PERCEPTION OF SYNTHETIC VOWELS BY MONOLINGUAL CANADIAN-ENGLISH, MEXICAN-SPANISH, AND PENINSULAR-SPANISH LISTENERS

Geoffrey Stewart Morrison
School of Language Studies, Australian National University, Canberra, ACT 0200, Australia
e-mail: geoff.morrison@anu.edu.au

ABSTRACT
Monolingual-Western-Canadian-English listeners, monolingual-Mexican-Spanish listeners, and monolingual-Peninsular-Spanish listeners classified stimuli from a synthetic vowel continuum which allowed for English /i/, /I/, /s/, /E/, and Spanish /i/, /s/, and /s/ responses. The continuum varied systematically in initial formant values, vowel inherent spectral change, and vowel duration. The portion of the stimulus space for which the English listeners’ modal response was English /i/ was identified as Spanish /i/ by both groups of Spanish listeners. Three quarters of the portion of the stimulus space for which the English listeners’ modal response was English /I/ was identified as Spanish /i/ and one-quarter as Spanish /s/ by Mexican-Spanish listeners, but almost all of this portion of the stimulus space was identified as Spanish /s/ by Peninsular-Spanish listeners. Spanish dialect may therefore have a substantial effect on first-language-Spanish listeners’ learning of the Western-Canadian-English /i/-/I/ contrast.

RÉSUMÉ

1. INTRODUCTION
Spanish speaking learners of English often have problems with the English /i/-/I/ contrast. Álvarez González (1980, ch. 5), Escudero (2005, §1.2.2), Flege (1991), and Moller Glasbrenner (2005) have reported that:

1. First-language Spanish second-language English listeners (L1-Spanish L2-English listeners) misidentify L1-English speakers’ productions of English /i/ as English /I/ and vice versa.

2. Monolingual-Spanish listeners assimilate the majority of tokens of English /i/ to the Spanish /i/ category.

3. Monolingual-Spanish listeners assimilate the majority of tokens of English /I/ to the Spanish /s/ category.

4. However, monolingual-Spanish listeners assimilate some tokens of English /I/ to Spanish /s/, and identify some tokens of English /I/ as English /E/.

These results were obtained for Peninsular- and American-Spanish speakers listening to English from South-Eastern England, and for American-Spanish speakers listening to English from the United States; however, there is evidence that the choice of English dialect can affect the extent to which tokens of English /I/ are assimilated to the Spanish /i/ category versus the Spanish /s/ category. Escudero & Boersma (2004) examined Peninsular- and American-Spanish listeners’ perception of two dialects of English: Compared to a dialect from the South-East of England, Scottish English has a larger spectral separation and smaller duration separation between /i/ and /I/. Thus L1-Spanish learners of Scottish English were expected to assimilate tokens of English /i/ and /I/ via a two-category assimilation to the Spanish /i/ and /s/ categories respectively, and to have little difficulty perceiving the difference...
between the two English categories. In contrast learners of
the dialect from South-Eastern England were expected to
assimilate tokens of English /i/ and /I/ via a single-category
or category-goodness-difference assimilation to the Spanish
/i/, and to have moderate to considerable difficulty
perceiving the difference between the two English
categories (see Best’s, 1995, Perceptual Assimilation
Model). The assimilation predictions were confirmed for
Peruvian-Spanish listeners (Escudero, 2005, §1.2.2).

There are clearly large differences in vowel pronunciation
across English dialects, but Spanish dialects appear to
be much more homogeneous in terms of vowel pronunciation
(Morrison & Escudero, 2007, failed to find significant
formant differences between the vowel systems of Spanish
speakers from Madrid and Lima). The present study
investigates whether there are differences in vowel perception
between monolingual speakers of two Spanish
dialects, Mexican Spanish (Mexico City) and Peninsular
Spanish (North-Central Spain). Specifically it investigates
whether there are perception differences between dialects
which could affect learning of the Western-Canadian-English
/i-/I/ contrast. Monolingual-Western-Canadian-English
listeners, monolingual-Mexican-Spanish listeners,
and monolingual-Peninsular-Spanish listeners were tested
on their perception of a set of synthetic vowels which
covered an acoustic space which allowed for the perception
of English /i/, /I/, /s/, /E/, and Spanish /i/, /s/, and /si/.

The synthetic stimuli in the present study included
vowel inherent spectral change (VISC), which has been found
to be an important factor in L1-English listeners’
vowel perception in Western-Canadian English, as well as
other dialects of North-American English (Andruski &
Nearey, 1992; Assmann & Katz, 2005; Assmann, Nearey,
& Hogan, 1982; Hillenbrand, Clark, & Nearey, 2001; Nearey
& Assmann, 1986). This contrasts with earlier synthe-
sic-speech studies and edited-natural-speech studies
(Escudero & Boersma, 2004; Flege, Bohn, & Jang, 1997;
Morrison, 2002, 2008), in which formant frequencies were
fixed over the timecourse of the vowel.

Note that Western-Canadian English /s/ is produced with
diverging VISC (F1 decreases and F2 increases over the
timecourse of the vowel), /I/ and /E/ are produced with
converging VISC (F1 increases and F2 decreases over the
timecourse of the vowel), and /u/ is produced with negligible
formant movement (Andruski & Nearey, 1992; Morrison,
2006b, §3.1; Nearey & Assmann, 1986). In Spanish, /s/ is
produced with diverging VISC, and /i/ and /s/ are produced with
negligible formant movement (Morrison, 2006b, §3.1).

2. METHODOLOGY

2.1 Listeners

Nineteen monolingual-Western-Canadian-English
speakers (eight men and eleven women) were recruited in
Edmonton, Alberta, Canada (one was from Saskatchewan
and all the others from Alberta). None reported an ability to
speak any language other than English. They ranged in age
from 18 to 54 with a median of 20.

Twenty monolingual-Mexican-Spanish speakers (ten
men and ten women) were recruited in Mexico City, Federal
District, Mexico. They were all speakers of Mexico-City
Spanish. Thirteen reported a limited ability to speak English
or French, but reported being unable to participate in a
conversation in these languages. They ranged in age from
18 to 31 with a median of 22.

Seventeen monolingual-Peninsular-Spanish speakers
(eight men and nine women) were recruited in Vitoria-
Gasteiz, Autonomous Region of the Basque Country,
Spain. They were speakers of North-Central Peninsular
Spanish (thirteen were from the Basque Country, and one
each from Navarre, Burgos, Leon, and Madrid). Seven
reported a limited ability to speak one or more of Basque,
French, and English, but reported being unable to participate
in a conversation in any of these languages. They ranged in
age from 25 to 53 with a median of 44.

2.2 Stimuli

A version of the Klatt synthesiser (Klatt & Klatt, 1999)
was used to create synthetic /βVñ/ stimuli, and the results were
inserted in to the natural Spanish and English carrier sentences “La próxima palabra es __pa” and “The next
word is __pa” (both sentences have the same meaning). The
final /ñ/ used in the English carrier sentence was actually
taken from the Spanish carrier sentence. In pilot tests the
unstressed utterance final Spanish /ñ/ was acceptable to
L1-English listeners, i.e., it was not perceived as
non-English like. In English-listening mode, the author
would transcribe the sound as English schwa; its mean F1,
F2, and F3 values were 696, 1357, and 2376 Hz. The natural
portions of the stimuli were produced by a male bilingual
speaker (the author).1 Care was taken to adjust synthe-
sis-parameter settings so as to produce synthetic speech
which (in the opinion of the author) was a good match for
the voice quality of the Spanish natural speech. The
speaker’s Spanish productions had a greater spectral tilt than
his English productions, and the spectral tilt of the English
carrier sentence was therefore increased so as to match the
voice quality of the Spanish-based synthetic speech.

A large stimulus space (1464 stimuli) was initially
constructed, and pilot studies were conducted in order to
find a smaller set of stimuli which included stimuli which
were acceptable as Spanish /ñ/, /ñ/, and /ñ/ to L1-Spanish
listeners, and stimuli which were acceptable as English /ñ/,
/ñ/, and /ñ/ to L1-English listeners. Figure 1 provides a
plot of the smaller stimulus set. The 90 stimuli selected had
ten sets of initial formant values along a diagonal in the
F1–F2 vowel space ranging from [F1, F2] of [283 Hz, 2090
Hz] to [580 Hz, 1730 Hz], in equal steps of [+33 Hz, −40
Hz]. At each start-point, stimuli were synthesised with three
levels of VISC: F1 and F2 either diverged, did not change
(we flat), or converged over the time-course of the vowel.
Formant movements [ΔF1, ΔF2] from the beginning to the

1
2.3 Procedures

Listeners were tested one at a time using custom-written software. Monolingual-Western-Canadian-English listeners were tested in the Centre for Comparative Psycholinguistics at the University of Alberta, Monolingual-Mexican-Spanish listeners were tested in the Phonics Studies Laboratory at El Colegio de México in Mexico City, and Monolingual-Peninsular-Spanish listeners were tested in the Phonetics Laboratory at the University of the Basque Country. In Spain and Canada testing took place in a sound booth using a Roland ED UA-30 USB Audio Interface and Sennheiser HMD 280 PRO headphones. In Mexico testing took place in the quietest room available using an Edirol UA-25 Audio Interface and AKG K701 headphones.

Listeners heard a stimulus sentence, and responded by clicking on the response button which corresponded to their identification of the synthetic vowel. A new stimulus was presented 500 ms after a response was given. In the Spanish experiment the response buttons were labelled BIPA, BEIPA, and BEPA representing /βipa/, /βi p a/, and /β ip a/ respectively, and in the English experiment the response buttons were labelled BEIPA, BIPPA, BAYPA, and BEPPA representing /βipa/, /βi p a/, /β ip a/, and /βip a/ respectively. The spelling-to-phoneme relationship is transparent in Spanish, but less clear in English. Prior to the English experiment, listeners were therefore trained on the English spelling-to-phoneme relationship. Listeners saw written sets of real words illustrating the four English vowel categories, and each set was followed by the corresponding response word. Listeners read the real and response words out loud, and the researcher monitored to ensure that they pronounced the same vowel sounds in the response words as in the real words. Any mismatches between the real and response words were corrected by asking the listeners to read the response word with the same vowel as in the appropriate set of real words. The researcher pointed at the written forms of the words but did not pronounce the words or model the vowels in isolation. Training was restricted to making sure that participants produced the same vowel sound in real and response words. The training continued until the researcher was confident that the listeners understood the spelling-to-phoneme relationships. The written sets of real and response words were also visible to the listeners during the experiment.

All 90 stimuli were presented in random order in two blocks, and in each of four subsequent randomised blocks an adaptive procedure selected 45 stimuli for presentation. In each of the last four blocks, category boundaries were estimated on the basis of the responses given in the earlier blocks, and stimuli in the vicinity of the category boundaries had the highest probability of being selected for presentation in the new block. This resulted in a total of 360 trials per listener, with each stimulus identified a minimum of twice and a maximum of six times. The procedure is described in detail in Morrison (2006a). It produces results which do not differ substantially from results obtained using six responses on each stimulus (540 trials), but within a time period which end of the vowel were [−99 Hz, +120 Hz], [0 Hz, 0 Hz], and [+99 Hz, −120 Hz] (minus three, zero, and plus three steps along the F1–F2 diagonal). Following Andruski & Nearcy (1992), the formant trajectories described straight lines in a log-hertz F1–F2–F3 space. Following Nearcy (1989), third-formant (F3) values were set using a formula based on a linear regression of F1 and F2 values onto F3 values from the model speaker’s vowel productions. Equation 1 provides the formula with F1, F2, and F3 values given in hertz.

\[
F3 = 4235 -2.427 \times F1 -0.272 \times F2.
\] (1)

Each of the 30 initial- and final-formant combinations was synthesised at three durations: 55, 70, and 85 ms (for each set of stimuli with the same initial- and final-target combinations, the shorter stimuli had steeper slopes than the longer stimuli). The synthetic stimuli also included a bilabial burst, bilabial onset and offset formant transitions, and a 90 ms long /n/ closure. The consonant transitions added an additional 25 ms to each vowel, resulting in total vowel durations of 80, 95, and 110 ms.

Audio recordings of the carrier sentences and stimuli are available in Morrison (2006b, p. 32).
does not lead to listener fatigue.

3. RESULTS & DISCUSSION

3.1 Statistical Modelling Procedures

Perception results were analysed using logistic regression. For an explanation of the type of logistic regression modelling applied here it is highly recommended that the reader refer to Morrison (2007a).

The logistic regression models estimated a set of coefficient values associated with each response category:

bias coefficients:
\[ \alpha_{0p}, \alpha_{0d}, \alpha_{0f}, \alpha_{0c} \]

initial-formant-tuned coefficients:
\[ \beta_{0/${\text{initial}}F_1}, \beta_{0/${\text{initial}}F_2}, \beta_{0/${\text{initial}}F_3}, \beta_{0/${\text{initial}}F_4} \]

duration-tuned coefficients:
\[ \beta_{0/${\text{dur}}_1}, \beta_{0/${\text{dur}}_2}, \beta_{0/${\text{dur}}_3}, \beta_{0/${\text{dur}}_4} \]

diverging-VISC-tuned coefficients:
\[ \beta_{0/${\text{div}}_1}, \beta_{0/${\text{div}}_2}, \beta_{0/${\text{div}}_3}, \beta_{0/${\text{div}}_4} \]

converging-VISC-tuned coefficients:
\[ \beta_{0/${\text{conv}}_1}, \beta_{0/${\text{conv}}_2}, \beta_{0/${\text{conv}}_3}, \beta_{0/${\text{conv}}_4} \]

Initial formant values and duration values were entered as continuous variables in just-noticeable-difference (JND) units. The JND scale for initial formant values was one-dimensional (F1 and F2 were 100% correlated in the synthetic stimuli) with its origin corresponding to the stimuli with the lowest F1 and highest F2 [283 Hz, 2090 Hz]. The JND used was 0.3 Bark (Kewley-Port, 2001). The conversion from hertz to the JND-formant scale (FJND) was performed using Equation 2 (which includes the hertz-to-bark formula from Traunmüller, 1990):

\[
F_{\text{JND}} = \frac{\sqrt{(\text{Bark}(F1) - \text{Bark}(283))^2 + (\text{Bark}(F2) - \text{Bark}(2090))^2}}{0.3} \times \frac{26.81F}{1960 + F} - 0.53
\]

The origin of the JND scale for duration corresponded to the stimuli with the shortest duration (80 ms), and the JND used was 5 ms on a base value of 90 ms (Noteboom and Doedeman, 1980, similar to the Weber fraction of 0.05 used by Smits, Sereno, and Jongman, 2006). The conversion from milliseconds to the JND-duration scale (durJND) was performed using Equation 3:

\[
dur_{\text{JND}} = \log_{1+0.90}(\text{dur}/90) - \log_{1+0.90}(80)/90
\]

Use of JND-scales allows initial-formant and duration results to be compared on an equal footing.

VISC was entered as three discrete levels, resulting in two dummy-coding coefficients \[[\beta_{0p}, \beta_{0c}] : [0.0] = \text{zero VISC}, [1.0] = \text{diverging VISC}, [0.1] = \text{converging VISC}\]. This encodes the onset + offset (or the onset + direction) hypothesis for the perceptually relevant aspects of VISC (Gottfried, Miller, & Meyer, 1993; Nearey & Assmann, 1986; Morrison, 2007b; Morrison & Nearey, 2007; Pols, 1977).

3.2 Statistical Modelling Results

Figures 2 through 4 provide population-average territorial maps and probability-surface plots based on logistic regression models fitted to monolingual-Western-Canadian-English listeners’ response data, monolingual-Mexican-Spanish listeners’ response data, and monolingual-Peninsular-Spanish listeners’ response data. Territorial maps indicate which category is the model’s predicted modal response in each part of the stimulus space (see Nearey, 1990, 1997). Probability-surface plots indicate the model’s predicted probability for each response category in each part of the stimulus space (see Morrison, 2007a, 2008). (Each category is shaded a different colour, the same colours are used in the territorial maps and probability surface plots). The population-average territorial maps and probability-surface plots were created by fitting a logistic regression model to each individual listener’s response data, then taking the mean of the logistic regression coefficient estimates across all listeners within each group. These mean coefficient values were then used to calculate the model’s predicted probability for each category response at each point in a fine grid of points covering the stimulus space.

Examination of Figure 2 indicates that English /ə/ is the modal response in approximately half the diverging-VISC portion of the stimulus space, consistent with its traditional description as a (diverging) phonetic diphthong. Western-Canadian-English /i/ and /E/ are produced with converging VISC, and consistent with this, English /I/ and /E/ were the modal responses in most of the converging-VISC portion of the stimulus space. Western Canadian English /I/ is produced as a monophthong, and consistent with this, English /I/ was the modal response in the low-F1 part of the zero-VISC portion of the stimulus space. Some parts of the stimulus space, e.g., low-F1 converging-VISC, do not correspond to the production values of any English vowel categories, but listeners extrapolated the neighbouring categories and gave responses in these areas. Note that the orientation of the boundary between the modal areas for /A/ and /I/ responses indicates that Western-Canadian-English listeners used a mixture of initial formant values, VISC, and duration to distinguish these two categories.

Examination of Figures 3 and 4 indicates that Spanish /ε/ is the modal response over about half of the diverging-VISC portion of the stimulus space. This is as expected given that Spanish /ε/ is a diverging diphthong. The zero-VISC stimulus space is divided between the two Spanish monophthongs /a/ and /e/. This is as expected
assuming that these two vowels are monophthongs. Monolingual-Spanish listeners also had Spanish /i/ and /e/ as the modal response in the converging-VISC portion of the stimulus space. Note that Spanish does not have any vowels with acoustic properties similar to those in the converging-VISC portion of the stimulus space, but the results indicate that the monolingual-Spanish listeners perceived these stimuli as more similar to their Spanish /i/ and /e/ categories than to their Spanish /ei/ category. The boundaries between /i/ and /e/ response categories were relatively close to parallel to the duration axis, suggesting that duration played little part in the monolingual-Spanish listeners’ perception of the contrast between these two vowels.

There were differences between Mexican- and Peninsular-Spanish listeners perception of the stimuli: The boundaries between Spanish /i/ and /e/ have noticeably higher F1 values for Mexican listeners (Figure 3) compared to Peninsular listeners (Figure 4).

### 3.3 Initial L2-perception predictions based on monolingual perception

Comparing the monolingual-Spanish and monolingual-English models, predictions can be made as to how L1-Spanish speakers just beginning to learn English would perceive the synthetic stimuli in terms of Spanish categories.

The area identified by English /I/ as the modal response in the monolingual-English listeners’ territorial map (Figure 2a), corresponded almost exclusively to areas which had Spanish /i/ as the modal response in the territorial maps of both groups of monolingual-Spanish listeners (Figures 3a and 4a).

Figures 3a and 4a include an overlay of the English /I/ modal response area from Figure 2a. Approximately two-thirds of the English /I/ modal response area corresponded to the Spanish /i/ modal response area in the monolingual-Mexican listeners’ territorial map (Figure 3a). This suggests that L1-Mexican-Spanish listeners will assimilate tokens of Western-Canadian-English /i/ and /I/ primarily via a category-goodness-difference assimilation to Spanish /i/, and may have difficulty distinguishing the two English vowels. The predictions for L1-Mexican-Spanish learners of English are consistent with the results of earlier studies of L1-Spanish listeners’ perception of Canadian-English /i/ and /I/, which suggested substantial confusion between /i/ and /I/ (Morrison, 2002, 2008).
In contrast, almost all the English /I/ modal response area corresponded to the Spanish /e/ modal response area in the monolingual-Peninsular listeners’ territorial map (Figure 4a), less than one-eighth corresponded to Spanish /a/. This suggests that L1-Peninsular-Spanish listeners will assimilate tokens of Western-Canadian-English /a/ and /I/ primarily via a two-category assimilation to the Spanish /a/ and /e/ categories respectively, and will therefore have little difficulty distinguishing /a/ and /I/.

4. SUMMARY & CONCLUSION

Earlier studies (Escudero, 2005, §1.2.2; Escudero & Boersma, 2004) have shown that L1-Spanish listeners’ perception of the English /a/-/I/ contrast is dependent on English dialect. This is not surprising given that across English dialects there can be substantial differences in the phonetic realisation of vowel phonemes. Compared to English there appears to be relatively little difference in vowel realisation across different dialects of Spanish, and several earlier studies (Escudero & Boersma, 2004; Flege, 1991; Flege et al., 1997; Morrison, 2008) have tacitly assumed that Spanish dialect will not have a major impact on the results of studies of L1-Spanish listeners’ perception of English /a/ and /I/. The present study tested monolingual-Western-Canadian-English, monolingual-Mexican-Spanish, and monolingual-Peninsular-Spanish listeners’ perception of a synthetic vowel continuum which varied systematically in initial formant values, vowel inherent spectral change, and vowel duration. Perception differences were found between Mexican and Peninsular listeners (one would also hypothesise that there are differences in production). In the portion of the stimulus space where Canadian-English listeners’ modal response was English /a/, the modal response for both Mexican- and Peninsular-Spanish listeners was Spanish /e/. In the portion of the stimulus space where Canadian-English listeners’ modal response was English /I/, the responses for the Mexican-Spanish listeners were approximately two-thirds Spanish /e/ and one-third Spanish /a/; whereas for the Peninsular-Spanish listeners the responses were almost all Spanish /a/. This leads to the prediction that whereas L1-Mexican-Spanish listeners are likely to perceive most tokens of Western-Canadian-English /a/ and /I/ via a category-goodness-difference assimilation to Spanish /e/, and to have difficulty learning the Western-Canadian-English /a/-/I/ contrast, L1-Peninsular-Spanish listeners are likely to perceive most tokens of Western-Canadian-Spanish /a/ and /I/ via a two-category assimilation to Spanish /a/ and /e/, and to have little difficulty learning the Western-Canadian-English /a/-/I/ contrast. L1-Spanish dialect may therefore have a substantial effect on L1-Spanish listeners’ ability to learn the English /a/-/I/ contrast.

ACKNOWLEDGEMENTS

This research was supported by the Social Sciences and Humanities Research Council of Canada (SSHRC). Collection and analysis of the Canadian and Peninsular data took place while the author was a SSHRC Doctoral Fellow at the Department of Linguistics, University of Alberta, and collection and analysis of the Mexican data took place while the author was a SSHRC Postdoctoral Fellow at the Department of Cognitive & Neural Systems, Boston University. The final version of this paper was completed at the School of Language Studies, Australian National University. My thanks to all the volunteers who provided data, and to those who assisted with recruiting and facilities in Edmonton, Mexico City, and Vitoria-Gasteiz. Thanks to my former PhD supervisor Terry Nearey for advice throughout the project, to two anonymous reviewers for comments on an earlier version of this paper, and to Takeki Kamiyama and Marie-Claude Tremblay for translating the abstract.

NOTE

1. The speaker’s first language was English. Although originally from the UK he had lived in Canada for over ten years. In Morrison (2006b, appendix 8) a control experiment was conducted in which a subset of the L1-English listeners also identified stimuli in a carrier sentence produced by a speaker from Edmonton with the synthetic stimulus voice properties matched to that speaker. There were no substantial differences between the listeners’ perception of the stimuli. The Speakers’ second language was Spanish. He began learning Spanish at age 13, had studied Spanish for many years, had visited Spain many times, had passed the Diploma Superior de Español como Lengua Extranjera [Advanced Diploma in Spanish as a Foreign Language], and had lived in Spain for a year. Even after prolonged conversations, Mexicans assumed the was Spanish. The Spanish carrier sentence did not contain any vocabulary or phonemes which would immediately mark the differences between Mexican and Peninsular Spanish.

REFERENCES


