial variability in target may be unproblematic if there are only a few widely separated targets, and thus little chance of confusion between them (Bladon, 1985). Contra the target plus slope hypothesis and pro the dual-target hypothesis, Bond (1978, 1982) found that changing the duration of the glide between initial and final targets had little effect on vowel identification, and in some cases even deleting the glide completely had no effect (for glide deletion see also Wise, 1965; Bladon, 1985; Nearey & Assmann,

[1] The terminology used here is that of Nearey & Assmann (1986). Gottfried, Miller, & Meyer's (1993) "onset + offset", "onset + slope", and "onset + direction" represent the same hypotheses, with the exception that they only included F2 slope in their onset + slope hypothesis; F1 and F2 slopes were used in studies conducted by Nearey and colleagues. In contrast to Lehiste & Peterson's (1961) use of the term *target*, Nearey & Assmann's (1986) term *dual target* does not imply that there must be steady states at the beginning and end of the diphthongs.

1986; Andruski & Nearey, 1992). Contra the target plus direction hypothesis, Bladon (1985) found that phonetically trained listeners transcribed truncated diphthongs with pairs of symbols appropriate for monophthongs at the initial and final formant values of the stimuli; the second symbol varied with the final formant values and was not invariant with direction. However, Bladon's (1985) choice of stimuli make the relevancy of the results questionable: He removed the latter portions of /ia, iɛ, ie/, all three have similar initial formant values and a similar direction, but different final targets. However, it is not clear that /ia, iɛ, ie/ really are phonemes, i.e., that they are perceived holistically as single units rather than as a sequence of two phonemes. Although there are clearly some similarities, findings based on a sequence of two vowels (or a glide plus vowel) may have little relevance for the perception of true diphthongs, and even less relevance for phonetic diphthongs and nominal monophthongs. A complication for the target plus direction hypothesis is the issue of whether some minimum magnitude of formant change is needed: instances of a vowel with negligible VISC may have random fluctuations in the direction of formant movement that are not perceptually pertinent (Nearey & Assmann, 1986). Note, however, that the same minimum formant movement threshold requirement could equally well apply to the dual-target and slope hypotheses. Perception of a vowel as a diphthong, as opposed to a monophthong, may also require some minimum duration for the glide portion of the vowel (see Nábělek, Czyzewski, & Crowley, 1994).

Some studies have found evidence in support of the slope hypothesis. Gay (1970) claimed that slope was the primary cue for distinguishing between different diphthongs, e.g., /ɔɪ/–/aɪ/; however, his synthetic stimuli confounded either target and slope or duration and slope, and his set of experiments did not allow full separation of the effects of slope from its covariants.[2] Assmann, Nearey, & Hogan (1982) applied pattern recognition models to

[2] The interpretation of Gay's (1970) results is hindered by contradictions between the description of his stimuli and the discussion of the results. Discussion and graphical results suggest that, in his Experiment II, F2 offset did not covary with duration so as to maintain a fixed slope, rather F2 offset stepped up at a slower rate than duration, e.g., for /ɔ/–/ɔɪ/ stimuli with an F2 onset of 840 Hz, the first three duration steps of 100, 110, and 120 ms all had an F2 offset of 1320 Hz, and thus progressively shallower slopes of 4.80, 4.36, and 4.00 Hz/ms; the next two duration steps of 130 and 140 ms both had an F2 offset of 1440 Hz, and thus slopes of 4.62 and 4.29; etc..

measurements of formant values of Canadian English nominal monophthongs and phonetic diphthongs, and obtained higher correct classification and higher correlation with listeners' response patterns when they included formant slope parameters in addition to midpoint formant parameters.

Other studies (Hillenbrand et al., 1995; Andruski & Nearey, 1992; Hillenbrand & Nearey, 1999; Hillenbrand, Clark, & Nearey, 2001) obtained higher correct classification or higher correlation with listeners' response patterns when they used dual-target parameterisations of Canadian and US English nominal monophthongs and phonetic diphthongs (see also Adank, van Hout, & Smits, 2004, for Dutch vowels). Andruski & Nearey (1992) conducted experiments using silent centre natural /bVb/ stimuli (short portions extracted from the beginning and end of natural productions), silent centre natural isolated vowel stimuli, and synthetic /bVb/ stimuli in which the vowel portion was a linear interpolation from initial to final target values. Since similar perceptual results were obtained for all three stimulus types, they argued that the perceptually relevant cues were the cues shared by all three, i.e., the initial and final target values (this is also a possible interpretation of the results of Strange, Jenkins, & Johnson, 1983).[3] Using a different methodology with US English true diphthongs, phonetic diphthongs, and nominal monophthongs, Fox (1983) also obtained results consistent with the dual-target hypothesis. In a multidimensional scaling experiment, Fox (1983) extracted four perceptual dimensions: the first dimension was most highly correlated with F2 formant values measured at the end of the vowels, and the third dimension with F2 formant values measured at the beginning of the vowels. Huang (1991, 1992) and Harrington & Cassidy (1994) found that pattern classifiers based on triple point models, e.g., measurements taken at 25%, 50%, and 75% of vowel duration, outperformed single point models, e.g., measurements taken at 50% of vowel duration, for non-back US English nominal monophthongs and /e/, and for Australian English diphthongs

[3] Fox (1989) presented evidence that when presented with very short extracts from consonant transitions, listeners extrapolate the trajectories of vowels from the dynamic information available in consonant transitions, rather than using absolute values immediately before and after silent centres. This is not necessarily inconsistent with the dual-target hypothesis if one assumes that the trajectories are extrapolated to include the targets. The initial and final portions in Andruski & Nearey's (1992) silent centre stimuli were relatively long and may therefore have actually reached the target values.

and nominal monophthongs. The authors did not claim that a triple target model was correct, only that more than a single target model was necessary.[4] Hillenbrand et al. (1995) compared one point, two point, and three point parameterisations of US English nominal monophthongs. Substantially higher correct classification rates were obtained for dual-target models as compared to single target models, but triple target models offered little or no improvement over dual-target models.

Nearey & Assmann (1986) tested the three VISC hypotheses using pattern recognisers fed with different parameterisations of Canadian English nominal monophthongs and phonetic diphthongs. Parameters were initial F1 and F2 values plus: final F1 and F2 values for the dual-target hypothesis; change in F1 and F2 values over the duration of glide for the target plus slope hypothesis; and change in F1 and F2 values each over the magnitude of the total change, e.g., $\Delta F1/\sqrt{\Delta F1^2+\Delta F2^2}$, for the target plus direction hypothesis (all formant values were transformed to natural logarithms prior to making any other calculations). Correlations with listeners' responses were slightly higher for the dual-target and target plus direction parameterisations than for the target plus slope parameterisations, but in general all three parameterisations provided adequate characterisations of listeners' response patterns. Gottfried, Miller, & Meyer (1993) compared the three hypotheses using pattern recognisers fed with different parameterisations of US English phonetic and true diphthongs. They used two sets of parameterisations: one was similar to that of Nearey & Assmann (1986) in that it used log F1 and log F2 measurements, but differed in that only the

[4] In a small-scale study Neel (2004) investigated the perception of synthetic 1, 2, 3, 5, and 11 point versions of US English phonetic diphthongs and nominal monophthongs. Each stimulus was based on the formant tracks from a single /dVd/ production from one of two speakers (problems with the study may be related to idiosyncrasies in the small number of productions). Two-point stimuli (based on formant measurements at 10% and 90% of duration) were poorly identified, typically at rates substantially worse than one-point stimuli (based on formant measurements at 50% of duration). A possible reason for this is that the 10% and 90% points may actually have been in the consonant transitions and therefore not representative of the initial and final target values: Identification rates were generally high for five-point stimuli (based on formant measurements at 10%, 30%, 50%, 70% and 90% of duration) which would be expected since these stimuli included some approximation of onset to initial target transition, initial target to final target transition (via a midpoint value), and final target to offset transition.

F2 slope was included,[5] and that the direction was specified as an angle in degrees, with adjustments made to avoid discontinuities at 0° / 360°. The second set of parameters transformed F1, F2, and F3 values into Miller's *auditory-perceptual space* (APS: Miller, 1989). Across speaking conditions (slow stressed, slow unstressed, fast stressed, and fast unstressed) the log formant parameterisations had slightly higher correct classification rates for the dual-target and slope hypotheses than the direction hypothesis, and in the APS parameterisations the dual-target hypothesis had higher correct classification rates than the slope and direction hypotheses. However, no one hypothesis was superior to the others in all contexts.

Different studies have selected different points at which to measure the initial and final target, e.g., at the earliest and latest measurable values (Nearey & Assmann, 1986), 40 ms after the initial consonant release and 40 ms before the final consonant closure (Andruski & Nearey, 1992), at 20% and 80% of the duration of the vowel (Hillenbrand & Nearey, 1999), and at 20% and 70% of the duration of the vowel (Hillenbrand, Clark, & Nearey, 2001). Gottfried, Miller, & Meyer (1993) measured at points immediately following and preceding the consonant transitions, which they determined on the basis of an algorithm which made use of the speed of formant movement. The choice of measurement points will clearly have an influence on the dual-target parameterisation. It may also affect the slope parameterisation: if there is any steady state between the measurement points then the true slope will be underestimated (most studies using data based on acoustic measurements of productions have not attempted to divide vowels into steady state and glide portions). If correct, the direction parameterisation is least likely to be affected.

The parameterisations above, based on formant measurements at two or three points in the vowel, could be criticised as being relatively crude measures incapable of capturing all the relevant details of inherently complex time-varying patterns (see Jenkins, Strange, &

[5] Assmann & Katz (2000) tested the perception of stimuli in which the F1 trajectory was flattened and F2 unchanged, and stimuli in which the F2 trajectory was flattened and F1 unchanged. Listeners' correct identification rates for the set of US English nominal monophthongs and phonetic diphthongs significantly decreased when either formant was flattened. Although some vowels were affected more by F1 flattening, and some were affected more by F2 flattening. The results indicate that a VISC theory applicable across vowel categories should refer to formant movement in both F1 and F2.

Miranda, 1994). Several studies have used more sophisticated curve-fitting parameterisations. Zahorian & Jagharghi (1991, 1993) fitted *discrete cosine transforms* (DCT) to the time-varying spectral properties of US English nominal monophthongs (although /e/ was excluded, /o/ was included). Static spectral slices were parameterised as formant values and as cepstral coefficients. For both spectral-slice parameterisations, the highest correct classification rates and highest correlations with listeners' responses were obtained for models that included the first two DCT coefficients, e.g., the five best predictors in the formant parameterisation were, in the following order, the first DCT coefficient for F1, F2, F3, F0, and the second DCT coefficient for F2 (the next best predictors were also second DCT coefficients). The first DCT coefficient gives the mean value of a formant/cepstral coefficient over time, and the second coefficient is a measure of time-normalised slope of the formant/cepstral coefficient trajectory: a half period of a cosine is fitted to the values of the formant/cepstral coefficient measured from the beginning to the end of the vowel. The value of the second DCT coefficient is therefore a symmetrically constrained measure of the direction and distance of the initial and final target from the mean value. This parameterisation is therefore similar to the dual-target parameterisation, but based on a curve fitted to the whole trajectory rather than only two points. Models including dynamic information outperformed models that did not, but no comparisons with dual-target, target plus slope, or target plus direction models were made. Watson & Harrington (1999) fitted DCTs to formant trajectories from Australian English vowels. Higher correct classification rates were obtained for models using the first and second DCT coefficients than for models using only the first, the differences were significant for vowels traditionally labelled as true and phonetic diphthongs but not for nominal monophthongs (see Harrington & Cassidy, 1994, for similar results based on a triple target model). No comparisons with slope, dual-target, or target plus direction parameterisation were made; however, some support for the superiority of the DCT parameterisation came from the observation that Australian English lax-tense vowel pairs had differences in the second DCT coefficient although they had very similar initial and final target values.[6] Hillenbrand, Clark, & Nearey

[6] Theoretically, there are some problems with this result: If lax and tense vowel pairs have the same initial and final targets, but the shape of the trajectory between those targets differs, then the second DCT should not provide a good fit for both of these shapes because the second DCT models only one shape, that of a half period

(2001) experimented with fitting polynomials and DCTs to formant trajectories from US English nominal monophthong and phonetic diphthongs, and concluded that these parameterisations were not superior to the simpler dual-target parameterisation. Thus, there has been no proof that more sophisticated curve-fitting parameterisations are superior to the dual-target parameterisation with respect to the substantive issues of correct classification and correlation with listeners' responses.

Since the dual-target hypothesis has proven to be successful in terms of correct classification of production data and correlation with perception data, and no worse than more sophisticated parameterisations based on curve fitting techniques, it is adopted in the present study.

of a cosine. For example, if the trajectory in tense vowels were a perfect half period of a cosine then the second DCT would fit this trajectory with zero error, and if the trajectory in the lax vowels were linear then there would be a large error in the fit of the second DCT (although in this case the second DCT coefficient value would be the same for the tense and lax vowel). Some pairs of shapes will result in different second DCT coefficients, but at least one will be a poor fit for the real shape of the trajectory. Therefore, a difference in the second DCT coefficient may indicate that a tense and lax vowel pair with the same initial and final target have different shaped trajectories, but it also indicates that a two-coefficient DCT parameterisation is not an ideal parameterisation of the shape of the trajectories. Adding third and higher order coefficients would allow the DCT model to fit different shaped trajectories.

Appendix 2.
Adaptive Sampling Procedure

A2.1 Introduction

A typical speech perception experiment involves creating a set of synthetic speech stimuli whose acoustic properties form a multidimensional matrix, randomly presenting each stimulus a fixed number of times, and, at each presentation, having listeners classify each stimulus as one of a number of speech sound categories. Data consist of the proportion of responses for each category given to each stimulus. A simple experiment might involve a two-dimensional matrix and two speech sound categories, e.g., equally spaced vowel duration steps on one dimension and equally spaced first formant (F1) steps on another dimension, covering the range of F1 and duration values between English /i/ and /ɪ/. More complex experiments may involve a larger number of response options and a larger number of stimulus dimensions. Several acoustic dimensions may be necessary to adequately model listeners' perception, but as the number of dimensions increases, the number of stimuli increases exponentially.

From the perspective of building an accurate unbiassed statistical model of listeners' speech categorisation, it is desirable to obtain a large number of responses for each stimulus from each participant. With a larger number of samples, there will be greater resolution in the proportional responses for each category. Unfortunately, collecting a large number of responses from human participants is time consuming, the participants can quickly become fatigued, and may be reticent to return to participate in subsequent sessions in longitudinal or multiple-condition experiments. The present paper describes an adaptive sampling procedure which was developed in order to make more efficient use of participants' time whilst still obtaining a reasonable degree of resolution in the proportional responses. For a quite different approach to adaptive sampling focussing on best exemplars rather than boundaries see Evans & Iverson (2004).

A2.2 Stimulus Set

The adaptive sampling procedure was initially developed for use with an experiment investigating the perception of English /i/, /ɪ/, /e/, /ɛ/, and Spanish /i/, /ei/, /e/. The procedure will be described using the stimulus set from this study as a concrete example. There were a total of 90 synthetic vowel stimuli covering three acoustic dimensions. The duration dimension had three points [80, 95, 110 ms]; the F1–F2 dimension had ten points, the first and second formants (F2) at the beginning of the vowel covaried forming a diagonal in the F1–F2 space [F1: 283–580 Hz in 33 Hz steps, F2: 2090–1730 in 40 Hz steps]; and the vowel inherent spectral change (VISC) dimension had three points, from the beginning of the vowel to the end F1 and F2 either diverged, remained flat, or converged [ΔF1: −-99, 0, +99 Hz, ΔF2: +120, 0, −120 Hz]. The number of stimuli had been winnowed from a larger stimulus space, by combining the F1 and F2 dimensions and reducing the number of points on each dimension; however, the stimuli were embedded in words in carrier sentences and in pilot tests it took listeners approximately half an hour to identify each stimulus four times (360 trials). The goal was to develop a sampling procedure which would give a resolution comparable to six responses per stimulus within the half hour time frame.

A2.3 Adaptive Sampling Procedure

A2.3.1 Basic procedure

The essential principle underlying the procedure is that certain stimuli will not need to be sampled a large number of times because they fall near the middle of a listener's perceptual space for a given category, and will therefore always be identified as that category. For example, if a stimulus is in the middle of the perceptual space for a listener's /i/ category, then the listener will always identify this stimulus as /i/; thus irrespective of the number of responses the listener gives to this stimulus, the proportion of /i/ responses for this stimulus will always be 1. Hence, once portions of the perceptual space which are far from boundaries have been located, there is no need to obtain further responses in those areas. On the other hand, stimuli near category boundaries may be identified as one category on one occasion, and as another category on another occasion. For example, a stimulus may be identified as /i/ two thirds of the time and as /ɪ/ one third of the time, and a neighbouring stimulus may be identified as /i/ half the time and as /ɪ/ half the time. In order to determine the proportion of /i/ responses with reasonable resolution such stimuli must be sampled a considerable number of times.

The procedure consists of the following steps:

1. All the stimuli are sampled twice, i.e., all the stimuli are presented in two blocks (once in each block) and the listener gives a identification response on each trial (180 responses).
2. A logistic regression model is fitted to the response data, and the predicted probabilities for each category are calculated for each stimulus.
3. The error between the predicted probability and observed proportion for each category for each stimulus is calculated.
4. Half of the stimuli, primarily those with the largest error scores, are resampled (45 responses, see Section 3.2).
5. Steps 2 through 4 are repeated three more times.

This procedure results in 360 trials, and each stimulus is sampled a minimum of twice and a maximum of six times. After two rounds, a stimulus which receives two /i/ responses and is surrounded by stimuli which receive two /i/ responses is unlikely to be near a category boundary. This stimulus will have an observed proportion of /i/ responses of 1, and a predicted probability for /i/ close to 1. This stimulus will therefore have a low error score, and is unlikely to be resampled in subsequent rounds. In contrast, a stimulus which receives two /i/ responses but is adjacent to stimuli which receive /ɪ/ responses, will have an observed proportion of /i/ responses of 1, but will have a predicted probability for /i/ that is somewhat less than 1. This stimulus will therefore have a higher error score, and is more likely to be resampled in subsequent rounds. A stimulus which receives one /i/ response and one /ɪ/ response could have a small error between observed and predicted values, but, especially in a multidimensional stimulus space and with multinomial response categories, it is more likely to have a relatively large error. In practice, the vast majority of stimuli near category boundaries receive relatively high error scores, and stimuli far from category boundaries receive low error scores.

An alternative procedure which resampled the stimuli with predicted probabilities furthest from 0 and 1 was also explored. Selecting stimuli using this criterion gave similar results to using the highest-error-score criterion, but the latter offered the advantage of a stronger mistake amelioration feature: A mistake where a listener accidentally presses the wrong button is likely to increase the error score for the stimulus on which the mistake was made. Using the highest-error-score criterion, that stimulus will therefore be resampled, leading to a reduction in the effect of the mistake.

The model fitted was a simple first-order model ($V + V\times\text{F1} + V\times\Delta\text{F1} + V\times\text{dur}$) containing one bias and three stimulus-tuning coefficients for each vowel category. Stimulus-tuning coefficients consisted of F1-tuning with initial formant values for F1 entered in Hertz (since F2 covaried with F1 it was redundant), ΔF1-tuning, with change in F1 value from the beginning to the end of the vowel entered in Hertz, and duration-tuning, with vowel duration values entered in milliseconds. All stimulus properties were treated as continuous variables. The number of each type of coefficient in the fitted model was actually one less than the number of categories, the coefficients for the last category being redundant and calculable as minus the sum of the coefficients for the other categories. A simple model is preferred to avoid overfitting the sparse data sets, especially near the beginning of the adaptive sampling procedure. An overfitted model may wrap around fluctuations in the data sets due to course sampling and give lower error scores to stimuli near boundaries than would the optimal model. In simulations, use of a quadratic model resulted in unstable results with high variances for the coefficients in the final model. Using an underfitted model during adaptive sampling will be less efficient than the optimal model, but will not obliterate more complex patterns in the data which may be captured by fitting a more complex model to the final results. If the model makes a linear approximation of a curved boundary then some stimuli will be a poor fit to the model because the model is underfitted; however, this will lead to these stimuli being resampled and the curved boundary will still be represented in the final data set.

A2.3.2 Selecting stimuli to resample

Rather than simply resampling the 45 stimuli with the highest error scores, the stimuli to resample were chosen such that those with higher error scores were most likely to be resampled but those with lower error scores also had some probability of being resampled. This ensured that listeners heard some reasonably good examples of the vowel categories in each round. Good examples provide the listeners with anchors against which to compare more ambiguous stimuli, good examples will also be easy to identify and thus be reassuring for the listeners. The stimuli to resample were selected stochastically in the following manner:

1. The stimuli were ranked in ascending order of their error scores, resulting in a sequence which increased in an approximately exponential manner (see Figure A2.1).
2. The error score of the 67th stimulus of the 90 ranked stimuli was obtained. (Vertical line in Figure A2.1)
3. Integers from 1 to 90 were randomly permuted then divided by 90 and multiplied by the error score of the 67th ranked stimulus. This generated a sequence of random numbers with the highest number being equal to the error score of the 67th ranked stimulus.
4. The sequence of ranked error scores and the sequence of random numbers were added. (Noisy line in Figure A2.1)
5. Stimuli with error-plus-random scores of greater than the median value were selected for resampling. (The median value is represented by the horizontal line in Figure A2.1. The stimuli selected for resampling are indicated by the bars at the bottom of the figure.)

Half the stimuli are resampled. All the stimuli have a non-zero probability of being resampled which increases with their error score, and the quarter of the stimuli with the worst fit are guaranteed to be resampled.

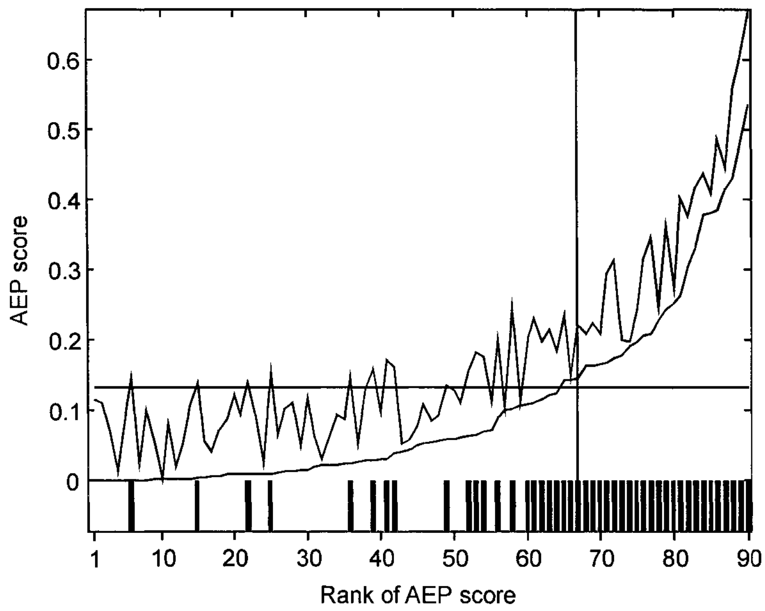


Figure A2.1 Example of selection of stimuli to be resampled on the basis of absolute errors in proportions (AEP) for a model fitted to two responses per stimulus.

A2.3.3 Error measures

Standard error measures such as *Root Mean Squared* (RMS) error are usually calculated assuming that each stimulus is sampled an equal number of times, which is not the case for the adaptive sampling procedure. Ad hoc error measures used instead were the *Absolute Errors in Proportions* (AEP) for individual stimuli, and the *Sum of the Absolute Errors in Proportions* (SAEP) for the stimulus set.

The AEP for a stimulus is calculated as half the sum of the absolute difference between the observed proportion of responses and the predicted proportion of responses for each category for that stimuli, or equivalently as half the sum of the absolute difference between the observed and predicted number of responses for each category divided by the total number of responses for that stimulus:

$$AEP_{stim} = \frac{\sum_{cat} \left| obs_{cat} - pred_{cat} \right|}{2 \times NumResponses_{stim}}$$

The theoretical minimum and maximum values for AEP are 0 and 1 (the scaling factor of ½ was introduced to make the maximum value 1). An AEP value of 0 indicates a perfect fit between the observed responses and the model's fitted responses, and an AEP value of 1 indicates a complete mismatch (e.g., if the participant always responded with one category, and the model predicted a probability of zero for that category). The SAEP for the stimulus set is calculated as the sum of the AEP for all stimuli. This measure is discussed in greater depth in Appendix 3.

An alternative error measure could have been to calculate errors of fit on the basis of differences between observed and predicted logit values. The error measure based on proportions was preferred since errors which would be the same size in logistic values, are smaller in proportion values when they are close to proportions of 0 and 1 relative to when they are near proportions of .5, and this weighting was advantageous because the error measures were being used as a criterion to select stimuli that were near category boundaries.

A2.4 Simulations

To obtain test data, the full set of stimuli were presented in random order in six blocks (540 trials), and on each trial the stimulus presented was identified by a single listener as one of the four English vowel responses. A first-order logistic regression model was fitted to the whole data set (a territorial map based on this model is given in Figure 2). The a posteriori probabilities from this model were used as population parameters in a multinomial sample generator which generated 100 simulated response sets of six responses per stimulus. Simulated responses were generated independently for each stimulus. To generate a single simulated response for a stimulus, the sample generator chose one of the four English vowels /i/, /ɪ/, /e/, /ɛ/, the probability of choosing a particular response category on each occasion being dependent on its a posteriori probability for that stimulus.

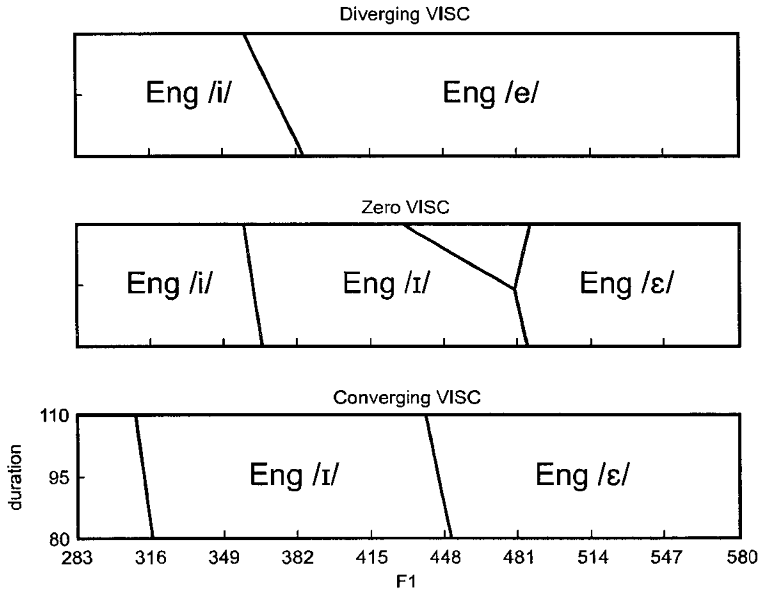


Figure A2.2 Territorial map based on logistic regression model fitted to original test data.

Whole-set logistic regression models were fitted to each of the 100 simulated response sets and the SAEP and coefficient values saved. Models based on the final set of responses selected by the adaptive sampling procedure were fitted to the same 100 simulated response sets. The first two simulated responses to each stimulus were used in both models, but subsequent simulated responses for a stimulus were only used in the adaptive model if that stimulus was selected for resampling. The whole-set models were compared with the adaptive models: for each sample set the difference between the logistic regression coefficient values for the whole-set model and the adaptive model were calculated, and these were used as the test statistic in paired-sample *t*-tests.

Different variants of the adaptive procedure were tested using different criteria for selecting the stimuli to resample and different levels of complexity for the logistic regression model. The version of the adaptive procedure described above was selected as giving the closest results to the whole-set model. Numerical comparisons between the whole-set model and this version of the adaptive model are presented below.

When the adaptive sampling procedure was applied to the original (as opposed to the simulated) test data, the resulting logistic regression coefficients were identical to the whole-set model. Table A2.1 presents the results of comparisons between the whole-set and the adaptive model for the simulated response sets. The difference in SAEP between the models was not significant. The differences between models for four coefficients were significant; however, the size of the difference was small, none of the mean differences were greater than 4.5%. Three of the four significant differences were related to a single response category, /i/, and were therefore not independent of each other: The magnitudes of the bias and the stimulus-tuned coefficients for /i/ all decreased by similar amounts (3.3–4.4%) indicating a slight reduction in the estimate of the rate at which responses changed from /i/ to other categories, but little change in the location of the boundary (if the size or direction of the change in the bias had differed from the size of the change in the stimulus-tuned coefficients, then the modelled location of the boundary would have changed).

Table A2.1 Comparison of error scores and coefficient values across sampling procedures.

Error	Sampling Method								
or	Six Samples		Adaptive		Difference				
Coefficient	Mean (sd)		Mean (sd)		Mean (sd)		%	*t*(99)	*p*
SAEP	6.813	(0.677)	6.793	(0.676)	-0.020	(0.426)	-0.3	-0.4711	.6386
i	34.113	3.187	32.923	3.201	-1.190	1.349	-3.5	-8.8228	.0000 **
ɪ	7.141	1.877	7.074	1.977	-0.068	0.906	-0.9	-0.7480	.4562
e	-8.147	2.123	-8.181	2.280	-0.034	0.792	+0.4	-0.4263	.6708
i×*F1*	-0.077	0.007	-0.074	0.007	0.003	0.003	-3.6	10.0470	.0000 **
ɪ×*F1*	-0.007	0.004	-0.007	0.004	0.000	0.002	-2.0	0.8498	.3975
e×*F1*	0.012	0.004	0.012	0.004	0.000	0.001	-1.8	-1.7419	.0846
i×*dip*	-2.028	0.353	-1.939	0.374	0.089	0.164	-4.4	5.4614	.0000 **
ɪ×*dip*	1.510	0.269	1.486	0.266	-0.025	0.137	-1.6	-1.8019	.0746
e×*dip*	-3.445	0.326	-3.384	0.359	0.061	0.189	-1.8	3.2167	.0018 **
i×*dur*	-0.037	0.016	-0.036	0.016	0.001	0.006	-3.3	2.0202	.0461 *
ɪ×*dur*	-0.020	0.011	-0.021	0.012	-0.001	0.004	+3.3	-1.6124	.1101
e×*dur*	0.044	0.013	0.045	0.013	0.001	0.005	+1.2	0.9991	.3202

% Percentage differences indicate differences in magnitude which are towards zero if negative and away from zero if positive

* significant at $\alpha = .05$, ** significant at $\alpha = .0038$ equal to .05 after a Bonferroni correction for 13 tests

The adaptive procedure required 360 trails. A non-adaptive model of 360 trails, four responses per stimulus, had a significantly larger SAEP than the original six-response model [mean 16.919 vs 6.813, $t(99) = 67.355, p < .0038$]. As in the case of the adaptive procedure, four coefficients (*i*, *ɪ*, *i×F1*, *i×dip*) had coefficients of significantly [$p < .0038$] smaller magnitude compared to the original model; however, the magnitude of the differences were greater than was the case for the adaptive model: these four coefficients were 4.3–5.9% smaller than those of the original model, and one non-significant difference (*ɪ×F1*) was 7.5% smaller.

In order to test the sampling method on a wider set of simulated data that might reflect a wider range of listeners, the data set was perturbed in several ways. The coefficient values from the logistic regression model based on the original data collected from the listener were reduced to 25% of their original values, and used to generate a further series of 100 sample sets. SAEP was significantly higher for the adaptive compared to the whole-set models [mean 20.425 vs 18.857, $t(99) = 18.823, p < .0038$], but none of the coefficient values had significant differences. Another series of 100 sample sets was generated on the basis of the original model, but 25% of the responses were replaced by responses generated at random with each response category having an equal probability irrespective of stimulus properties. SAEP was significantly higher for the adaptive compared to the whole-set models [mean 25.426 vs 24.114, $t(99) = 5.438, p < .0038$]. The mean difference in *i*, and *i*×F1 coefficient values between the adaptive and the whole-set models were also significantly different [*i* mean 7.612 vs 7.129, $t(99) = 5.208, p < .0038$; *i*×F1 mean −0.016 vs −0.017, $t(99) = 5.208, p < .0038$], the magnitude of both these differences was 6.8%.

A2.5 Conclusion

On the basis of the simulations, it was decided that any small differences in the accuracy of results were immaterial compared to the benefits accrued by presenting the participants with a shorter experiment, 360 trials rather than 540. The adaptive sampling procedure as described above was therefore adopted for use in data collection in the study of the perception of English /i/, /ɪ/, /e/, /ɛ/, and Spanish /i/, /ei/, /e/. Individual participants took between 20 and 40 minutes to complete the perception experiment, and participant retention was very high: of the 95 participants who were asked to participate in two or more experiment sessions (e.g., one experiment giving English responses and one experiment giving Spanish responses to the same stimuli), only 3 dropped out after the first session.

Appendix 3.
AEP & SAEP Error Measures

A standard intuitive error measure for goodness-of-fit is *Root Mean Squared* (RMS) error. A conservative variant of RMS error (that used in Nearey, 1990, 1997; and Morrison, 2005b) sums the difference between observed and predicted values over all the stimuli, and uses the degrees of freedom in the denominator which includes an adjustment for the number of parameters in the fitted model:

$$RMS = \sqrt{\frac{\sum_{stim} \left(obs_{stim} - pred_{stim}\right)^2}{df}}$$

$$df = NumStimuli \times (NumCategories - 1) - NumParameters$$

Percentage RMS takes into account the number of responses, assuming an equal number of responses for each stimulus:

$$\%RMS = 100 \times (RMS \,/\, NumResponsesPerStimulus)$$

Another initiative goodness-of-fit measure is *Mean of Absolute Errors* (MAE):

$$MAE = \frac{\sum_{stim} \left|obs_{stim} - pred_{stim}\right|}{NumStimuli}$$

MAE can also be scaled by number of responses:

$$\%MAE = 100 \times (MAE \,/\, NumResponsesPerStimulus)$$

Note that this version of MAE is less conservative than the version of RMS above because the number of stimuli, rather than the degrees of freedom, are used in the denominator. However, the difference between using number of stimuli and degrees of freedom is simply a scaling factor.

The measures above are appropriate when all stimuli are sampled an equal number of times, but not when different stimuli are sampled a different number of times. In the latter case, one appropriate measure would be *Sum of Absolute Errors in Proportions* (SAEP), in which the absolute difference between the observed and predicted value for each stimulus is divided by the number of responses for that stimulus:

$$SAEP = \tfrac{1}{2} \times \sum_{stim} \left(\frac{\left| obs_{stim} - pred_{stim} \right|}{NumResponses_{stim}} \right)$$

The reason for the scaling factor of ½ is explained below. The SAEP equation above can be applied directly to a binomial model, a multinomial version takes the difference between the observed and predicted value for each response category for each stimulus. The *Absolute Errors in Proportions* (AEP) for a single stimulus is:

$$AEP_{stim} = \frac{\sum_{cat} \left| obs_{cat} - pred_{cat} \right|}{2 \times NumResponses_{stim}}$$

The SAEP over all the stimuli in the experiment is then the sum of the AEP:

$$SAEP = \sum_{stim} AEP_{stim}$$

Several mathematically equivalent methods of calculating AEP and SAEP are possible, and in the present study they were actually calculated using the difference between the observed proportions and predicted proportions for each response category for each stimulus (hence the name).

An alternative error measure could have been to calculate errors of fit on the basis of differences between observed and predicted logit values. The error measurement based on proportions was preferred since errors which would be the same size in logistic values, are smaller in proportion values when they are close to proportions of 0 and 1 relative to when they are near proportions of .5, and this weighting was advantageous since the error measures were being used as a criterion to select stimuli that were near category boundaries.

AEP and SAEP are ad hoc measures developed to deal with a problem in error measurement introduced by the efficient sampling method which did not obtain an equal number of responses from each stimulus. The theoretical minimum and maximum values for AEP are 0 and 1 (the scaling factor of ½ was introduced to make the maximum value 1). An AEP value of 0 indicates a perfect fit between the observed responses and the model's fitted responses, and an AEP value of 1 indicates a complete mismatch (e.g., if the participant always responded with one category, and the model predicted a probability of zero for that category). An intuitive way to conceptualise a SAEP value of, say, 5 is to imagine that the model was a perfect fit for most of the stimuli but that it was a perfect mismatch for 5 stimuli. In practice SAEP will be distributed over the set of stimuli, some will have larger and some smaller AEP, but values of exactly 0 or 1 are unlikely. SAEP can also be scaled as a percentage of the number of stimuli.

Appendix 4.
Statistical Comparisons of L1 Vowel Productions

Table A4.1a ANOVA comparing monolingual versus bilingual L1-English groups on all English vowels. Dependent variable: F1.

Source	*df*	*F*	*p*
Group	1, 45.001	.283	.597
Gender	1, 45.001	87.751	.000
Vowel	3, 138.013	1046.053	.000
Speaker (Group × Gender)	45, 138.020	5.256	.000
Group × Gender	1, 45.001	.040	.842
Group × Vowel	3, 138.043	6.197	.001
Gender × Vowel	3, 138.032	3.156	.027
Vowel × Speaker (Group × Gender)	138, 1787	15.631	.000

Table A4.1b Follow-up ANOVAs comparing monolingual versus bilingual L1-English groups on individual English vowels. Only results for Group main effects are reported in the table, Gender and Speaker (Group × Gender) effects were significant ($\alpha = .05$) and Group × Gender effects were not significant for all vowels. Dependent variable: F1.

Vowel	Marginal Means (Hz)		*df*	*F*	*p*
	Monolingual	Bilingual			
Eng /i/	335	363	1, 45.011	3.112	.085
Eng /ɪ/	517	530	1, 44.998	.006	.939
Eng /e/	491	510	1, 45.016	1.055	.310
Eng /ɛ/	683	673	1, 45.005	2.252	.140

Table A4.2a ANOVA comparing monolingual versus bilingual L1-English groups on all English vowels. Dependent variable: F2.

Source	*df*	*F*	*p*
Group	1, 45.000	2.784	.102
Gender	1, 45.000	93.319	.000
Vowel	3, 138.017	1369.941	.000
Speaker (Group × Gender)	45, 138.027	27.201	.000
Group × Gender	1, 45.000	.153	.698
Group × Vowel	3, 138.057	4.366	.006
Gender × Vowel	3, 138.043	5.890	.001
Vowel × Speaker (Group × Gender)	138, 1787	11.59	.000

Table A4.2b Follow-up ANOVAs comparing monolingual versus bilingual L1-English groups on individual English vowels. Only results for Group main effects are reported in the table, Gender and Speaker (Group × Gender) effects were significant (α = .05) and Group × Gender effects were not significant for all vowels. Dependent variable: F2.

Vowel	Marginal Means (Hz)		*df*	*F*	*p*
	Monolingual	Bilingual			
Eng /i/	2520	2473	1, 45.003	4.854	.033
Eng /ɪ/	1941	1924	1, 44.999	1.643	.207
Eng /e/	2251	2193	1, 45.004	5.046	.030
Eng /ɛ/	1779	1788	1, 45.004	.322	.573

Table A4.3 ANOVA comparing monolingual versus bilingual L1-English groups on all English vowels. Dependent variable: ΔF1.

Source	*df*	*F*	*p*
Group	1, 45.005	.145	.706
Gender	1, 45.005	.295	.590
Vowel	3, 138.024	192.784	.000
Speaker (Group × Gender)	45, 138.038	2.296	.000
Group × Gender	1, 45.005	.002	.969
Group × Vowel	3, 138.081	1.279	.284
Gender × Vowel	3, 138.061	5.215	.002
Vowel × Speaker (Group × Gender)	138, 1787	8.225	.000

Table A4.4 ANOVA comparing monolingual versus bilingual L1-English groups on all English vowels. Dependent variable: ΔF2.

Source	*df*	*F*	*p*
Group	1, 45.004	.649	.425
Gender	1, 45.004	2.783	.102
Vowel	3, 138.039	467.082	.000
Speaker (Group × Gender)	45, 138.061	4.143	.000
Group × Gender	1, 45.004	.320	.575
Group × Vowel	3, 138.130	1.622	.187
Gender × Vowel	3, 138.097	2.141	.098
Vowel × Speaker (Group × Gender)	138, 1787	5.125	.000

Table A4.5a ANOVA comparing monolingual versus bilingual L1-English groups on all English vowels. Dependent variable: duration.

Source	*df*	*F*	*p*
Group	1, 45.001	.189	.666
Gender	1, 45.001	10.068	.003
Vowel	3, 138.044	573.031	.000
Speaker (Group × Gender)	45, 138.070	23.967	.000
Group × Gender	1, 45.001	.036	.851
Group × Vowel	3, 138.149	5.755	.001
Gender × Vowel	3, 138.112	.170	.916
Vowel × Speaker (Group × Gender)	138, 1787	4.459	.000

Table A4.5b Follow-up ANOVAs comparing monolingual versus bilingual L1-English groups on individual English vowels. Only results for Group main effects are reported in the table, Gender and Speaker (Group × Gender) effects were significant (α = .05) and Group × Gender effects were not significant for all vowels. Dependent variable: duration.

Vowel	Marginal Means (ms)		*df*	*F*	*p*
	Monolingual	Bilingual			
Eng /i/	91	94	1, 45.012	.057	.813
Eng /ɪ/	68	74	1, 44.998	2.033	.161
Eng /e/	122	120	1, 45.009	.737	.395
Eng /ɛ/	89	93	1, 45.009	.449	.506

Table A4.6a ANOVA comparing monolingual versus bilingual L1-Spanish groups on all Spanish vowels. Dependent variable: F1.

Source	*df*	*F*	*p*
Group	1, 55.001	.319	.574
Gender	1, 55.001	39.170	.000
Vowel	2, 112.027	1161.250	.000
Speaker (Group × Gender)	55, 112.004	7.886	.000
Group × Gender	1, 55.001	.022	.882
Group × Vowel	2, 112.040	4.375	.015
Gender × Vowel	2, 111.996	6.830	.002
Vowel × Speaker (Group × Gender)	112, 1590	9.291	.000

Table A4.6b Follow-up ANOVAs comparing monolingual versus bilingual L1-Spanish groups on individual Spanish vowels. Only results for Group main effects are reported in the table, Gender and Speaker (Group × Gender) effects were significant (α = .05) and Group × Gender effects were not significant for all vowels. Dependent variable: F1.

Vowel	Marginal Means (Hz)		*df*	*F*	*p*
	Monolingual	Bilingual			
Sp /i/	353	350	1, 55.000	.090	.765
Sp /ei/	510	509	1, 55.000	.001	.974
Sp /e/	499	520	1, 55.012	5.095	.028

Table A4.7 ANOVA comparing monolingual versus bilingual L1-Spanish groups on all Spanish vowels. Dependent variable: F2.

Source	*df*	*F*	*p*
Group	1, 55.000	1.147	.289
Gender	1, 55.000	111.031	.000
Vowel	2, 112.034	657.498	.000
Speaker (Group × Gender)	55, 112.005	25.422	.000
Group × Gender	1, 55.000	.210	.649
Group × Vowel	2, 112.050	2.956	.056
Gender × Vowel	2, 111.995	.097	.908
Vowel × Speaker (Group × Gender)	112, 1590	7.461	.000

Table A4.8a ANOVA comparing monolingual versus bilingual L1-Spanish groups on all Spanish vowels. Dependent variable: ΔF1.

Source	*df*	*F*	*p*
Group	1, 55.007	2.135	.150
Gender	1, 55.006	.099	.754
Vowel	2, 112.030	726.868	.000
Speaker (Group × Gender)	55, 112.005	1.550	.026
Group × Gender	1, 55.006	.353	.555
Group × Vowel	2, 112.044	5.535	.005
Gender × Vowel	2, 111.996	.632	.533
Vowel × Speaker (Group × Gender)	112, 1590	8.557	.000

Table A4.8b Follow-up ANOVAs comparing monolingual versus bilingual L1-Spanish groups on individual Spanish vowels. Only results for Group main effects are reported in the table, Gender and Speaker (Group × Gender) effects were significant (α = .05) and Group × Gender effects were not significant for all vowels. Dependent variable: ΔF1.

Vowel	Marginal Means (ΔHz)		*df*	*F*	*p*
	Monolingual	Bilingual			
Sp /i/	-7	-10	1, 55.000	1.216	.275
Sp /ei/	-129	-113	1, 55.001	5.121	.028
Sp /e/	+2	+4	1, 55.041	.144	.706

Table A4.9a ANOVA comparing monolingual versus bilingual L1-Spanish groups on all Spanish vowels. Dependent variable: ΔF2.

Source	*df*	*F*	*p*
Group	1, 55.004	.660	.420
Gender	1, 55.003	1.091	.301
Vowel	2, 112.029	533.669	.000
Speaker (Group × Gender)	55, 112.004	2.522	.000
Group × Gender	1, 55.004	.223	.638
Group × Vowel	2, 112.042	4.637	.012
Gender × Vowel	2, 111.996	.532	.589
Vowel × Speaker (Group × Gender)	112, 1590	8.848	.000

Table A4.9b Follow-up ANOVAs comparing monolingual versus bilingual L1-Spanish groups on individual Spanish vowels. Only results for Group main effects are reported in the table, Gender and Speaker (Group × Gender) effects were significant (α = .05) and Group × Gender effects were not significant for all vowels. Dependent variable: ΔF2.

Vowel	Marginal Means (ΔHz)		*df*	*F*	*p*
	Monolingual	Bilingual			
Sp /i/	+30	+45	1, 55.000	2.249	.139
Sp /ei/	+303	+269	1, 55.001	2.935	.092
Sp /e/	+31	+30	1, 55.043	.108	.743

Table A4.10 ANOVA comparing monolingual versus bilingual L1-Spanish groups on all Spanish vowels. Dependent variable: duration.

Source	Marginal Means (ms)		*df*	*F*	*p*
	Monolingual	Bilingual			
Group	93	102	1, 55.002	5.467	.023
Gender			1, 55.001	15.354	.000
Vowel			2, 112.027	479.099	.000
Speaker (Group × Gender)			55, 112.004	6.086	.000
Group × Gender			1, 55.002	.012	.914
Group × Vowel			2, 112.040	.464	.630
Gender × Vowel			2, 111.996	.540	.584
Vowel × Speaker (Group × Gender)			112, 1590	9.338	.000

Table A4.11 ANOVA comparing Peninsular versus Mexican L1-Spanish groups on all Spanish vowels. Dependent variable: F1.

Source	*df*	*F*	*p*
Group	1, 37.000	.370	.546
Gender	1, 36.999	27.356	.000
Vowel	2, 75.995	908.416	.000
Speaker (Group × Gender)	37, 76.004	7.975	.000
Group × Gender	1, 36.999	.000	.988
Group × Vowel	2, 76.009	1.975	.146
Gender × Vowel	2, 75.994	3.284	.043
Vowel × Speaker (Group × Gender)	76, 1103	9.497	.000

Table A4.12a ANOVA comparing Peninsular versus Mexican L1-Spanish groups on all Spanish vowels. Dependent variable: F2.

Source	*df*	*F*	*p*
Group	1, 37.000	1.226	.275
Gender	1, 37.000	95.007	.000
Vowel	2, 75.993	597.226	.000
Speaker (Group × Gender)	37, 76.006	27.898	.000
Group × Gender	1, 37.000	.932	.988
Group × Vowel	2, 76.013	4.991	.009
Gender × Vowel	2, 75.992	.220	.803
Vowel × Speaker (Group × Gender)	76, 1103	6.416	.000

Table A4.12b Follow-up ANOVAs comparing Peninsular versus Mexican L1-Spanish groups on individual Spanish vowels. Only results for Group main effects are reported in the table, Gender and Speaker (Group × Gender) effects were significant (α = .05) and Group × Gender effects were not significant for all vowels. Dependent variable: F2.

Vowel	Marginal Means (Hz)		*df*	*F*	*p*
	Peninsular	Mexican			
Sp /i/	2298	2282	1, 37.000	1.324	.257
Sp /ei/	2043	2057	1, 37.000	.138	.712
Sp /e/	1951	1903	1, 36.999	2.819	.102

Table A4.13a ANOVA comparing Peninsular versus Mexican L1-Spanish groups on all Spanish vowels. Dependent variable: ΔF1.

Source	*df*	*F*	*p*
Group	1, 36.998	4.480	.041
Gender	1, 36.997	.214	.647
Vowel	2, 75.995	562.400	.000
Speaker (Group × Gender)	37, 76.004	1.515	.064
Group × Gender	1, 36.997	1.110	.299
Group × Vowel	2, 76.010	3.628	.031
Gender × Vowel	2, 75.994	.477	.623
Vowel × Speaker (Group × Gender)	76, 1103	9.018	.000

Table A4.13b Follow-up ANOVAs comparing Peninsular versus Mexican L1-Spanish groups on individual Spanish vowels. Only results for Group main effects are reported in the table, Gender and Speaker (Group × Gender) effects were significant ($\alpha = .05$) and Group × Gender effects were not significant for all vowels. Dependent variable: ΔF1.

Vowel	Marginal Means (Hz)		*df*	*F*	*p*
	Peninsular	Mexican			
Sp /i/	-7	-9	1, 37.000	.289	.594
Sp /ei/	-129	-114	1, 37.002	4.499	.041
Sp /e/	-1	+9	1, 36.997	4.469	.041

Table A4.14 ANOVA comparing Peninsular versus Mexican L1-Spanish groups on all Spanish vowels. Dependent variable: ΔF2.

Source	*df*	*F*	*p*
Group	1, 36.998	.054	.818
Gender	1, 36.998	.294	.591
Vowel	2, 75.994	433.260	.000
Speaker (Group × Gender)	37, 76.005	2.331	.001
Group × Gender	1, 36.998	.615	.438
Group × Vowel	2, 76.011	1.846	.165
Gender × Vowel	2, 75.994	.613	.544
Vowel × Speaker (Group × Gender)	76, 1103	8.213	.000

Table A4.15 ANOVA comparing Peninsular versus Mexican L1-Spanish groups on all Spanish vowels. Dependent variable: duration.

Source	Marginal Means (ms)		*df*	*F*	*p*
	Peninsular	Mexican			
Group	92	101	1, 36.999	4.114	.050
Gender			1, 36.999	7.387	.010
Vowel			2, 75.995	414.111	.000
Speaker (Group × Gender)			37, 76.004	6.225	.000
Group × Gender			1, 36.999	.181	.673
Group × Vowel			2, 76.010	1.471	.236
Gender × Vowel			2, 75.994	.362	.698
Vowel × Speaker (Group × Gender)			76, 1103	9.066	.000

Table A4.16 Multivariate Hotelling's T^2 tests, and follow-up univariate t-tests, on ΔF1 and ΔF2 fo individual L1 vowels.

Vowel	Variable(s)	Mean (ΔHz)	*df*	T^2	*F*	*t*	*p*
Sp /i/	ΔF1, ΔF2		2, 57	99.330	48.809		.000
	ΔF1	-9	58			-7.428	.000
	ΔF2	+40	58			9.043	.000
Sp /ei/	ΔF1, ΔF2		2, 57	1180.510	580.078		.000
	ΔF1	-118	58			-30.699	.000
	ΔF2	+279	58			27.245	.000
Sp /e/	ΔF1, ΔF2		2, 57	36.507	17.939		.000
	ΔF1	+3	58			1.800	.077
	ΔF2	+30	58			5.929	.000
Eng /i/	ΔF1, ΔF2		2, 47	4.564	2.2234		.118
	ΔF1	+0	48			.084	.934
	ΔF2	+13	48			2.123	.039
Eng /ɪ/	ΔF1, ΔF2		2, 47	368.444	180.384		.000
	ΔF1	+30	48			10.382	.000
	ΔF2	-79	48			-18.021	.000
Eng /e/	ΔF1, ΔF2		2, 47	397.473	194.596		.000
	ΔF1	-56	48			-13.579	.000
	ΔF2	+138	48			18.647	.000
Eng /ɛ/	ΔF1, ΔF2		2, 47	275.508	134.884		.000
	ΔF1	+45	48			13.609	.000
	ΔF2	-74	48			-13.463	.000

Table A4.17 Paired-sample *t*-tests comparing durations of pairs of L1 vowels.

Dependent Variable	Mean Duration (ms)		*df*	*t*	*p*
	Fist Vowel	Second Vowel			
Sp /i/ – Sp /e/	81	86	58	-7.840	.000
Sp /ei/ – Sp /e/	86	138	58	22.922	.000
Eng /i/ – Eng /e/	93	121	48	-27.151	.000
Eng /i/ – Eng /ɪ/	93	71	48	20.251	.000
Eng /i/ – Eng /ɛ/	93	91	48	1.686	.098
Eng /ɛ/ – Eng /ɪ/	91	71	48	18.735	.000

Table A4.18 Univariate ANOVAs comparing Spanish /i/ and English /i/, with Speaker as a random factor nested within Language and Gender. Only results for Language main effects are reported in the table.

Dependent Variable	Marginal Means (Hz / ΔHz / ms)		*df*	*F*	*p*
	Sp /i/	Eng /i/			
F1	351	350	1, 103.999	2.421	.123
F2	2324	2495	1, 104.000	12.440	.001
ΔF1	-9	+0	1, 103.996	14.789	.000
ΔF2	+40	+13	1, 103.978	12.157	.001
duration	81	93	1, 103.999	17.103	.000

Table A4.19 Univariate ANOVAs comparing Spanish /i/ and English /ɪ/, with Speaker as a random factor nested within Language and Gender. Only results for Language main effects are reported in the table.

Dependent Variable	Marginal Means (Hz / ΔHz / ms)		*df*	*F*	*p*
	Sp /i/	Eng /ɪ/			
F1	351	524	1, 103.999	662.850	.000
F2	2324	1932	1, 104.000	235.164	.000
ΔF1	-9	+30	1, 103.997	156.123	.000
ΔF2	+40	-79	1, 103.997	357.912	.000
duration	81	71	1, 103.999	23.791	.000

Table A4.20 Univariate ANOVAs comparing Spanish /e/ and English /ɪ/, with Speaker as a random factor nested within Language and Gender. Only results for Language main effects are reported in the table.

Dependent Variable	Marginal Means (Hz / ΔHz / ms)		*df*	*F*	*p*
	Sp /e/	Eng /ɪ/			
F1	513	524	1, 103.993	.335	.564
F2	1950	1932	1, 103.998	4.497	.036
ΔF1	+3	+30	1, 103.984	55.112	.000
ΔF2	+30	-79	1, 103.979	241.258	.000
duration	86	71	1, 103.994	45.026	.000

Table A4.21 Univariate ANOVAs comparing Spanish /e/ and English /ɛ/, with Speaker as a random factor nested within Language and Gender. Only results for Language main effects are reported in the table.

Dependent Variable	Marginal Means (Hz / ΔHz / ms)		*df*	*F*	*p*
	Sp /e/	Eng /ɛ/			
F1	513	677	1, 103.979	10.367	.002
F2	1950	1784	1, 103.992	51.781	.000
ΔF1	+3	+45	1, 103.952	183.407	.000
ΔF2	+30	-74	1, 103.946	116.198	.000
duration	86	91	1, 103.981	122.864	.000

Table A4.22 Univariate ANOVAs comparing Spanish /ei/ and English /e/, with Speaker as a random factor nested within Language and Gender. Only results for Language main effects are reported in the table.

Dependent Variable	Marginal Means (Hz / ΔHz / ms)		*df*	*F*	*p*
	Sp /ei/	Eng /e/			
F1	509	501	1, 103.985	5.112	.026
F2	2072	2218	1, 103.995	10.001	.002
ΔF1	-118	-56	1, 103.970	145.763	.000
ΔF2	+279	+138	1, 103.979	118.356	.000
duration	138	121	1, 103.989	21.088	.000

Appendix 5.
L1-English Production Controls

Eighteen (18) monolingual English participants took part in an L1-English vowel production control experiment in which they produced vowel in isolated words without the carrier sentence. In one control condition they produced /bVpə/ words identical to those in the original experiment, and in another control condition they produced /bVp/ words. The mean values of the acoustic properties measured are given in Tables A5.1 through A5.3 for the original and each of the control conditions. There were no significant differences between the vowels in /b_pə/ context when produced within the carrier sentence compared to when produced in isolated words (see Table A5.4). There were significant differences between the vowels when produced in isolated /b_pə/ context compared to when produced in isolated /b_p/ context (see Table A5.5, most of the univariate tests on each acoustic measure for each vowel were also significant, the results of these tests are not reported). Notable differences were that all vowels were longer in /b_p/ context, most had higher F1 and F2, and greater magnitude of VISC. Plots of vowels in isolated /b_pə/ and /b_p/ contexts are given in Figure A5.1.

Table A5.1 Mean values of acoustic measurements of L1 English vowels produced by 18 monolingual English speakers in /b_pə/ context with carrier sentence. Geometric means calculated on log scales then converted back to Hertz and milliseconds.

Vowel	F1 (Hz)	ΔF1 (Hz)	F2 (Hz)	ΔF2 (Hz)	duration (ms)
Eng /i/	339	-2	2517	+29	93
Eng /ɪ/	509	+34	1973	-78	69
Eng /e/	499	-54	2256	+156	122
Eng /ɛ/	677	+53	1804	-74	90

Table A5.2 Mean values of acoustic measurements of L1 English vowels produced by 18 monolingual English speakers in /b_pə/ context without carrier sentence. Geometric means calculated on log scales then converted back to Hertz and milliseconds.

Vowel	F1 (Hz)	ΔF1 (Hz)	F2 (Hz)	ΔF2 (Hz)	duration (ms)
Eng /i/	342	-4	2527	+35	94
Eng /ɪ/	508	+39	1990	-80	72
Eng /e/	499	-55	2257	+163	122
Eng /ɛ/	677	+51	1812	-75	89

Table A5.3 Mean values of acoustic measurements of L1 English vowels produced by 18 monolingual English speakers in /b_p/ context without carrier sentence. Geometric means calculated on log scales then converted back to Hertz and milliseconds.

Vowel	F1 (Hz)	ΔF1 (Hz)	F2 (Hz)	ΔF2 (Hz)	duration (ms)
Eng /i/	350	-5	2553	+76	126
Eng /ɪ/	530	+63	2060	-85	102
Eng /e/	495	-66	2297	+210	144
Eng /ɛ/	687	+75	1878	-75	114

Table A5.4 Hotelling's T^2 tests on multivariate paired samples: monolingual English speakers' vowels produced in /b_pə/ context with carrier sentences vs without carrier sentences.

Vowels	T^2	F	df	p
Eng /i/	7.58	1.16	5, 13	.3798
Eng /ɪ/	6.34	0.97	5, 13	.4714
Eng /e/	3.78	0.58	5, 13	.7166
Eng /ɛ/	3.13	0.48	5, 13	.7865

Table A5.5 Hotelling's T^2 tests on multivariate paired samples: monolingual English speakers' vowels produced without carrier sentences, /b_pə/ context vs /b_p/ context.

Vowels	T^2	F	df	p
Eng /i/	388.95	59.49	5, 13	.000
Eng /ɪ/	501.40	76.69	5, 13	.000
Eng /e/	202.78	31.01	5, 13	.000
Eng /ɛ/	175.10	26.78	5, 13	.000

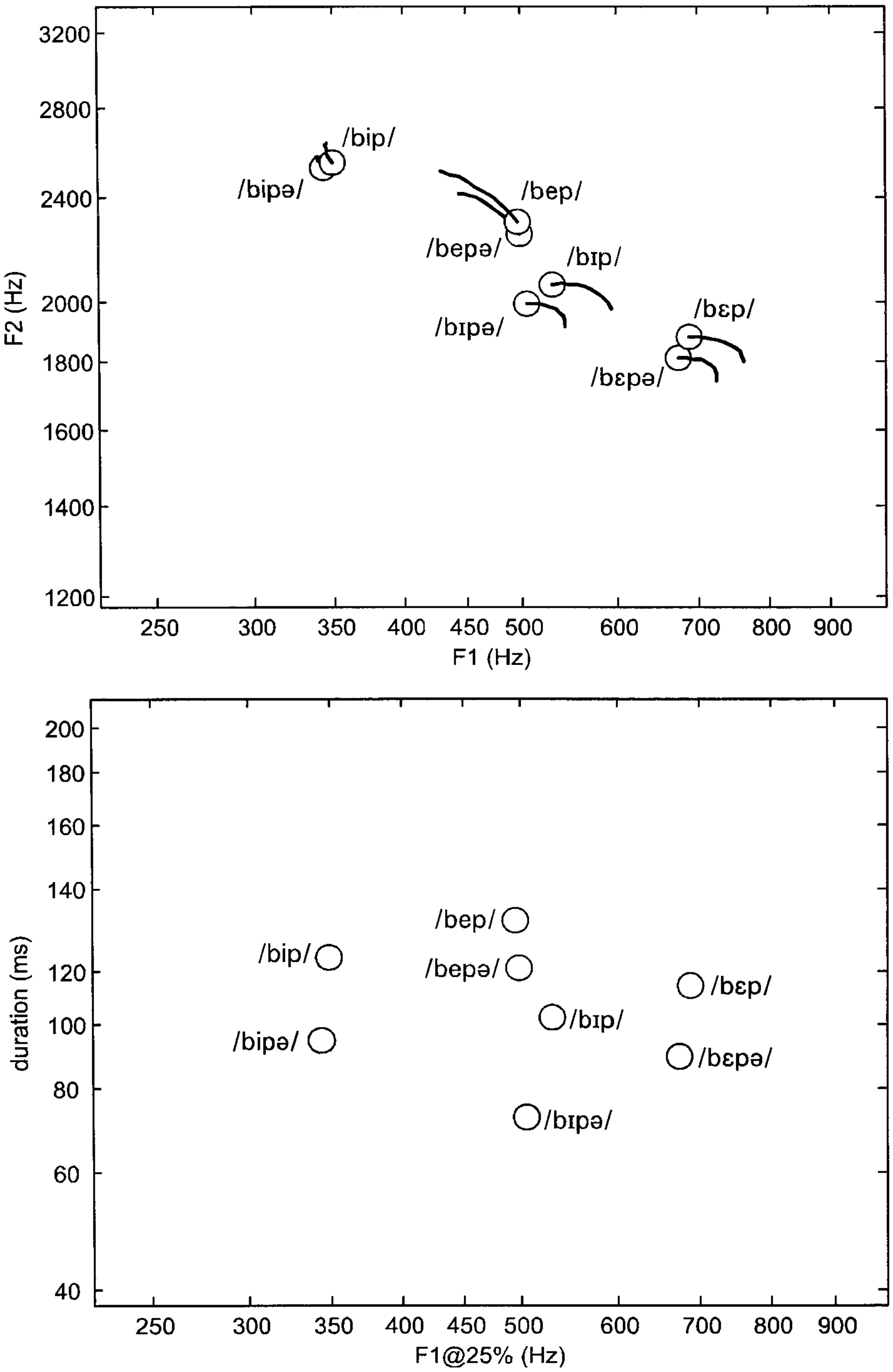


Figure A5.1 Non-normalised mean acoustic properties of L1-Spanish and L1-English vowels produced in isolated words. Top: F1, F1, ΔF1, and ΔF2. Comet heads indicate formant values at 25% of the duration of the vowel, ends of comet tails indicate formant values at 75% of the duration of the vowel. Bottom: F1 at 25% of the duration of the vowel and vowel duration.

Appendix 6.
Vowel Normalisation

Prior to the canonical discriminant function analyses, formant values were normalised using a variant of log mean normalisation (Morrison & Nearey, In press; Nearey, 1978, 1989; Nearey & Assmann, In press):

- For each speaker, a single mean, $\overline{G_s}$, was calculated for the log-Hertz values of F1 and F2 in all L1 vowels. Means were first calculated over the entire formant tracks from 25–75% of the duration of each vowel, then over each vowel category, and then over all L1 vowel categories. If more than one set of L1 vowels were available from a given speaker, only the first set of vowels was used to calculate the normalisation factor. $\overline{G_s}$ can be considered a default position for this set of vowels for speaker *s*, which is dependent on non-linguistic factors such as differences due to vocal tract length.
- Each speaker's mean, $\overline{G_s}$, was subtracted from each individual F1 and F2 measurement for that speaker, resulting in each speaker's formant measurements being centred around their own mean value. Individual formant values are now expressed as deviations from $\overline{G_s}$, coded as deviations from zero. Assuming that all speakers of a given language / dialect have the same vowel F1–F2 ratio pattern, individual differences have now been removed, and only the vowel formant ratio pattern remains.

Because the vowel formant ratio patterns are expected to vary across languages, differing in number of vowels, symmetry, or formant range, an ideal bilingual speaker (vowel patterns identical to vowel patterns of monolingual speakers of each language) would not be expected to have the same $\overline{G_s}$ in both languages. For example, in the present study, four English vowels are examined compared to only three in Spanish, and (considering only F1) English /ɛ/ has higher F1 values than any Spanish vowel, so the ideal bilingual's $\overline{G_s}$ for English would be expected to be higher than their $\overline{G_s}$ for Spanish. Therefore, when L2-

English vowels are compared with L1-English vowels, and L2-Spanish vowels are compared with L1-Spanish vowels, additional normalisation steps are required:

- Within each L1, the mean of the $\overline{G_s}$ are calculated, $\overline{G}_{Spanish}$ and $\overline{G}_{English}$. This is then used to calculate an inter-language normalisation factor: $\overline{G}_{Spanish}$ - $\overline{G}_{English}$.
- Each speaker's L2 vowels are speaker-normalised via subtraction of the $\overline{G_s}$ calculated using only the vowels of their L1, then inter-language-normalised via addition or subtraction, as appropriate, of the inter-language normalisation factor.

The calculation of the inter-language normalisation factor assumes that the only differences between the speakers in each L1 group are inter-language differences. In the present study, the ratio of male to female participants was 27:32 for L1-Spanish and 17:30 for L1-English; therefore, the inter-language normalisation factor was calculated on using data from all L1-English participants and 17 male and 30 female L1-Spanish participants selected at random.

The synthetic stimuli happened to have been based on a male voice; therefore, to facilitate comparison of production results with results for perception of the synthetic stimuli, normalised formant frequencies were adjusted so that they were centred around the mean of the male speakers' $\overline{G_s}$ within each language. Following the subtraction of each speaker's $\overline{G_s}$ from their individual F1 and F2 measurements, and when appropriate the addition of the inter-language normalisation factor, $\overline{G}_{male}$ for the target language was added.

Since female speakers produced longer vowels than male speakers (see Appendix 4, Tables A4.5a and A4.10), vowel duration was independently normalised using the same procedure as used to normalise vowel formants. Means were calculated on log-millisecond values.

Appendix 7.
Logistic Regression Modelling

Logistic regression is a statistical method suitable for analysing identification response data from speech perception experiments.[1] This appendix is intended to be an introduction to understanding logistic regression applied to first and second language (L1 and L2) speech perception research. I will illustrate some of the ways in which logistic regression can be applied using relatively simple data sets, readers should then find it easier to understand the more complex analyses in speech perception papers such as Benkí (2001) Maddox, Molis, & Diehl (2002), Nearey (1990, 1997), and Morrison (2005a, 2005b), as weel as the present study. For general introductions to applied logistic regression see Hosmer & Lemeshow (2000), Menard (2001), and Pampel (2000).

A7.1. Fitting A Logistic Regression Model

A7.1.1. One stimulus dimension, binomial responses

In speech perception research, the basic goal of logistic regression analysis is to fit a sigmoidal (S-shaped) curve to categorical response data. Consider a classic voice onset time (VOT) experiment in which there is a single acoustic dimension, VOT ranging from 0 to 60 ms in 10 ms intervals, and there are two response categories, voiced or voiceless (one stimulus dimension and binomial/dichotomous responses). Imagine the following idealised response data: A participant hears the stimuli 10 times in random order and gives 10 voiceless responses for all the stimuli with VOT < 20 ms, 8 voiceless and 2 voiced responses for the stimulus with VOT = 20 ms, 2 voiceless and 8 voiced responses for the stimulus with VOT = 30 ms, and 10 voiced responses for all the stimuli with VOT > 30 ms. This is proportional data: the proportion of voiced responses is 0 for all stimuli with VOT < 20 ms, .2 for the stimulus with VOT = 20 ms, .8 for the stimulus with VOT = 30 ms, and 1 for all stimuli with VOT > 30 ms. The observed proportions of voiced responses are plotted in Figure A7.1, as are the sigmoidal curves fitted via a logistic regression analysis to the

[1] Although less flexible, another suitable method is probit analysis.

proportions of voiced and voiceless responses.

The fitted curves are not a perfect fit to the data, for example the predicted probability of a voiced response at 20 ms is .172 rather than the observed value of .200. However, the curve is generally very close to the data points. Goodness of fit can be assessed in several ways. A standard method is to measure the distance between the observed and predicted values for each stimulus and take an average over all the stimuli: Root-mean-squared (RMS) error is the sum of the squares of the differences between the observed and predicted values (sum of squared errors), divided by the residual degrees of freedom in the model, then square rooted.[2] RMS error can be scaled by the number of responses per stimulus to give a percentage root-mean-squared error (%RMS). The RMS error for the logistic regression model fitted to the data in Figure A7.1 is 2.6%. Another intuitive measure of goodness of fit is the percentage modal agreement (%MA), the percentage of times, over all the stimuli, that the most likely response predicted by the model matches the most common (the modal) response of the listener. If getting the category right is what counts, then %MA may be a more meaningful measure. The MA for the logistic regression model fitted to the data in Figure A7.1 is 100%. The goodness of fit measure actually used when fitting logistic regression models is the deviance statistic (G^2).[3] Compared to RMS error, the G^2 statistic is less intuitively meaningful, but, like RMS error, it decreases as goodness of fit improves.

[2] The number of residual degrees of freedom is the number of independent pieces of information in the model. For the models here, this is the number of stimuli multiplied by one less than the number of response categories, minus the number of non-redundant coefficients estimated in the model. (Since the responses are proportions, they must sum to 1 and the proportions for the last category are redundant.) There are 7 stimuli and 2 response categories in the VOT data, and 2 coefficients/parameters in the logistic regression model fitted to the data; therefore there are 5 residual degrees of freedom in the model.

[3] G^2: For each stimulus, the observed values of the responses multiplied by the natural logs of the observed values of the response divided by the model's predicted values, then summed over all stimuli and multiplied by two.

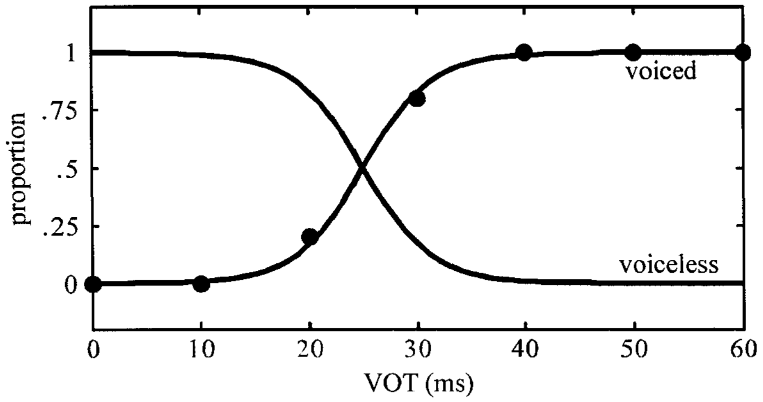


Figure A7.1 Sigmoidal logistic regression curves fitted to idealised VOT data. Dots represent proportions of voiced responses observed in the data.

Several factors can affect goodness of fit. One factor is the appropriateness of the model: Clearly the sigmoidal curve of a logistic regression model is a better fit to our data than would be the straight line of a linear regression model. In some cases the appropriateness of the model, or lack thereof, may not be so apparent, Hosmer & Lemeshow (2000: §5.3) discuss this issue in detail. For formant values in vowel stimuli, goodness of fit typically improves when frequency is entered in to the model in log Hertz (or mel, Bark, or ERB) rather than in Hertz. Since human frequency perception is logarithmic rather than linear, a model fitted to log Hertz values is usually more appropriate than a model fitted to Hertz values. Another factor which can decrease goodness of fit is noise in the data: If the listener is occasionally distracted, they may fail to hear a stimulus and press a response button at random. A certain number of responses in the data will then be from a random distribution which does not reflect the listener's perception of the stimuli. If the number of random responses is relatively small, they may have relatively little effect on the location and shape of the fitted curve; however, the random responses will cause the observed values for some stimuli to be further from the curve than they would otherwise be, and so decrease the goodness of fit (noise will also usually cause the slopes of the curves to be shallower). Yet another factor that can decrease goodness of fit, is the use of data pooled across listeners. It could be that a logistic regression model fits each individual's data well, but that the exact location of the category boundaries vary across listeners, and hence the boundaries in the pooled data are fuzzier than each individual listener's boundaries. Although problematic for statistical analysis, use of pooled data may be justified on linguistic grounds: If the listeners are all native speakers of the same dialect then they will have similar pronunciation and perception, and any interlistener differences will be negligible for communication purposes. A population average model based on data pooled across listeners may reasonably characterise the perception of a group of native speakers of a given dialect.

A7.1.2. Multiple stimulus dimensions, multinomial response categories

Let us turn to some real data. Álvarez González (1980: Ch. 3) investigated native Spanish listeners' perception of a synthetic vowel space in which F1 varied from 250–800 Hz in 9 steps, F2 varied from 750–2700 Hz in 8 steps, and F3 varied from 2300–2900 Hz in 2 steps (a total of 231 stimuli since the corner where F1 would have been higher than F2 was excluded). Fifty listeners heard each stimulus once in random order in the context /_ɾa/, and responded by circling orthographic 'ara', 'era', 'ira', 'ora', or 'ura' on an answer sheet, thereby identifying each synthetic vowel as one of the Spanish vowels, /a/, /e/, /i/, /o/, or /u/. This constitutes three stimulus dimensions and five response categories.

The software that we will use to build logistic regression models of multinomial/polytomous response data was implemented by Terrance M. Nearey based on an algorithm described in Haberman (1979), and is available as Matlab code upon request either from Nearey <t.nearey@ualberta.ca> (with some additional effort, most of the analyses described below could also be conducted using commercial software such as SPSS or STATA, or free software such as R). Logistic regression operates in a logistic (log odds) space,[4] and fits a model by maximising the G^2 goodness of fit to the data using an iterative maximum likelihood technique. This results in a series of logistic regression coefficients. For models that will be fitted to the Álvarez González data, these will be:[5]

[4] Non-linear probability values can be transformed into linear logit values, see Pampel (2000: Ch. 1). In the case of the VOT data, the odds of a voiced response is the ratio of the probability of a voiced response to the probability of a voiceless response *odds*(voiced) = *p*(voiced) / *p*(voiceless). The logit is the natural logarithm of the odds *L*(voiced) = *log*(*odds*(voiced)).

[5] It is also possible to build more complex models including coefficients for quadratic and crossproduct terms, etc.

bias coefficients:	$\alpha_{/a/}$, $\alpha_{/e/}$, $\alpha_{/i/}$, $\alpha_{/o/}$, $\alpha_{/u/}$
F1-tuned coefficients:	$\beta_{/a/F1}$, $\beta_{/e/F1}$, $\beta_{/i/F1}$, $\beta_{/o/F1}$, $\beta_{/u/F1}$
F2-tuned coefficients:	$\beta_{/a/F2}$, $\beta_{/e/F2}$, $\beta_{/i/F2}$, $\beta_{/o/F2}$, $\beta_{/u/F2}$
F3-tuned coefficients:	$\beta_{/a/F3}$, $\beta_{/e/F3}$, $\beta_{/i/F3}$, $\beta_{/o/F3}$, $\beta_{/u/F3}$

These include redundant coefficients, the value of the fifth coefficient in each family of coefficients (α, β_{F1}, β_{F2}, β_{F3}) is known once the values of the other four coefficients are known: We use deviation-from-mean coding, hence the sum of the values of the coefficients in each family is zero, and the value of the fifth coefficient is minus the sum of the other four coefficients.

Questions which we will ask about the perception of Spanish vowels in the data from Álvarez González are:

– Does the Spanish listeners' vowel perception depend on F1 and F2?

– Does the Spanish listeners' vowel perception depend on F3 in addition to F1 and F2?

We will answer the questions by comparing the difference in goodness of fit between different logistic regression models fitted to the response data. If a model that contains F1 and F2 fits the data better than a model which does not contain F1 and F2, then this indicates that the listeners' vowel perception depends on F1 and F2. Likewise, if a model that contains F3 fits the data better than a model which does not contain F3, then this indicates that the listeners' vowel perception depends on F3. The models we will fit are the following:

$$\mathrm{V} \tag{1a}$$

$$\mathrm{V} + \mathrm{V}\times F1 + \mathrm{V}\times F2 \tag{1b}$$

$$\mathrm{V} + \mathrm{V}\times F1 + \mathrm{V}\times F2 + \mathrm{V}\times F3 \tag{1c}$$

Model 1a contains only the bias coefficients for each vowel response, Model 1b contains bias coefficient and F1 and F2 stimulus-tuned coefficients for each vowel, and Model 1c additionally contains F3 stimulus-tuned coefficients for each vowel. The difference in goodness of fit of nested models (models where the smaller model contains a subset of the parameters in the larger model) can be statistically assessed using the difference in the G^2 statistic (ΔG^2, also known as the -2 log likelihood ratio). Assuming pure multinomial error, ΔG^2 is asymptotically distributed as a χ^2 with degrees of freedom equal to the difference in degrees of freedom between the two models. However, if there is overdispertion/heterogeneity in the data, such as may arise when data is pooled over participants, then the ΔG^2 test may suffer from a serious Type II error and indicate a significant difference when the difference is in fact not significant. One approach to dealing with this problem (provided in Nearey's software), is to use a quasi-likelihood F-test, where the F-ratio is the result of dividing the ΔG^2 by the overdispersion factor (the overdispersion factor is calculated as the ratio of the Pearson χ^2 to the residual degrees of freedom)[6] and the degrees of freedom in the F-test are the difference in degrees of freedom between the two models and the residual degrees of freedom of the larger model (see McCullagh & Nelder 1983; and Nearey 1990, 1997).[7]

Table A7.1 shows the G^2, %RMS error, and %MA for each model fitted to the response data. F1 and F2 were converted to the natural logarithms of their Hertz values before fitting the logistic regression models. Table A7.2 shows the ΔG^2, overdispertion, and quasi-likelihood F-ratio for comparisons of model 1b with 1a, and 1c with 1b. Adding F1 and F2 stimulus tuning to a model containing only bias coefficients (1b vs 1a) resulted in a large (22.8 percentage point) decrease in %RMS error, and a large (54.9 percentage point) increase in %MA, and the increase in goodness of fit was statistically significant on the quasi-likelihood F-test. Therefore it can be concluded that the listeners' vowel responses did depend on F1 and F2.

Adding F3 stimulus tuning to a model already containing bias coefficients and F1 and

[6] The Pearson χ^2: For each stimulus, the square of the difference between the observed values of the responses and the model's predicted values, then divided by the model's predicted values, then summed over all stimuli.

[7] McCullagh & Nelder (1983) advise using a fixed overdispertion, typically from the largest model considered. The 1b versus 1a comparison would still be significant on the quasi-likelihood F test if the overdispertion from Model 1c were used.

F2 tuning (1c vs 1b) resulted in a small (0.2 percentage point) decrease in %RMS error, and a small (0.8 percentage point) decrease (rather than increase) in %MA, and the increase in goodness of fit was not statistically significant on the quasi-likelihood *F*-test. There is therefore little reason to believe that the listeners' vowel responses depended on F3 in addition to F1 and F2.

Table A7.1 Goodness of fit measures for models fitted to the vowel perception data from Álvarez González.

Model	*df*	G^2	χ^2	%RMS	%MA
1a	920	43647	54031	34.6	35.1
1b	912	8105	125912	11.8	90.0
1c	908	7911	130478	11.6	89.2

Table A7.2 Comparisons of goodness of fit measures for models fitted to the vowel perception data from Álvarez González.

Models compared	Δdf	*df* residual	ΔG^2	$p(\Delta G^2)$	over-dispertion	*F*	$p(F)$
1b vs 1a	8	912	35542	.000	58.7	75.65	.000
1c vs 1b	4	908	194	.000	138.1	0.35	.843

A7.2. Interpreting Logistic Regression Coefficients

A7.2.1. Graphical representations

A third question that will be asked about the perception of Spanish vowels in the data from Álvarez González is:

– How do F1 and F2 affect Spanish listeners' vowel perception?

One way to answer this question is via graphical representations of the logistic regression model of listeners' perception. The estimated logistic coefficient values calculated for Model 1b are shown in Table A7.3 and the stimulus-tuned coefficient values are plotted in Figure A7.2. The relative locations of the perceptual vowel response categories in the F1-tuned-coefficient–F2-tuned-coefficient space in Figure A7.2 is reminiscent of the distribution of vowel production values in the F1–F2 space; correlation of coefficients with production patterns are frequently found in logistic regression analyses. The direct interpretation of the stimulus-tuned coefficients will be discussed below in section 2.2.

Table A7.3 Estimated values of logistic regression coefficients for Model 1b fitted to the vowel perception data from Álvarez González.

bias coefficients		F1-tuned coefficients		F2-tuned coefficients	
α/a/	-35.667	β/a/F1	6.804	β/a/F2	-1.059
α/e/	-40.832	β/e/F1	1.240	β/e/F2	4.774
α/i/	1.519	β/i/F1	-5.664	β/i/F2	4.561
α/o/	14.618	β/o/F1	3.982	β/o/F2	-5.405
α/u/	60.362	β/u/F1	-6.362	β/u/F2	-2.870

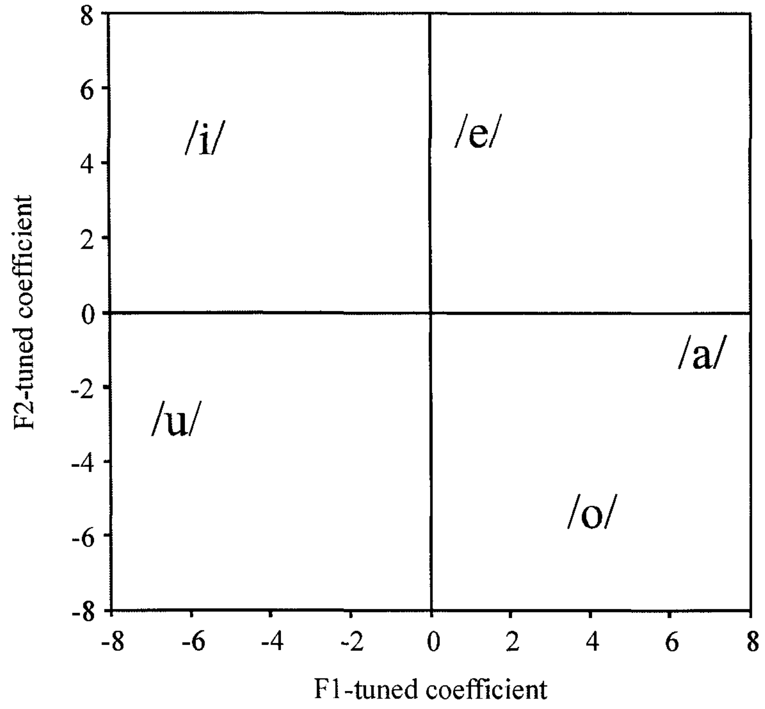


Figure A7.2 Plot of estimated values of stimulus-tuned logistic regression coefficients for Model 1b fitted to the vowel perception data from Álvarez González.

In order to obtain a predicted logistic value for a given category at a given set of stimulus values, the F1 and F2 values and logistic regression coefficient values are substituted into Equation 1b and all coefficients that do not correspond to the given category are set to zero. For example, to obtain the predicted logistic value for the response /u/, $L_{/u/}$, at F1 = 250 Hz, F2 = 800 Hz, the values would be substituted into the following equation:

$$L_{/u/} = \alpha_{/u/} + \beta_{/u/F1}\times F1 + \beta_{/u/F2}\times F2 \qquad (2)$$

$$L_{/u/} = 60.362 - 6.362\times\log(250) - 2.870\times\log(800) \quad = \quad 6.050$$

The predicted probability of /u/, $p_{/u/}$ is given by:

$$p_{/u/} = \frac{e^{L_{/u/}}}{\sum_{x} e^{L_x}} \qquad (3)$$

$$p_{/u/} = e^{6.050} / (e^{-5.178} + e^{-2.073} + e^{0.734} + e^{0.474} + e^{6.050}) \quad = \quad 0.991$$

where x takes on the values of all the response categories {/a/, /e/, /i/, /o/, /u/} and all L_x are calculated for the same set of F1 and F2 values.

If a range of F1 and F2 values covering the stimulus space are substituted into equations of the type given in Equations 2 and 3, the predicted probability of each vowel response category can be calculated over the two-dimensional stimulus space and plotted in a three-dimensional probability surface plot as in Figure A7.3. The height of a surface above the base of the plot indicates the predicted probability of the response associated with that surface. The predicted probability of an /u/ response is close to 1 for low-F1–low-F2 values and decreases sigmoidally as either F1 or F2 or both increase. Response categories /i/, /e/, and /o/ have their highest predicted probabilities in the other corners of the stimulus space. The predicted probability of an /a/ response is highest for high-F1 and intermediate-F2 values. The maximum predicted probability of an /a/ response is quite low compared to the maximum predicted probabilities of the other response categories (the number of /a/ responses in the raw data is low, this is not an analytical error).

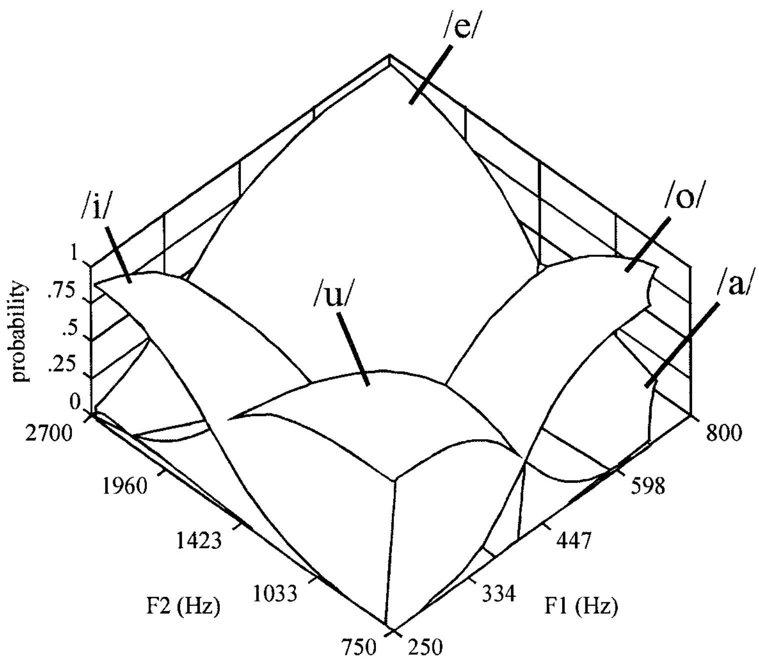


Figure A7.3 Probability surface plot based on logistic regression Model 1b fitted to the vowel perception data from Álvarez González. The height of a surface about the base of the plot indicates the predicted probability of the corresponding response category.

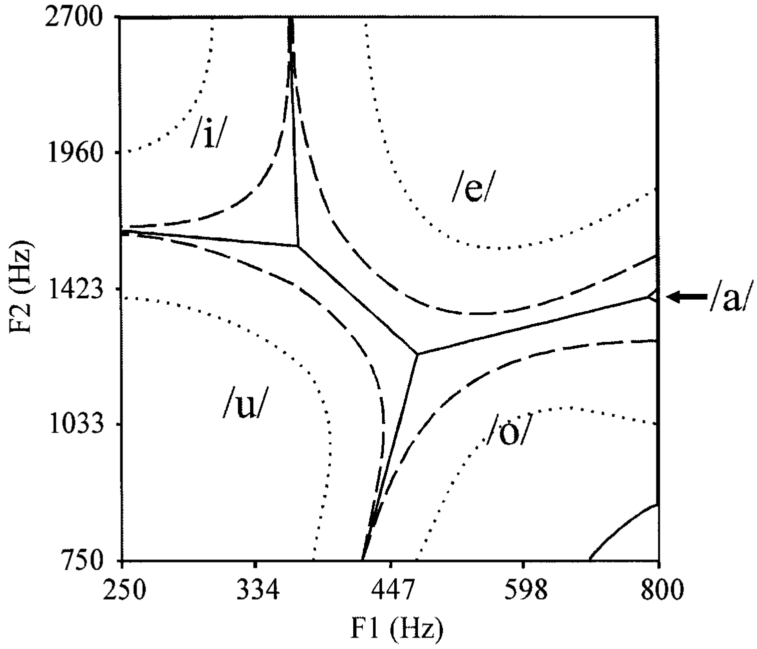


Figure A7.4 Territorial map based on logistic regression Model 1b fitted to the vowel perception data from Álvarez González.

Figure A7.4 is a two-dimensional territorial map, it is equivalent to a view of the three-dimensional probability surface plot (Figure A7.3) from directly above the stimulus plane. Only the response with the highest predicted probability is visible in any part of the stimulus space. The solid lines represent the location of perceptual boundaries between vowels, on one side of the boundary one vowel is the more probable response, on the other side another vowel is more probable. The dashed and dotted lines represent the .5 and .75 predicted probability contours for the locally dominant categories. The /i/–/e/ boundary is at lower F1 values than the /u/–/o/ boundary; this perceptual result corresponds to the finding that Spanish speakers produce /e/ with lower F1 than /o/ (Álvarez González 1980: §2.7).

A7.2.2. Boundary crispness

A stimulus-tuned logistic regression coefficient represents the slope of a line in the logistic space. With deviation-from-mean coding, the rate of change from one category to another along a dimension in the logistic space is the difference between the stimulus-tuned logistic regression coefficient for each category (the distance between the centres of the vowel labels in Figure A7.2). For example, in Model 1b fitted to the Álvarez González data, the rate of change from /i/ to /e/ as F1 increases is $\beta_{/e/F1} - \beta_{/i/F1} = -1.240 - 5.664 = 4.424$ per log Hertz (the rate of change from one category to another will be referred to below as the *contrast coefficient*).[8]

The contrast-coefficient slope in the logistic space is related to the slope of the sigmoidal curve representing the rate of change from one category to another in the probability space. For expository purposes, we will return to the binomial VOT example. In a binomial model the slope of the steepest tangent to the sigmoidal rate of change curve in the probability space is one-quarter the slope of the contrast-coefficient[9] line in the logistic

[8] Rates of change for any category contrast can be calculated along any line in the stimulus space. For example, the rate of change from back vowel to front vowel identification as F2 increases: $(\beta_{/i/F2} + \beta_{/e/F2}) - (\beta_{/u/F2} + \beta_{/o/F2})$ per log Hertz. Or the rate of change from /i/ to /e/ for a one log Hertz increase in F1 and a two log Hertz decrease in F2: $(\beta_{/e/F1} - 2\times\beta_{/e/F2}) - (\beta_{/i/F1} - 2\times\beta_{/i/F2})$.

[9] In the binomial case, one would usually use reference-category rather than deviation-from-mean coding. The coefficient values for one category would be fixed at zero and (what I call) the contrast coefficients would be the only coefficients reported by the software. If reference-category coding had been adopted in the multinomial

space.[10] The size of the contrast coefficient and the corresponding steepness of the steepest tangent to the sigmoidal curve in the probability space are indicators of the crispness of the boundary between the two categories. The logistic regression model fitted to the idealised VOT data has a voiceless to voiced contrast coefficient, $\beta_{(\text{voiced-voiceless})\text{VOT}}$ (hereafter β_{VOT}), of 0.314 logits per millisecond = a maximum rate of change in the probability of .079 per millisecond. Figures A7.5a and A7.5b show plots of the linear slope in the logistic space and the sigmoidal curve in the probability space, based on a contrast coefficient value four times that of the contrast coefficient value from the model fitted to the VOT data. The sigmoidal curve is almost steplike: as the VOT increases, the probability of a voiceless response is essentially 0 until very close to the boundary, then jumps to essentially 1. This is therefore a very crisp categorical boundary. Figures A7.5c and A7.5d show plots of the linear slope in the logistic space and the sigmoidal curve in the probability space, based on a contrast coefficient value one fourth that of the contrast coefficient value from the model fitted to the VOT data. The sigmoidal curve is almost linear with a gradual increase in the probability of a voiced response from 0 to 60 ms VOT. This is therefore a very fuzzy categorical boundary.

model of the Álvarez González data, the reference category, e.g., /u/, would have been at the origin of Figure A7.2, and the other categories would have been shifted but maintained the same relative locations.

[10] The instantaneous value of the probability slope is the (partial) derivative of the probability with respect to the dimension of interest. Using the binomial VOT example, this is: $dp/d\beta_{\text{VOT}} = \beta_{(\text{voiced-voiceless})\text{VOT}} \times p(\text{voiced}) \times p(\text{voiceless})$ (see Pampel 2000: 24). The steepest tangent occurs at the intersection between the lines/surfaces representing the probability of each category. In the binomial case each category has a probability of .5 at the intersection, hence the instantaneous slope at this point is: $\beta_{(\text{voiced-voiceless})\text{VOT}} \times .5 \times .5 = \beta_{(\text{voiced-voiceless})\text{VOT}} \times .25$. In multinomial cases, the calculation of the slope of the maximum tangent to the sigmoidal rate of probability change curve between two categories is complicated by the fact that other categories may have non-zero predicted probabilities at the intersection of the two categories of interest, thus each category of interest will not have .5 probability at the intersection. However, a larger contrast coefficient value will still indicate a larger value for the maximum slope of a tangent to the sigmoidal rate of probability change curve.

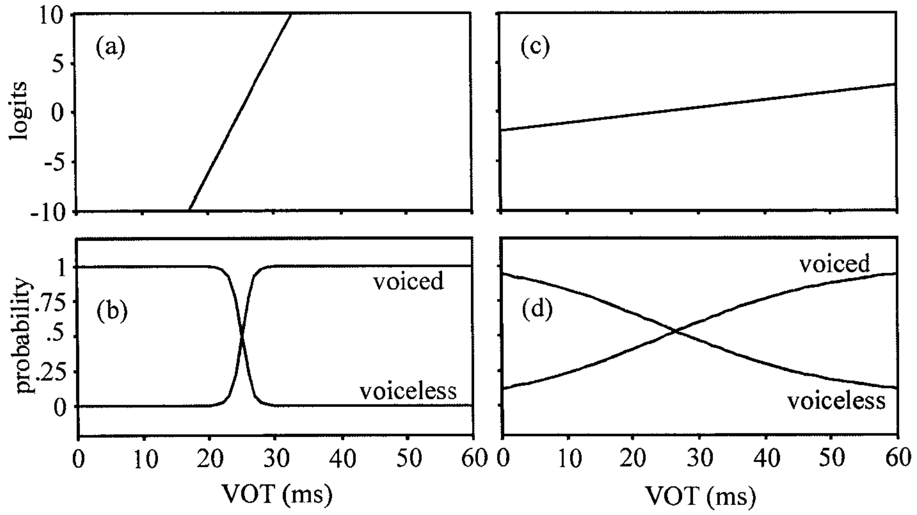


Figure A7.5 Linear slopes in the logistic space (a and c) and the corresponding sigmoidal curves in the probability space (b and d) for contrast coefficient values of 1.256 logits/ms (a and b) and 0.079 logits/ms (c and d).

Measures of boundary crispness or fuzziness are useful when analysing L2 perception data. Native speakers typically have crisp boundaries between categories, similar to Figure A7.5b. L2 learners may not have L1 categories distinguished by the same acoustic cues as the L2 categories, the L1 may not use an acoustic dimension that is used in the L2, or the range of values sampled along the dimension may all fall within a single L1 category. In such cases, the L2 learners would be expected to have very fuzzy boundaries, similar to Figure A7.5d. Even though their L1 may not provide them with a crisp categorical boundary, they may still be able to hear differences along the acoustic dimensions under study and respond in a gradient manner, e.g., giving more voiced responses for longer VOT, and thus have a non-zero contrast coefficient. As they learn the L2, they would be expected to approximate the perception of native speakers of the L2, their categorical boundaries would become crisper, and this would be reflected in the contrast coefficient values from logistic regression models fitted to their perception data.

A7.2.3. Polar-coordinate contrast coefficients

We will now turn to an example of the use of logistic regression contrast coefficients applied to real L2 perception data. In Escudero & Boersma (2004), L1-English and L1-

Spanish L2-English listeners gave English /i/ or /ɪ/ responses to a synthetic vowel continuum that varied orthogonally in spectral and duration properties. Morrison (2005b) fitted logistic regression models to individual participant's responses in Escudero & Boersma's data and derived /i/–/ɪ/ contrast coefficients β_{spec} and β_{dur} along the spectral and duration dimensions. The contrast coefficient values for the 20 L1-English speakers from the south of England, and for the 14 L1-Spanish listeners learning the Southern England dialect of English, are plotted in Figure A7.6.

The L1-Spanish listeners had significantly larger duration-tuned contrast coefficients and significantly smaller spectral-tuned contrast coefficients: Welch's *t* tests β_{dur} $t(26.589) = 3.951$, $p < .01$, β_{spec} $t(27.858) = -4.742$, $p < .001$. This was taken as evidence that, compared to the L1-English listeners, the L1-Spanish listeners made greater use of duration and less use of spectral properties when distinguishing English /i/ and /ɪ/.

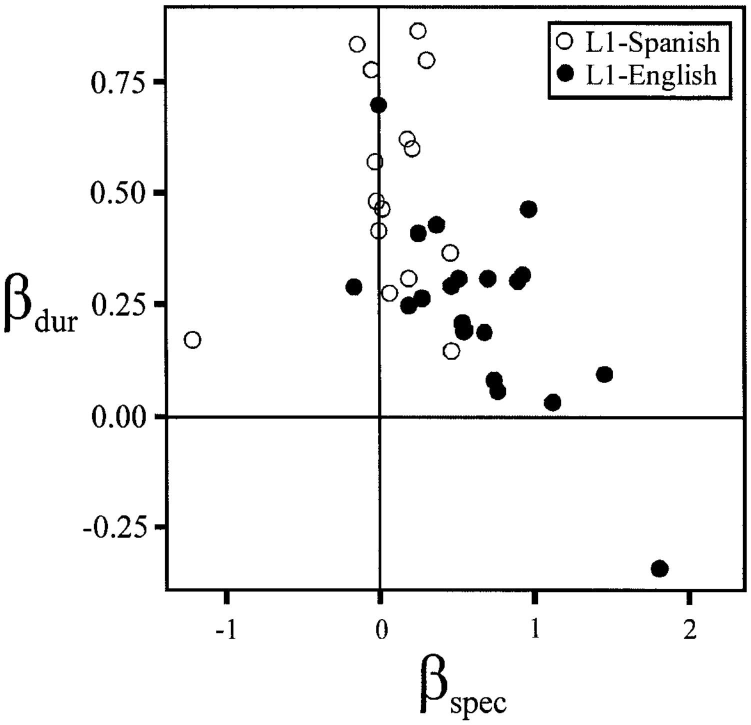


Figure A7.6 Contrast coefficients from logistic regression models fitted to individual participant data from Escudero & Boersma.

Boersma & Escudero (2005) pointed out that because of constraints imposed by the edges of the stimulus space, the spectrally-tuned and duration-tuned contrast coefficients were partially correlated, and recommended using the ratio of the two contrast coefficients in the same manner as they had used the ratio of their spectral and duration reliance measures. The ratio of the spectrally-tuned and duration-tuned contrast coefficients give the orientation of the /i/–/ɪ/ boundary in the spectral–duration stimulus space (the ratio is a gradient which can be converted to an angle in degrees). Rather than simply taking the ratio of the two contrast coefficients, they can be converted into polar coordinates to provide orthogonal measures of 1. the orientation of the boundary in the spectral–duration stimulus space (polar-coordinate angle), and 2. the boundary crispness (polar-coordinate magnitude), i.e., the rate of change from one category to the other in the direction perpendicular to the orientation of the boundary.[11] The use of polar coordinates provides intuitive numerical descriptors for the boundary. Figure A7.7 provides probability surface plots which give examples of different boundary angles and magnitudes. The angles were calculated such that an angle of 90° would indicate that the listener used only spectral cues, and an angle of 0° would indicate that the listener used only duration cues.

The L1-English listeners /i/–/ɪ/ boundary angles were significantly greater than those of the L1-Spanish L2-English listeners, $t(32) = 5.503$, $p < .001$, but the /i/–/ɪ/ boundary magnitudes were not significantly smaller, $t(32) = 1.367$, $p = .181$. Again we conclude that, compared to the L1-English listeners, the L2-English listeners made greater use of duration, but we did not find evidence in support of the hypothesis that the L2 learners would have fuzzier boundaries.

[11] angle = $arctan(\beta_{spec}/\beta_{dur})$ magnitude = $\sqrt{\beta_{spec}^2 + \beta_{dur}^2}$

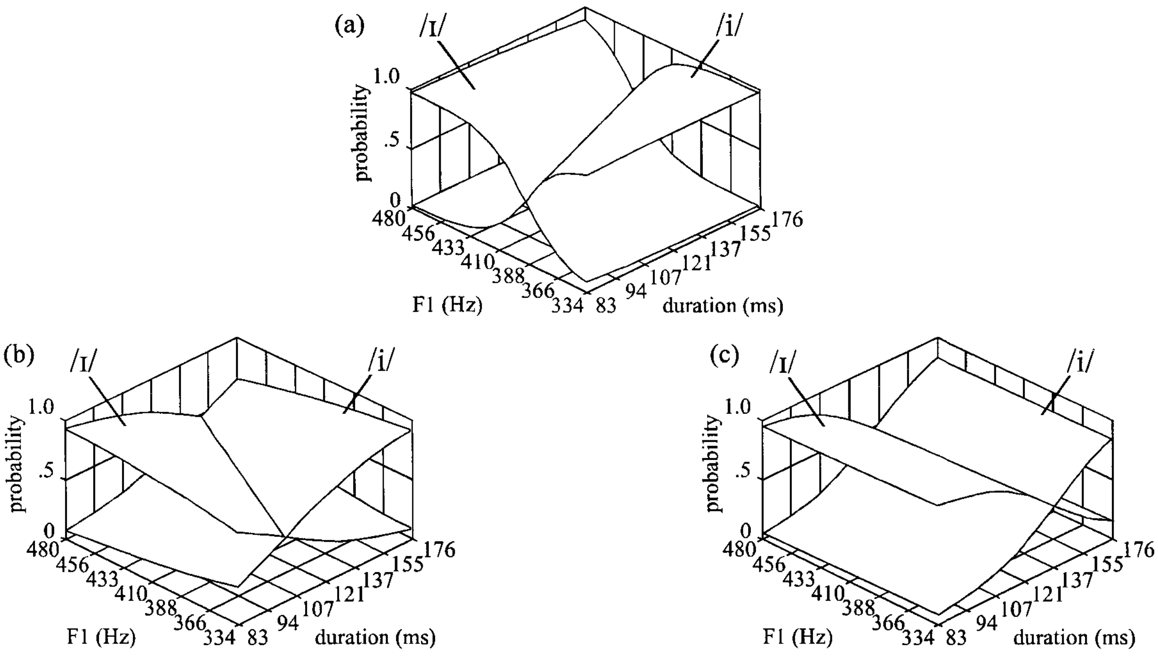


Figure A7.7 Probability surface plots illustrating different boundary angles and magnitudes.

(a) L1-English listener, angle 70° magnitude 0.88

(b) L2-English listener, angle 27° magnitude 0.35

(c) L2-English listener, angle -2° magnitude 0.46

A7.3 Conclusion

This appendix introduced logistic regression analysis as applied to the type of speech perception data collected in identification experiments. Comparison of goodness of fit of different logistic regression models was demonstrated as a means of determining which acoustic cues listeners use when identifying stimuli. This chapter also demonstrated the use of logistic regression coefficients to describe listeners' perceptual use of acoustic cues. Logistic regression coefficients can be used to produce intuitive detailed graphical representations of listeners' use of perceptual cues, they provide a metric of intercategory boundary orientation and crispness, and can also be used as statistics in secondary analyses which test the differences in perception between predefined groups.

Appendix 8.
L1-English Perception Controls

In order to obtain differences between L1-mode and L2-mode perception, bilingual participants completed two perception experiments, one in their L1 and another in their L2. However, some differences between the two sets of response data could be due to non-language mode effects, and control experiments were conducted to ascertain the magnitude of any such effects.

Repeating the same task again provides a second sample of the listeners true perception, and random variation in sampling could result in differences in the results. On the second repetition a practice effect could also affect the results. To test for such differences, five monolingual English participants completed a control experiment which was an exact replication of the original English perception experiment.

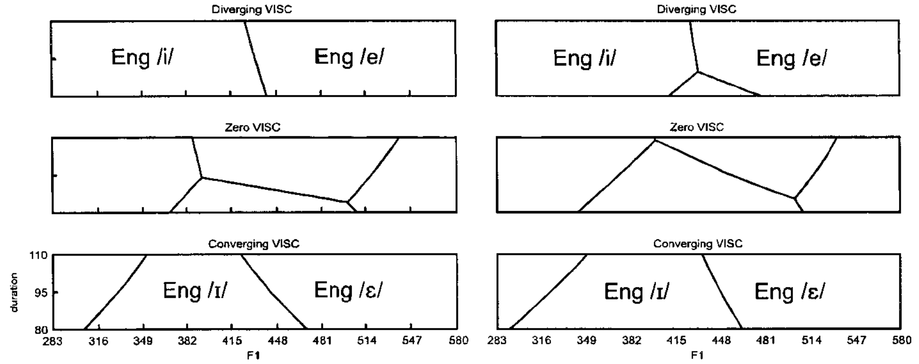
For the bilingual participants, the same stimuli were presented in an English carrier sentence on one occasion and in a Spanish carrier sentence on the other occasion. The carrier sentences were designed to elicit L1- and L2-mode perception, but differences in acoustic properties in the carrier sentences could lead to contrast effects on the stimuli. To test for such differences, four monolingual English participants completed a control experiment in which the carrier sentence was in Spanish (but the response options were in English as in the original English perception experiment). Also, four monolingual English participants completed a control experiment in which there was no carrier sentence.

The bilingual speaker who spoke the original carrier sentences was not a first dialect speaker of Canadian English; therefore, five monolingual English participants completed a control experiment in which the carrier sentence was read by a monolingual English speaker who had grown up in Edmonton, and in which the voice quality characteristics of the synthetic stimuli were matched to this speaker.

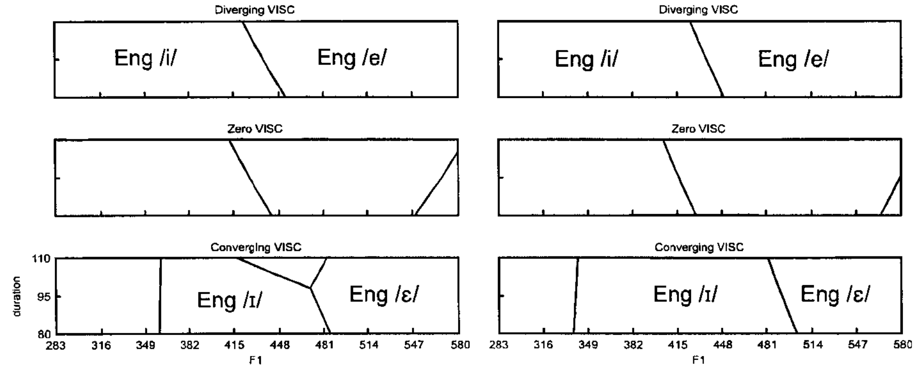
Figures A8.1–A8.4 provide territorial maps based on logistic regression models of each individual's vowel identification data in the original experiments (left) and control experiments (right).

Figure A8.1 Territorial maps from logistic regression models based on monolingual English listeners' original experiment (left) and replication experiment (right).

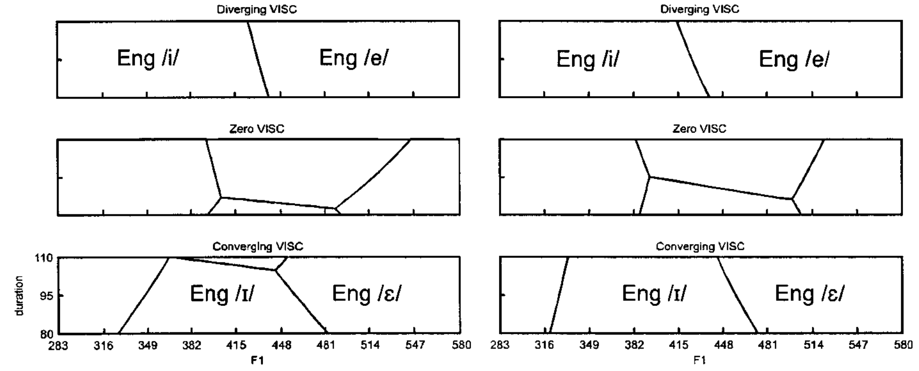
me099



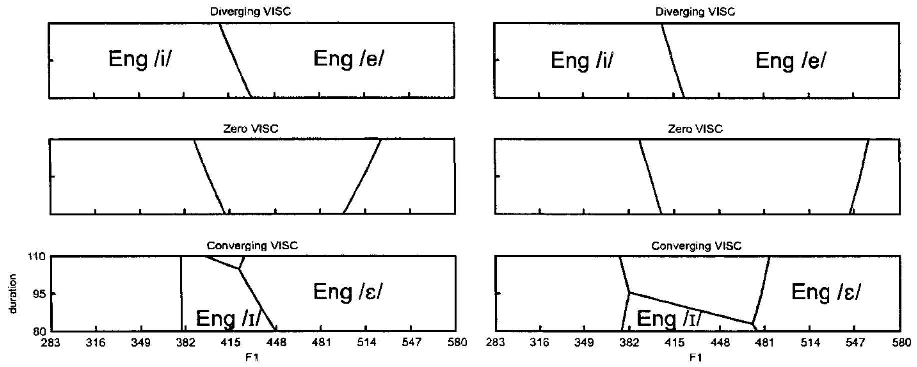
me103



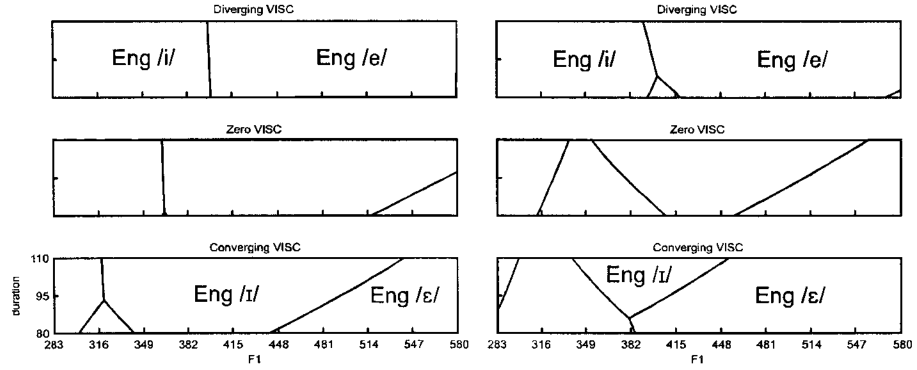
me107



me111

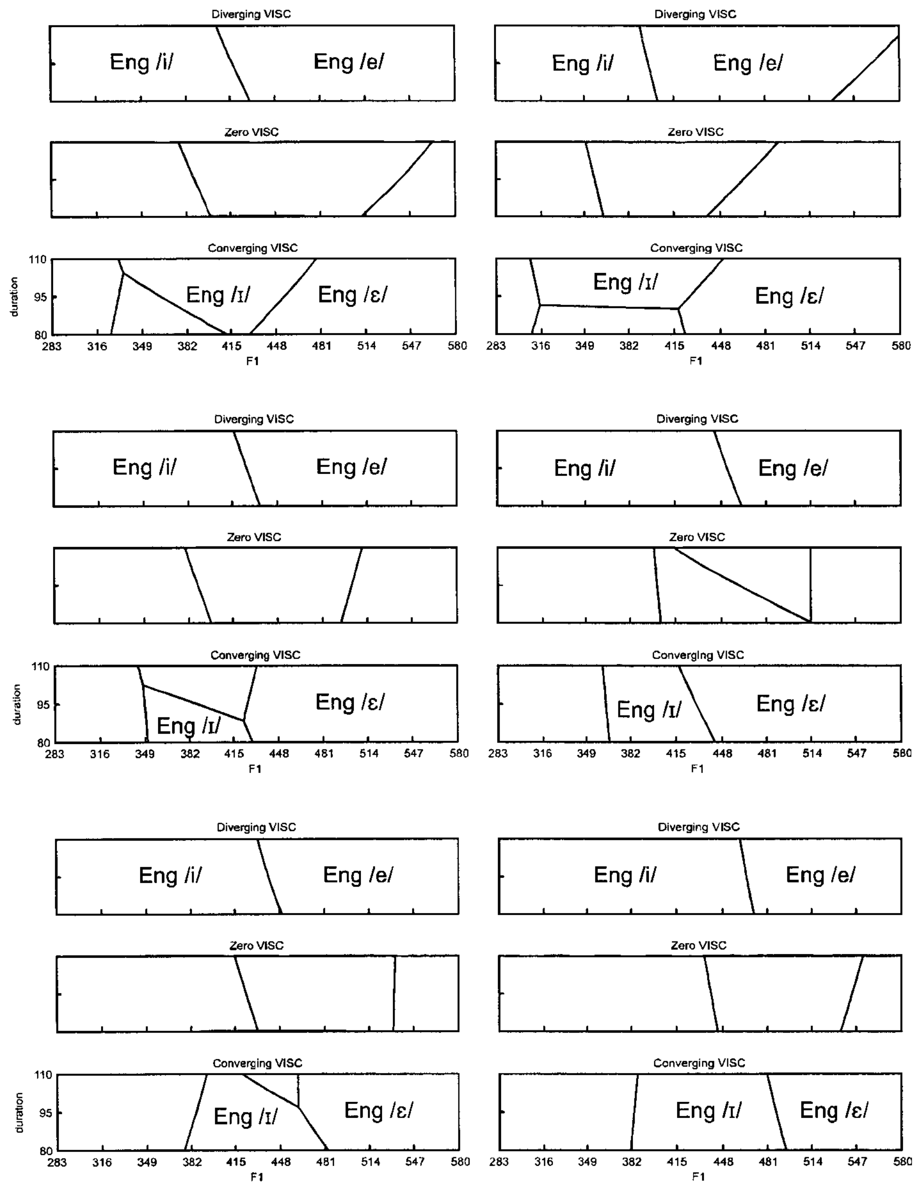


me115



Between the original and the replication conditions, results for four of the five listeners were almost identical, except for an approximately 20 Hz shift in the /i/–/ɪ/ boundary towards lower F1 and an approximately 20 Hz shift in the /ɪ/–/ɛ/ boundary towards higher F1 for me103, and an approximately 25 Hz shift in the /ɪ/–/ɛ/ boundary towards higher F1 and a similar shift in the /e/–/ɛ/ boundary and increase in /e/ responses in the converging-VISC subspace for me111. Participant me115 had a substantial increase in /ɪ/ responses between the original and replication experiment.

Figure A8.2 Territorial maps from logistic regression models based on monolingual English listeners' original experiment (left) and Spanish-carrier-sentence experiment (right).

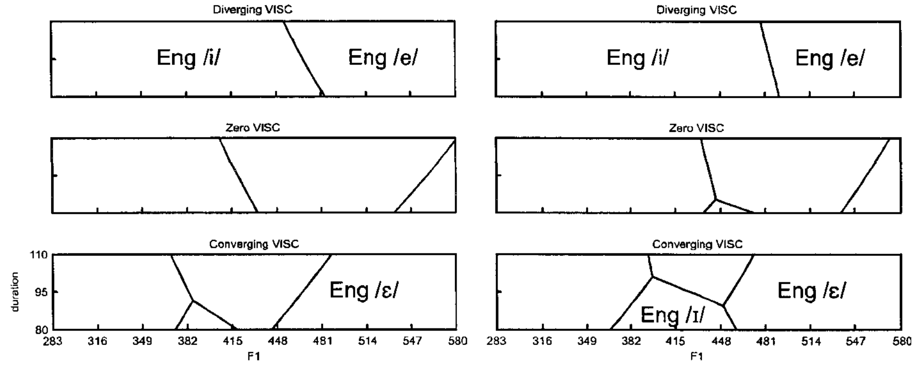


me096

me100

me108

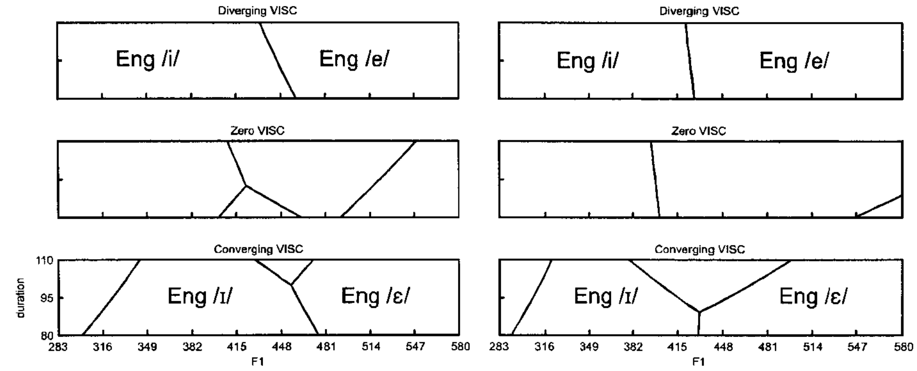
me112



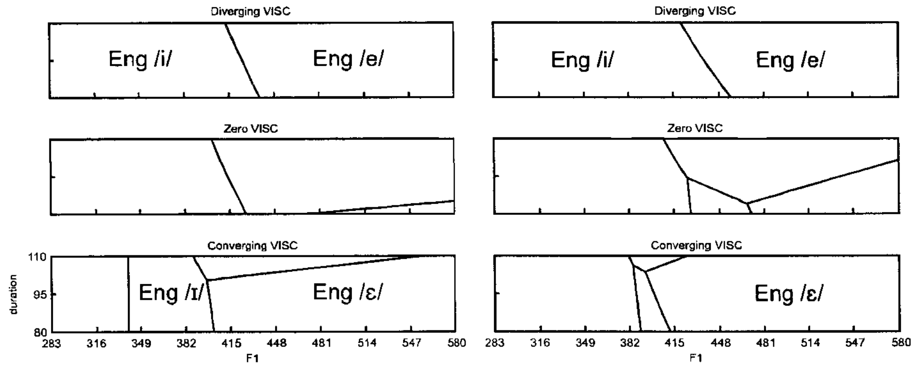
Between the original and Spanish-carrier-sentence conditions there were shifts towards longer durations for the /ɪ/–/e/ boundaries for most participants. This fairly consistent result may be due to contrast effects with the carrier sentences, possibly due to rhythmic differences between English and Spanish. The largest shifts in other boundaries were an increase in F1 of approximately 25 Hz for the /i/–/e/ boundary and approximately 15 Hz for the /i/–/ɪ/ boundary for me100.

Figure A8.3 Territorial maps from logistic regression models based on monolingual English listeners' original experiment (left) and no-carrier-sentence experiment (right).

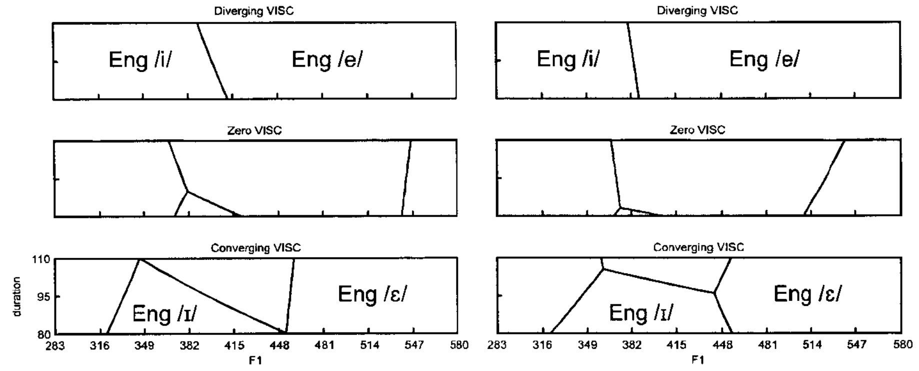
me098



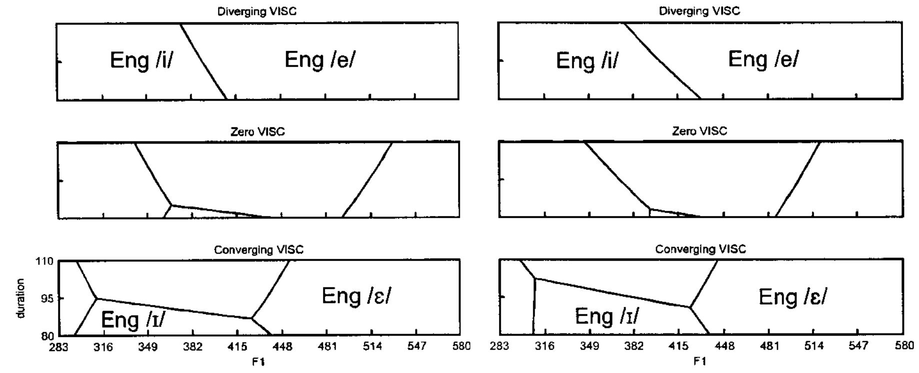
me106



me117

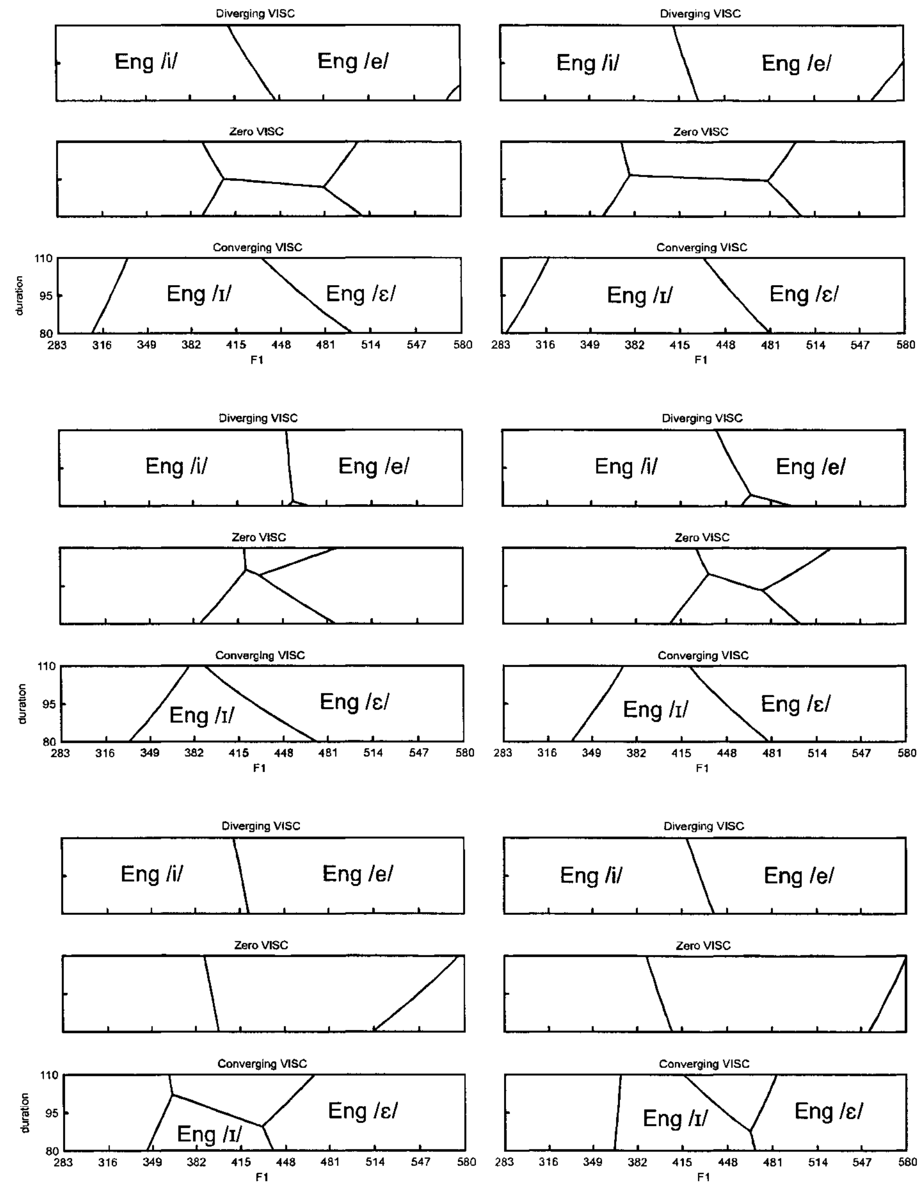


me118



Between the original and no-carrier-sentence conditions most participants had very similar results, although there was a substantial reduction in /ɪ/ responses for me106 with an approximately 45 Hz shift towards higher F1 for the /i/–/ɪ/ boundary. In general, the results indicate that the existence versus the absence of the English carrier sentence had little effect on the monolingual English listeners' perception.

Figure A8.4 Territorial maps from logistic regression models based on monolingual English listeners' original experiment (left) and Edmonton-voice experiment (right).

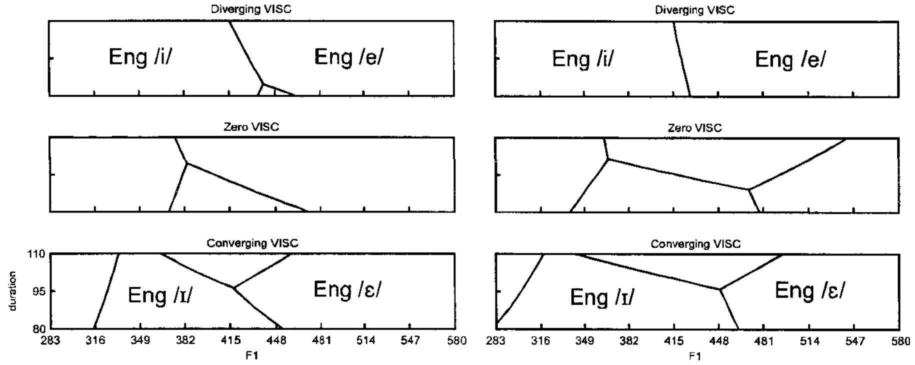


me097

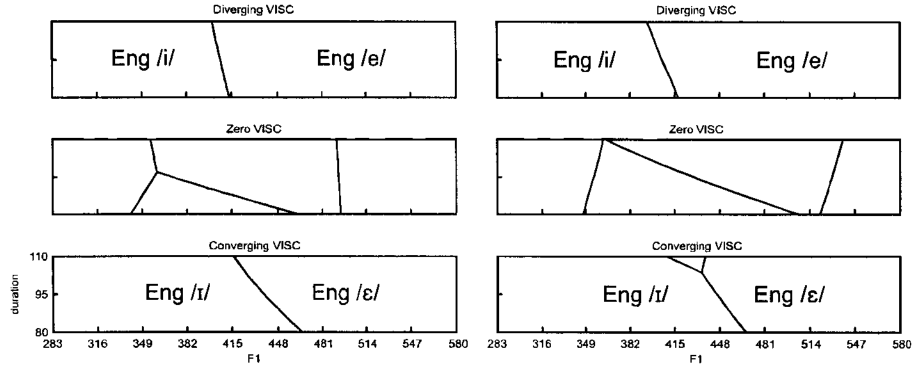
me101

me105

me109



me113



Between the original and no-carrier-sentence conditions most participants had very similar results, indicating that differences in the speakers' voice qualities had little effect on the monolingual English listeners' perception.

Given the results of these comparisons between original and control-condition experiments, differences in bilingual listeners' boundaries along the F1 dimension between pairs of English and pairs of Spanish vowel categories will not be considered to be due to L1 versus L2 language-mode perception unless they exceed approximately 25 Hz.