Japanese Listeners’ Use of Duration Cues in the Identification of English High Front Vowels

GEOFFREY STEWART MORRISON
Simon Fraser University

0. Introduction
General Canadian English (along with several other English dialects) has two high front vowel phonemes differing in both spectral and duration properties. In identical contexts, the tense /i/ is both more peripheral in the vowel space and longer than the lax /ɪ/ (Nearey & Assmann 1985, Morrison 2002b). Japanese also has two high front vowel phonemes: short /i/ and long /iː/. Japanese long and short vowels are traditionally described as differing only in terms of duration with identical spectral properties (Akamatsu 1997).

Ingram & Park (1997) found that Japanese listeners identified Australian English /i/ and /ɪ/ at near perfect levels (an error rate of only 1%). In an L1 assimilation test, Japanese listeners classified English /i/ as Japanese /iː/ at a rate of 100% and classified English /ɪ/ as Japanese /i/ at a rate of 90% (10% as /iː/). The stimuli had been produced by two native English speakers, one of whom spoke faster than the other resulting in different vowel durations: one speaker’s /i/ and /ɪ/ vowels were approximately 75 and 255ms long respectively, and the other speaker’s vowels were 165 and 390ms. English vowels were consistently classified as long or short Japanese categories irrespective of the inter-speaker duration differences. Ingram & Park hypothesised that the Japanese listeners were able to normalise for speaking rate differences because of their experience with the phonemic vowel duration contrast in Japanese. Korean listeners tested on the same stimuli were not able to normalise for speaking rate differences (modern Korean lacks a vowel duration contrast). The implication is that the Japanese listeners identified English /i/ and /ɪ/ using the same duration criteria that they used to identify Japanese /i/ and /iː/, and normalised for speaking rate differences as they would when identifying Japanese long and short vowels. Presumably the Japanese participants also made some use of spectral cues, since English /ɪ/ was more sometimes correctly identified as /i/.

1This paper is based on part of my forthcoming MA thesis. My thanks to Dr. Murray J. Munro for comments on an earlier version of the paper, and to the anonymous participants and model speakers.
The results of several studies, however, cast doubt on the ability of Japanese listeners to normalise for speaking rate. Strange et al. (1998) found that in US English /h_\_ba/ words, long English vowels /i e æ a o u/ and short English vowels /ɪ ɛ æ o u/ were assimilated to Japanese long and short vowels at rates of 39% and 92% respectively. When the final /a/ in the /h_\_ba/ words was truncated to the same duration as /A/ in order to simulate a faster speaking rate, there was little difference in the duration assimilation pattern: Long and short English vowels were assimilated to Japanese long and short vowels at rates of 43% and 94% respectively. Likewise, Guion et al. (2002) found no significant difference in the assimilation of English vowels to Japanese long and short categories in isolated /b_\_bo/ words in which /bo/ had two durations. Toda (1999) measure native Japanese listeners’ long-short boundary in the final vowel of (C)VCC - (C)VCVV continua. Altering the duration of the first vowel so that it was 100% longer in one series of continua than the other resulted in only a 10% increase in the location of the boundary. Given the small degree of normalisation for simulated speaking rate in these studies, it seems unlikely that normalisation could account for the ability of the Japanese listeners in Ingram & Park (1997) to correctly identify English /i/ and /ɪ/ irrespective of large speaking rate differences. It seems especially unlikely given that the vowels were presented in isolated /h_\_d/ words in which only duration differences in the two consonants would be available as speaking rate cues.

Morrison (2002a) had Japanese listeners identify Scottish / Northern English vowels in isolated CVC words in which both consonants were either voiced or voiceless. The correct-identification rate for English /i/ was 93%, and for English /ɪ/ in the voiced consonant condition it was 95%. In contrast, English /i/ in the voiceless consonant condition had correct-identification rate of only 27%; it was identified as /ɪ/ at a rate of 72%. Given that English vowels are typically longer before voiced than before voiceless consonants (Chen 1970), Morrison (2002a) proposed that: English /i/ is assimilated to Japanese /i:\:/ before a voiced consonant but assimilated to Japanese /ɪ/ before a voiceless consonant; that English /ɪ/ is assimilated to Japanese /i/ before both a voiced and voiceless consonant; that English vowels assimilated to Japanese /i:\:/ are identified as English /i/, and that English vowels assimilated to Japanese /i/ are identified as English /ɪ/. This requires that the Japanese long-short vowel boundary fall between the longest duration for an English /i/ stimulus in the voiced consonant condition or for an English /ɪ/ stimulus in either condition (whichever be the greater), and the shortest duration for an English /i/ stimulus in the voiceless consonant condition. Comparison of Morrison’s (2002b) stimulus durations with Japanese vowel production data from Tsukada (1999) and the English stimulus durations from Ingram & Park (1997) (Figure1) suggested that the Japanese long-short vowel duration boundary does indeed fall within this range.

Under this hypothesis, Morrison (2002a) proposed that the near perfect correct identification rates for English /i/ and /ɪ/ in Ingram & Park (1997) were not due to
Japanese Listeners’ Use of Duration Cues

(Figure 1) Duration ranges for English /i/ before voiced consonants and English /ɪ/ before voiced and voiceless consonants (left), and for English /i/ before voiced consonants (right) from Morrison (2002a). Duration ranges for Japanese /i/ (left) and /iː/ (right) from Tsukada (1999); and for English /i/ (left) and /i/ (right) from Ingram & Park (1997). The dashed line represents a categorical boundary that would account for the identification patterns in Ingram & Park (1997) and Morrison (2002a).

the Japanese listeners normalising for speaking rate. Rather, the results may be due to the English /ɪ/ stimuli in Ingram & Park (1997) being short enough to be assimilated to Japanese /i/, and the English /i/ stimuli being long enough to be assimilated to Japanese /iː/. Although there was a large difference in vowel durations, the duration ranges for vowels within each English category did not straddle the Japanese long-short categorical boundary due to the fact that Ingram & Park had used context in which the vowels occurred before a voiced consonant.

The present paper will test Morrison’s (2002a) hypothesis by determining whether Japanese listeners identify English /i/ and /ɪ/ using the same duration criteria that they use to identify Japanese /iː/ and /i/.

1. Method
   1.1. Participants
Canadian English participants (3 women, 4 men) had lived until age 16, and immediately prior to data collection, in an Anglophone region of Canada west of Quebec. The group had a mean age of 30 (range 25-38).

Japanese participants (4 women, 3 men) were undergraduate students attending an academic exchange in Vancouver. Their programme included content-based language instruction but no courses in English pronunciation. They had lived in Vancouver for a period of less than 2 months prior to data collection. They had lived until age 16 in Japan, and as children, had not been immersed in a language other than Japanese. They had never lived outside Japan for a continuous period of more than three months, and had not lived in an English speaking country during the year prior to their arrival in Canada. The group had a mean age of 20 (range 19-22). They had studied English in school for 7-9 years, starting at age 13 or 14.
1.2. Stimuli
1.2.1 Target segments and carrier sentences
The English target words (“bit, beat, bid, bead” /bɪt bɪt bɪd bɪd/) were chosen to exemplify the English vowels /i/ and /u/ in a stressed position followed by both a voiced and voiceless plosive.

Since responses to English and Japanese stimuli were to be compared, the segment immediately following the English target word had to match the parallel segment in the Japanese stimuli. The segment selected was /s/, resulting in voiceless affricated release for English /t/ and /d/, corresponding to the Japanese affricate [ts]. It was not possible in Japanese to exactly match the English voiced alveolar plosive with voiceless affricated release; in Japanese both parts of the affricate are voiced, i.e. [dz]. The Japanese target words read by the model speaker were therefore “bitsu, biitsu, bidzu, biidzu” exemplifying possible combinations of short and long vowels with voiced and voiceless singleton and geminate consonants. Although phonotactically rare, the words “biddzu, biittsu, biiddzu” were added as possible response categories in the perception test since some of the stimuli had properties that might suggest these responses.

Matching English and Japanese carrier sentences were “What they’re wearing are beat suits,” and “motteiru mono wa bitsuutsu” (‘What I have is a [nonsense word]-suit’). Strange et al. (1998) found that the use of a carrier sentence, as compared to the presentation of target words in isolation, can have a significant impact on perception, such that long English vowels were assimilated to Japanese long vowels more often in a sentence condition than in an isolated word condition (85 versus 42%). They hypothesised that the participants had used the rhythmic pattern of the sentence to calibrate their duration-based responses.

1.2.2 Acoustic properties of stimuli
Perceptual stimuli were constructed using edited natural speech based on the productions of two model speakers, a 34 year-old male monolingual English speaker who had lived most of his life in Vancouver, and a 38 year-old male native speaker of Japanese who had lived most of his life in Fukuoka. Recordings were made in a soundproofed room using a Sony MZS-R5ST Mini Disc recorder and a Sony ECM-MS907 microphone. The speakers read randomised lists of stimulus sentences at normal, slow, and fast speaking rates. The recordings were digitally transferred to computer, were analysed, and were used to create the stimuli using Praat computer software. Details of the procedure are given in Morrison (2002b).

A multidimensional continuum was created in which acoustic properties varied along the following dimensions: vowel spectra (5 points), vowel duration (7 points), plosive closure duration (5 points), and speaking rate (2 points). The dimensions can be visualised as the cuboids shown in Figure 2.

The F1, F2, and F3 values at the midpoints of the vowels in the English stimuli
Japanese Listeners’ Use of Duration Cues

(Figure 2) Cuboids representing the dimensions along which the English stimuli vary. Vowel duration (durv) from front to back: 1 is shortest and 7 longest. Vowel spectral properties (spec) from right to left: 1 is most /i/-like and 5 most /i/-like. Consonant closure duration (durc) from bottom to top: 1 is shortest and 5 longest. Speaking rate: left cube is slow and right cube is fast. The triangles represent the position of typical “bit, beat, bid, bead” produced by the model speaker.

are given in Table 1. The spectral properties of spectral points 1 and 5 were the mean spectral values from the model English speaker’s /i/ and /i/'. The five points were evenly spaced on the mel scale.

The vowel and consonant durations of the stimuli are given in Table 2. The range of durations for the English stimuli was expanded to cover the same range as the Japanese durations. The properties of the duration endpoints were based on the mean values from contexts resulting in the longest and shortest vowels and consonants produced by the Japanese model speaker. The duration points were evenly spaced on a logarithmic scale. Based on mean values from the English and Japanese model speakers, sentence durations for the slow and fast speaking rates were 1.541 and 1.265 seconds respectively (excluding the duration of the target segments).

The Japanese continua had the same duration properties as the English continuum, but had no spectral dimension. The spectral values for the Japanese stimuli (F1, F2, F3 = 318, 2126, 3115Hz) were those of a model vowel which had spectral values close to the mean values produced by the model speaker. Since Japanese phonemic voicing is always cued by phonetic voicing, two continua were created, one with, and one without phonetic voicing. The dimensions can be visualised as the rectangular matrices shown in Figure 3.

Stimulus vowels had durations and spectral properties within 2.5% of the values given in Tables 2 and 3. Consonant closure durations for English consonants and Japanese voiceless consonants consisted of silence.
Geoffrey Stewart Morrison

(Table 1) Spectral values at the midpoints of the vowels in the English stimuli.

<table>
<thead>
<tr>
<th>Endpoint vowel</th>
<th>Dimension point</th>
<th>F1 Hz</th>
<th>F2 mel</th>
<th>F3 Hz</th>
<th>F3 mel</th>
</tr>
</thead>
<tbody>
<tr>
<td>i</td>
<td>1</td>
<td>410</td>
<td>496</td>
<td>1700</td>
<td>1433</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>370</td>
<td>454</td>
<td>1834</td>
<td>1503</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>330</td>
<td>412</td>
<td>1974</td>
<td>1572</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>292</td>
<td>370</td>
<td>2121</td>
<td>1642</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>255</td>
<td>328</td>
<td>2275</td>
<td>1712</td>
</tr>
</tbody>
</table>

(Table 2) Durations of the vowels and consonants in the English and Japanese stimuli.

<table>
<thead>
<tr>
<th>Dimension point</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vowel duration (ms)</td>
<td>60</td>
<td>72</td>
<td>86</td>
<td>104</td>
<td>124</td>
<td>149</td>
<td>179</td>
</tr>
<tr>
<td>Consonant closure duration (ms)</td>
<td>40</td>
<td>58</td>
<td>84</td>
<td>122</td>
<td>177</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(Figure 3) Rectangular matrices representing the dimensions along which the Japanese stimuli vary. Vowel duration (durv) from left to right: 1 is shortest and 7 longest. Consonant closure duration (durc) from bottom to top: 1 is shortest and 5 longest. Speaking rate: left rectangles are slow right rectangles fast. Consonant voicing: top rectangles are voiceless and bottom rectangles voiced. The triangles represent the position of typical “bittsu, bitsu, biitsu, bidzu, biidzu” produced by the model speaker.
1.3 Data Collection
Participants were tested individually in a soundproofed room. They listened to the stimuli presented in random order via MEDS computer software over Optimus HP340 headphones, and responded to each English stimulus by clicking on one of five pictures on the computer screen. Four of the pictures represented the words “bit, beat, bid, bead.” The fifth picture, an “X,” was a null response which the participants were instructed to use if they heard a word other than one of the four target words. Pictures were used in order to avoid potential confusion from orthographic representations of the target words. Participants were trained to interpret the pictures before the perception test. The computer played a stimulus once and did not proceed until the participant had given a response. The inter-trial interval was 500ms. The order of the pictures was assigned randomly for each trial. A subset of 24 stimuli was used as a warm-up before the full set of 350 stimuli, each of which was identified once.

The Japanese perception test was conducted approximately one month after the English test using the same procedures as the English test, except that the eight response options “bittsu, biitsu, bitsu, biitsu, bidduz, bidzu, biddu, biiddzu” were written in katakana (a Japanese orthographic system). The full set of 140 stimuli were presented in two randomisations.

2. Results and Discussion
Categorical boundaries in the listeners’ responses were modelled using discriminant analyses. Vowel spectral properties, vowel duration, consonant duration, and speaking rate were candidate variables entered stepwise into the model if they produced a between-group $F$ ratio significance of .05 or less. Discriminant function coefficients and group centroid values were used to calculate 0.1, 0.5, and 0.9 probability of /u/ identification points which were then graphed in terms of the original variables. The 0.5 probability points indicate the position of the categorical boundary. A small distance between the 0.1 and 0.9 probability points indicates a sharp boundary, and a large distance indicates a fuzzy boundary. Due to space limitations, only probability point values, variable weightings expressed as percentages, and cross-validated correct classification rates are reported here. Details of the analysis procedure and full statistical results are given in Morrison (2002a).

2.1. Canadian English Listeners
The discriminant analysis was highly successful at categorising the English listeners’ responses (the cross-validated correct classification rate was 95.5%). All variables except speaking rate were entered into the model. This suggests that speaking rate did not affect the English listeners’ vowel perception, and that they identified the vowels primarily according to their spectral properties. The relative weighting given to the variables were 84% for vowel spectral properties, 11% for vowel duration, and
5% for consonant duration. The small effect for vowel duration is predictable since tense /i/ is longer than lax /I/ all else being equal. The slight effect for consonant closure duration is also predictable since a long consonant closure will create the perceptual contrast effect of a shorter vowel (Kluender, Diehl, & Wright, 1988).

The categorical boundary is graphed in Figure 4. The planes are almost perpendicular to the spec axis, indicating that the native English participants based their vowel identification primarily on spectral properties. The mean spectral value of the boundary between /i/ and /I/, was 3.07 (F1, F2, F3 = 415, 1568, 1875mel; 333, 1964, 2668Hz). The spectral distance between 0.1 and 0.9 probability of /I/ identification was 0.44 (ΔF1, ΔF2, ΔF3 = 18, -30, -19mel; 17, -62, -47Hz, Hertz values calculated at the centre of the planes) indicating a sharp boundary.

2.2. Japanese Listeners
2.2.1. English Test
The discriminant analysis was moderately successful at categorising the Japanese listeners’ responses to the English stimuli (the cross-validated correct classification rate was 85.9%). All variables except speaking rate were entered into the model. This suggests that speaking rate did not affect the Japanese listeners’ vowel perception, and is in accordance with the results of other studies (Strange et al., 1998; Guion et al., 2002, Ingram & Park, 1997) that found little effect for speaking rate on Japanese listeners’ identification of English lax versus tense vowels. The Japanese listeners identified the vowels primarily according to vowel duration but also made considerable use of vowel spectral properties. The relative weightings were 63% for
vowel duration, 28% for vowel spectral properties, and 9% for consonant duration. The predominantly duration-based identification pattern was radically different to the almost exclusively spectral identification pattern used by the native English speakers.

The Japanese listeners’ English /i/-/i/ categorical boundary is graphed in Figure 4. The orientation of the planes indicates that the Japanese listeners based their identification primarily on vowel duration. The mean duration of the boundary was 4.22 (108ms). The vowel duration distance between 0.1 and 0.9 probability of /i/ identification was 1.49 (29ms at the centre of the planes) indicating a relatively sharp boundary. The slopes due to the use of secondary cues resulted in a considerable duration range for the boundary: The duration range for the 0.5 plane was 2.98 to 5.46 (86 to 135 ms), and for the 0.1 and 0.9 planes 2.24 to 6.20 (75 to 155ms).

The Japanese participants’ use of spectral properties may be a result of learning English. Although still relying primarily on transfer of L1 duration criteria they may have realised that English /i/ and /i/ differ spectrally and shifted their cue weighting towards greater use of spectral properties. L1 experience, however, may also have caused sensitivity to spectral properties: Fitzgerald (1996) found that Japanese speakers produced long vowels that were more peripheral in the vowel space than short vowels. He found that, with only F1 and F2 as independent variables, a discriminant analysis on all vowels correctly classified 77% of /i/ and 69% of /i/ vowels. The existence of spectral differences leads to the possibility that spectral properties may in fact be a secondary cue differentiating long-short vowel pairs in Japanese.

2.2.2. Japanese Test
The discriminant analysis was moderately successful at categorising the Japanese listeners’ responses to the Japanese stimuli (the cross-validated correct classification rate was 88.6%). The Japanese listeners based their identification of /i/ and /i/ (Figure 5) Points describing the categorical boundary between Japanese /i/ and /i/ derived from the discriminant analysis carried out on the identification responses of Japanese participants.

Japanese test

<table>
<thead>
<tr>
<th></th>
<th>fast voiceless</th>
<th>fast voiced</th>
<th>slow voiceless</th>
<th>slow voiced</th>
</tr>
</thead>
<tbody>
<tr>
<td>probability of</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>/i/ identification</td>
<td>0.1</td>
<td>0.5</td>
<td>0.9</td>
<td></td>
</tr>
</tbody>
</table>

Japanese test
primarily on vowel duration. The relative weightings given to the variables were 81% for vowel duration, 13% for consonant voicing, and 5% for speaking rate. The small effect due to consonant voicing is consistent with the finding of Tsukada (1999) that the consonant voicing effect on vowel duration in Japanese was small compared to the size of the effect in English. The small effect for speaking rate is consistent with the results of Toda (1999) who found only a small effect for speaking rate on Japanese listeners’ identification of Japanese long versus short vowels.

The categorical boundary points are graphed in Figure 5. The mean duration of the boundary between /i/ and /iː/ was 4.07 (105ms). The vowel duration distance between 0.1 and 0.9 probability of /i/ identification was 1.16 (22ms centred on the mean 0.5 probability) indicating a relatively sharp boundary. Due to the effects from consonant voicing and speaking rate, the duration range for the 0.5 planes was 3.89 to 4.25 (102 to 108 ms), and for the 0.1 and 0.9 planes 3.31 to 4.83 (91 to 121ms).

2.2.3. Comparison of English and Japanese Boundaries
Since the English stimuli all had silent consonant closures, comparison was made between the English boundary and the Japanese /iː/-/i/ boundary based only on the responses to silent-consonant-closure stimuli. Likewise, the Japanese boundary was compared with the English /i/-/iː/ boundary based on stimuli with the same spectral properties as the Japanese stimuli. The formant frequencies of the Japanese stimuli were closest in terms of F1 and F2 to the English stimuli with spectral value 4.

A discriminant-analysis model based on responses to Japanese silent-consonant-closure stimuli resulted in a cross-validated correct classification rate of 89.9%, and a cue weighting of 92% for vowel duration and 8% for consonant duration. The mean Japanese /i/-/iː/ categorical boundary was 4.13 (106ms). The distance between 0.1 and 0.9 probability of /i/ identification was 0.99 (19ms centred on the mean 0.5 probability) indicating a relatively sharp boundary. Due to the effects from consonant closure, the duration range for the 0.5 line was 3.98 to 4.29 (103 to 109ms), and for the 0.1 and 0.9 lines 3.31 to 4.83 (94 to 120ms).

A discriminant-analysis model based on the Japanese listeners’ responses to English /i/ and /iː/ stimuli with spectral value 4 resulted in a cross-validated correct
classification rate of 87.9%, and a cue weighting of 84% for vowel duration and 16% for consonant duration. The mean English /i/-/i/ categorical boundary was 3.82 (100ms). The distance between 0.1 and 0.9 probability of /i/ identification was 1.20 (22ms centred on the mean 0.5 probability) indicating a relatively sharp boundary. Due to the effects from consonant closure, the duration range for the 0.5 line was 3.45 to 4.18 (94 to 107 ms), and for the 0.1 and 0.9 lines 2.85 to 4.78 (84 to 119 ms).

A graphical representation of the Japanese listeners’ mean Japanese /i/-/i/ and English /i/-/i/ boundary is given in Figure 6. The 6ms difference between the mean English and Japanese 0.5 probability points falls within both the English and Japanese 0.1 to 0.9 probability ranges. From these results it can be concluded that the Japanese participants’ identified English /i/ and /i/, and Japanese /i/ and /i:/ using a very similar or identical categorical vowel duration boundary.

2.3. Conclusion

The present study found that Japanese listeners used the same duration criteria to identify English /i/ and /i/ and Japanese /i/ and /i/. This finding supports the hypothesis that Japanese listeners identify English /i/ and /i/ via assimilation to Japanese /i/ and /i:/ proposed by Morrison (2002b).

Speaking rate was found to have no effect on the Japanese listeners’ English /i/-/i/ categorical boundary and only minimal effect on their Japanese /i/-/i:/ boundary. This supports the proposal that the near perfect correct identification rate for English /i/ and /i/ in Ingram & Park (1997) was not due to speaking rate normalisation.

References

Geoffrey Stewart Morrison


Department of Linguistics
University of Ottawa
P.O. Box 450, Stn. A
Ottawa, Ontario K1N 6N5
CANADA

gsm@alumni.sfu.ca