Perception of English /i/ and /ɪ/ by Japanese listeners

Geoffrey Stewart Morrison
Simon Fraser University

This paper reports on a subset of results from a preliminary study in which Japanese listeners identified English vowels in terms of English vowel categories. Between voiced consonants /i/ was primarily identified as /i/, but /i/ between voiceless consonants and /ɪ/ in both voiced and voiceless context were primarily identified as /ɪ/. English tense vowels are longer than lax vowels, but English vowels are also longer before voiced consonants than before voiceless consonants. It is hypothesised that Japanese listeners perceive English /i/ and /ɪ/ according to the phonemic long-short vowel distinction of Japanese, that they assimilate English /i/ in voiced context to Japanese long /i:/, and assimilate English /i/ in voiceless context and English /ɪ/ in both voiced and voiceless context to Japanese short /i/. Spanish listeners, whose first language does not have phonemic duration distinctions, were also tested. Their identification pattern did not match that of the Japanese listeners.

1. Introduction

The present paper reports on a subset of the results of an exploratory pilot study aimed at developing hypotheses regarding the identification of English vowels by Japanese learners of English. The results reported here relate to one of the hypotheses developed from the study, namely that Japanese listeners use the Japanese long-short vowel distinction when attempting to distinguish the English tense-lax vowel pair /i/ and /ɪ/. It is further hypothesised that the use of the Japanese long-short vowel distinction prevents Japanese listeners from adjusting their perception criteria to account for duration differences in English vowels which are due to voicing contrasts in the following consonant. Results from the remainder of the vowels tested in this study are presented in Morrison (2001).

The structure of this paper will reflect the order in which the experiments were carried out and analysed. Results were related to existing literature and hypotheses were developed after conducting the first experiment; hence the literature review will be included in the discussion section of the paper, rather than in the introduction. The paper ends with an outline of future research planned to test the hypothesis developed as a result of the present study.

Although the experiment was not designed to test any detailed hypotheses, certain general hypotheses were implicit in the design of the experiment. It was hypothesised that Japanese learners of English, even at an advanced level of studies, would have some

difficulties with English vowel identification. Also that consonant context may have some effect on the Japanese listeners' identification pattern.

1.1 Japanese vowels

The following description covers the majority of Japanese dialects, including Tokyo and Osaka-Kyoto area dialects. Japanese has 10 vowel phonemes, 5 short vowels /i e a o u/ and 5 long vowels /iː eː aː oː uː/. Japanese is a mora-timed language and long vowels are phonologically analysed as a sequence of two identical vowels associated with adjacent timing slots. The first vowel may be preceded by a consonant associated with the same timing slot. The quality of long vowels are generally considered to be identical to the quality of the corresponding short vowels. Pitch changes, however, may occur at the mora boundary within a long vowel. There may also be slight differences between morphologically derived long vowels and underlyingly long vowels, see Shibatani (1990, p160ff).

1.2 English vowels

The following description describes the English vowels of the speaker (a 29 year-old male) who produced the stimuli for the perceptual experiments outlined below. As a child, the speaker was exposed to English accents from Eastern Scotland and North-Eastern England, and as an adult to General-Canadian English and Canadian-Maritime English. The Speaker's vowel system can best be described as that of Scottish English at the most English end of the Scots-English continuum. The speaker's monophthong phonemes consist of /i, ɪ, e, ɛ, æ, ɜ, ɑ, ɒ, ʌ, ɔ, o, u/ including several vowels which Wells (1982) lists as occurring in some but not all accents of Scottish English. The speaker's English does include a duration distinction between tense and lax vowels.

2 Experiment 1 - English and Japanese listeners

2.1 Methodology

2.1.1 Participants

Two groups of participants were included in the first experiment, a group of Japanese learners of English and a group of native English speakers.

The Japanese group consisted of 51 students from a professional interpreting and translation School in Tokyo. All the Japanese participants were taking interpreting or translation courses (English-Japanese or vice-versa) or advanced level English courses. All were considered to have an advanced level of English, the entry level for the programme being a TOEIC score of 800 (Test of English for International Communication, The Chancery Group International Ltd.).

The native English group consisted of 17 participants: 5 Anglophone Canadians, 1 American, and 1 Scot who were teachers of English resident in Japan; and 6 Scots and 4 Northern English who were university students in Scotland.

Table 1. Words in standard orthography used in the answer booklet to represent the English monophthongs /i, ɪ, e, ɛ, æ, ɜ, ɑ, ɒ, ʌ, ɔ, o, u/ in the consonantal contexts /t _ k/, /k _ t/, /d _ g/, and /g _ d/.

teek took	keet koot	deeg doog	geed gooed
tik	kit	dig	gid
take toak	kate koat	dage dowg	gade goed
tek turktuk tork	ket kertkut kort	degderg dugdorg	gedgerd gudgord
tak tarktok	kat kartkot	dag darg dog	gad gard god

2.1.2 Stimuli and answer booklet

Stimuli consisted of the English vowels /i, ɪ, e, ɛ, æ, ɜ, ɑ, ɒ, ʌ, ɔ, o, u/ spoken by the speaker, details of whose speech are outlined in section 1.2. Each vowel was spoken in four different contexts: /t _ k/, /k _ t/, /d _ g/, and /g _ d/. The speaker produced each stimulus word twice, in random order, blocked by consonant context (four blocks). Each stimulus was numbered, the speaker read the stimulus number, paused for ½ second, read the stimulus word in isolation, and paused for 4 seconds before continuing with the next stimulus. The speaker was recorded in a quiet room using a Sony MZS-R5ST Mini Disc system and a Sony ECM-MS907 microphone. If the speaker misarticulated, the misarticulated stimulus was re-recorded before proceeding with the next stimulus. Instructions for the listeners regarding page turning in the answer booklet and announcements of the beginning of new blocks were also recorded. A warm-up, consisting of one word from each block was included at the beginning of the recording.

The answer booklet contained numbered sections corresponding to the numbered stimuli on the recording. Each section had a group of ten words in standard English orthography corresponding to the audio stimuli for that block. See table 1 for a complete list of words and an example of how the words were arranged in the answer booklet. Whilst some of the words corresponded to real English words, the majority were non-existent but phonotactically possible English words.

The recorded stimuli included two productions of each vowel in each block, both productions will now be referred to together as a stimulus type.

2.1.3 Procedure

The Japanese group was tested in classrooms in class-sized subgroups. The stimuli were played to them via a Sony TCM-1390 cassette / public address system, half the participants were played the original Mini Disc using the Sony MZS-R5ST Mini Disc system, but due to lack of equipment the remaining students were played a first generation cassette copy. First generation cassette copies of the stimuli were mailed to individual native English participants who were instructed to listen to the cassette in a quiet room using the best quality cassette player available to them.

Instructions were presented in writing on the answer booklet and were spoken on the same recording as the stimuli. The participants were instructed to read the four sets of

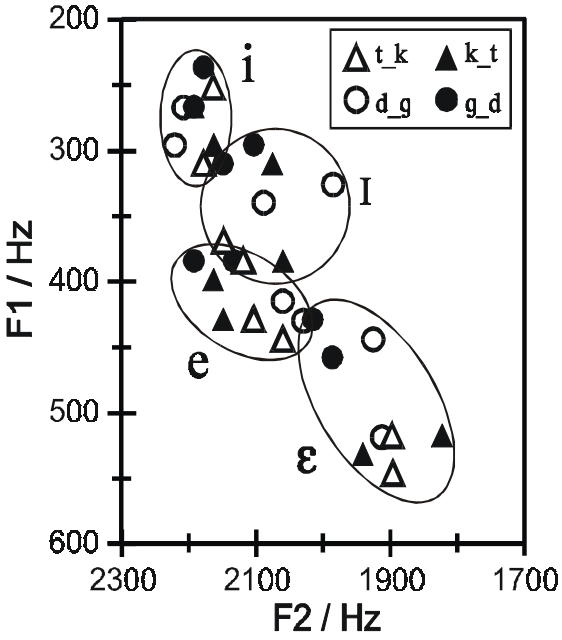

Figure 1. First and second formants of the vowel stimuli used in the identification task.

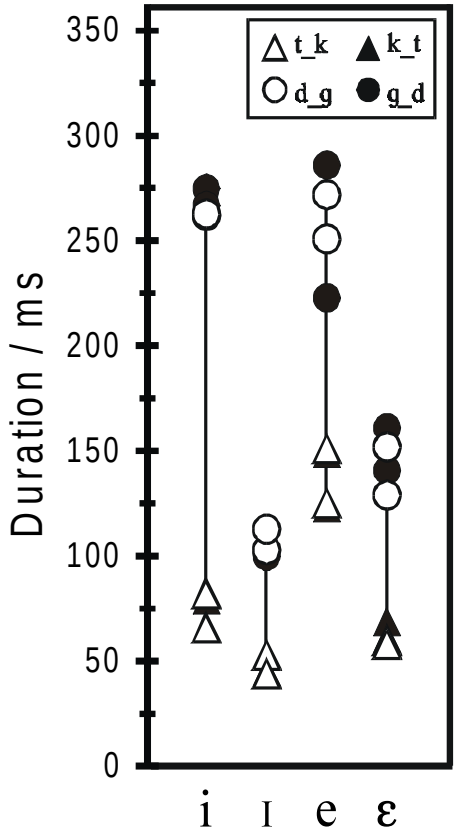

Figure 2. Durations of the vowel stimuli used in the identification task.

words presented in the warm-up section of the answer booklet, and to think about how each word would be pronounced. The Japanese participants were given 3 minutes to do this. The experimenter did not provide the participants with spoken models or explain the pronunciation of each word. The participants were instructed to circle the word in the answer booklet that corresponded to the word they heard on the recording. The warm-up section of the recording was then played. For the Japanese participants, the recording was then stopped and the experimenter (the class teacher) checked that the participants understood the procedure. The remainder of the recording was then played without intervention from the experimenter.

2.1.4 Analysis of identification data

The response data were input to a spreadsheet which was designed to automatically tally the number of times each stimulus type on the recording was identified with a particular word in the answer booklet. The spreadsheet produced confusion matrices for each group of participants and for each consonant context.

2.1.5 Acoustic analysis

The durations and the first two formant frequencies of the vowel stimuli were measured using Speech Analyser 1.06a (SIL, 1998). Vowel duration measurements were taken from the beginning of vowel periodicity following the release burst, and any aspiration, of the initial plosive, to the sudden reduction in periodicity amplitude at the beginning of the closure for the final plosive. Formants were measured from the LPC smoothed Spectrum calculated from a window covering the steady state section of the vowel observed in the spectrogram.

2.2 Results

2.2.1 Native English listeners

Confusion matrices for vowel identification by native English listeners are presented in table 2. As might be expected both /i/ and /ɪ/ were correctly identified at rates of either 100% or close to 100%. The only responses used by native English listeners were non-low front vowels.

2.2.2 Japanese listeners

Confusion matrices for vowel identification by Japanese listeners are presented in table 2. Only non-high front vowel responses are included in the matrices, maximum

Table 2. Confusion matrices for identification of stimuli by native English listeners and by Japanese listeners. Numbers indicate the percentage of particular responses out of the total of responses for particular stimulus types. Percentages are based on the total number of responses per row, 34 for native English listeners and 102 for Japanese listeners if no responses are missing, missing responses are not included in the percentage calculations. Percentage values are rounded to the nearest integer, row totals may not be 100 due to rounding errors. A reported percentage value of 3 for native English listeners and 1 for Japanese listeners represents a single identification response. Only non-low front vowel responses are included in the matrices, native English listeners did not use other responses, Japanese listeners' maximum response scores for other vowels did not exceed 1% except for 3% for /u/ in the /t _ k/ context. Numbers in boxes correspond to correct identifications of the stimuli.

Native English Listeners

/t _ k/

stimuli \ response	i	ɪ	e	ɛ
i	100			
ɪ		97		3

/k _ t/

stimuli \ response	i	ɪ	e	ɛ
i	94			6
ɪ		100		

/d _ g/

stimuli \ response	i	ɪ	e	ɛ
i	94			6
ɪ		100		

/g _ d/

stimuli \ response	i	ɪ	e	ɛ
i	100			
ɪ	3	94	3	

Japanese Listeners

/t _ k/

stimuli \ response	i	ɪ	e	ɛ
i	28	70		
ɪ		95		1

/k _ t/

stimuli \ response	i	ɪ	e	ɛ
i	25	75		
ɪ		91		9

/d _ g/

stimuli \ response	i	ɪ	e	ɛ
i	89	10		1
ɪ		98		1

/g _ d/

stimuli \ response	i	ɪ	e	ɛ
i	100			
ɪ	1	86	1	10

response scores for other vowels did not exceed 1% except for 3% for /u/ in the /t _ k/ context. Spectral and durational properties of the stimuli are presented in figures 1 and 2 respectively.

The Japanese listeners' correct response rates for /ɪ/ are almost as high as those of native English speakers, overall 92% versus 98%. Scores appear to be slightly lower in the velar-alveolar contexts with /ɛ/ as the modal incorrect response. The choice of /ɛ/ rather than spectrally intermediate /e/ may be due to the similarity of duration between relatively short /ɪ/ and /ɛ/ compared to longer /e/.

The Japanese listeners' correct response rates for /i/ in voiceless contexts are much lower than in voiced contexts. Mean correct response rates are 95% in voiced contexts, but only 27% in the voiceless contexts. The modal incorrect response is /ɪ/. The /ɪ/ stimuli are spectrally adjacent to /i/ and have a similar duration (in both voiced and voiceless contexts) to /i/ in voiceless contexts. The /i/ stimuli are substantially longer in voiced contexts which may preclude their identification with /ɪ/.

2.3 Discussion

2.3.1 Theoretical models

The current theoretical frameworks dealing with second language speech perception are Best's Perceptual Assimilation Model (PAM) (Best, 1994, 1995a, 1995b) and Flege's Speech Learning Model (SLM) (Flege, 1981, 1987a, 1991a, 1992a, 1995a). The PAM provides a theory explaining how listeners interpret novel speech sounds and how this is related to the development of first language speech perception. The SLM provides a theory relating to how perception of second language speech sounds changes as learners become more experienced with the language. The SLM also addresses the influence of speech perception on second language speech production.

The Perceptual Assimilation Model (Best, 1994, 1995a, 1995b) deals primarily with the assimilation of foreign language speech sounds to first language speech sounds. A foreign language speech sound may be assimilated to a native language speech sound, that is, perceived as being an example of the native language speech sound. In this case the foreign sound may be perceived as a good example, or an acceptable but not perfect example, or a poor example of the native sound. Alternatively, the foreign sound may be perceived as a speech sound that is not identifiable with any native sound, or may be perceived as a non-speech sound. This results in three assimilation types: assimilated as a native category, assimilated as an uncategoriseable speech sound, and not assimilated to speech, with three levels of goodness of fit to the native category in the first type. Pairs of foreign speech sounds may be assimilated differently leading to different patterns of assimilation including *two category assimilation* in which each foreign speech sound is assimilated to a different native category, *single-category assimilation* in which both sounds are assimilated as equally good examples of a single native category, and *category-goodness assimilation* in which both sounds are assimilated to a single native category but are not equally good examples. In the first case the listener's ability to distinguish the foreign sounds is expected to be excellent, in the latter cases the ability to distinguish the

foreign sounds will depend on the degree to which each speech sound is a good example of the native category.

The Speech Learning Model (Flege, 1981, 1987, 1991a, 1992a, 1995) predicts that when a second language speech sound is perceived as a *new* speech sound, this will lead to the development of a new speech sound category. Since the new category will only be based on the properties of the second language speech sound, it will be learnt accurately according to the norms of the second language. When a second language sound is perceived as *similar* to a native speech sound (equivalence classification) this will not result in a new speech category but rather a modification of the existing native category. Since this category will now be influenced by both the native and second language sounds, the resulting pronunciation will ultimately be intermediate between the native and second language sounds. Production of second language speech sounds will be based on the learner's perceptual categories. The SLM emphasises that perception of second language speech sounds is not based on counterparts in the phoneme inventories of the first and second languages, but in the properties of the phonetic realisations of speech sounds in particular contexts. Hence, different allophones of a single phoneme in the second language may be perceived in different ways depending on their similarity to allophones in parallel contexts in the native language. Whilst the PAM operates under a direct realism model and specifically claims that the speech sound properties perceived are articulatory gestures, research based on the SLM has tended to work on the assumption that the acoustic properties of the speech sounds are the basis of perception.

The Perceptual Magnet Effect (Kuhl, 1991; Kuhl et al, 1992; Kuhl & Iverson, 1995) applies particularly to vowel perception. Examples of vowels in the vicinity of a vowel prototype are perceived as closer to the vowel prototype than would be suggested by the spectral properties of the vowel. Whilst the Perceptual Magnet Effect aids in the classification of native language vowels, it may hinder the development of second language vowel perception by making it more likely that non-native vowels will be assimilated as good examples of native vowel categories. Second language learners would then be more likely to identify vowels as similar or even identical to native vowel categories, and would not learn to perceive or pronounce the vowels in the same way as native speakers.

According to the SLM, second language learners may base their perceptual categories on criteria that are different, or differently weighted, to those of native speakers. With reference to vowels, primary criteria for identification are considered to be spectral properties, such as the first two formant frequencies, and duration. Bohn & Flege (1990), Bohn (1995) and Bohn & Flege (1996) found that German listeners primarily made use of duration in identifying English /ɛ/ - /æ/ continua, whereas native English listeners primarily used spectral cues. This was not unexpected since German lacks a vowel in the space occupied by English /æ/, and does make use of phonemic duration, having both an /ɛ/ and /ɛː/ phoneme. However, Bohn (1995) and Flege, Bohn & Jang (1997) also found that Spanish, Mandarin, and Korean listeners made substantial use of duration to distinguish members of English /i/ - /ɪ/ continua. Native English listeners overwhelmingly used spectral differences to distinguish members of the same continua. The Mandarin and Korean listeners may have had exposure to contrastive length differences in their native languages, but the result was totally unexpected for the Spanish

listeners since Spanish does not use duration contrastively. This led Bohn (1995) to develop the Desensitisation Hypothesis which states that should listeners be unable to make use of spectral cues to distinguish non-native vowels (because their native language has not sensitised them to spectral differences in that part of the vowel space) they will instead use duration, irrespective of whether duration is used in their native language.

2.3.2 Relation between results and theoretical models

The most striking result in the present study is the very high (73%) level of identification of /i/ with /ɪ/ in the voiceless contexts, whilst /i/ in voiced contexts and /ɪ/ in both voicing contexts were correctly identified at rates of over 90%. The pattern can be explained in terms of assimilation to Japanese vowel categories. In terms of spectral properties, Japanese only has one vowel in the high front part of the vowel space, hence it might be expected that both English /i/ and /ɪ/ would be assimilated to this vowel (single-category or category-goodness assimilation). If both English vowels were assimilated to a single Japanese category, the Japanese listeners' ability to distinguish them would be expected to be poor. However, Japanese makes a phonemic duration distinction resulting in two high front vowels, one long (/i:/) and one short (/i/). Since the English tense-lax pair /i/ and /ɪ/ differ in terms of duration as well as spectral properties, it might be expected that the Japanese listeners, being unable to make use of spectral cues, would assimilate English /i/ to long Japanese /i:/ and English /ɪ/ to short Japanese /i/ (two-category assimilation). However, this reasoning is based on phonemic categories and does not consider allophonic differences as required by the SLM. Vowels in many languages have been found to be longer before voiced consonants than before voiceless consonants. In English this duration difference is particularly large, sufficient to cue a phonemic voicing distinction for the following consonant (Kluender, Diehl, & Wright, 1988; Crowther & Mann, 1992). Duration differences are evident in the vowel stimuli used in the present study, vowels in voiced contexts are longer than in voiceless contexts (see figure 2).

Whilst the /i/ stimuli in voiced contexts have durations in the range 262 - 275ms, /i/ in voiceless contexts and /ɪ/ in both voiced and voiceless contexts can be grouped together in the range 44 - 113ms. This leads to the hypothesis that English /i/ in voiced context is assimilated to Japanese long /i:/, and English /i/ in voiceless contexts and /ɪ/ in both voiced and voiceless contexts are assimilated to Japanese short /i/. Also it is argued that the Japanese listeners selected the English /i/ responses (*deeg* and *geed*) to identify the vowels which they has assimilated to Japanese long /i:/, and selected the English /ɪ/ responses (*dig*, *gid*, *tik*, *kit*) to identify the vowels which they had assimilated to Japanese short /i/.

This hypothesis can only be maintained if the categorical boundary between short Japanese /i/ and long Japanese /i:/ lies between 113ms (the upper duration bound for the English /i/ stimuli in voiceless contexts and /ɪ/ stimuli in voiced and voiceless contexts) and 262ms (the lower duration bound for the English /i/ stimuli in voiced contexts). Toda (1999) used duration continua to investigate categorical thresholds for short versus long

Japanese vowels (/e/, /a/ and /o/) in the final position of isolated CVCV words. When different speech rates were stimulated by varying the duration of the first vowel to 70% and 130% of its original duration, the long-short threshold for the second vowel also varied. The average threshold durations, however, varied over a relatively small range from 231ms (the shortest /e/ threshold) to 267ms (the longest /a/ threshold). Toda's results are not incompatible with the listeners in the present study having a categorical threshold around the lower bound of the present study's English /i/ stimuli in voiced context. A boundary at this point would even account for the fact that the /i/ stimuli in /d _ g/ context are shorter than in /g _ d/ and are 10% identified as /ɪ/ whereas the /i/ stimuli in /g _ d/ are 100% identified as /i/. However, there is a difference in the contexts and vowels tested in Toda (1999) compared to the contexts and vowels tested in the present study. English vowels in word-final open syllables tend to be longer than vowels in word-internal syllables, this might lead to the expectation of a shorter long-short vowel threshold for the CVC contexts used in the present study.

Tsukada (1996) measured the duration of long and short Japanese vowels, including /i/ and /i:/, produced by native Japanese speakers in isolated /CVto/ and /CVdo/ words. This is a closer match for the English vowels and contexts used in the present study. In Tsukada (1996), Japanese vowels were longer before a voiced consonant than before a voiceless consonant by an average ratio of 1:1.06. The mean duration for "kido" was 60ms and the mean duration for "shiito" was 190ms (values estimated from figure 2 of Tskuda, 1996). This suggests a threshold value well below the duration of the present study's English /i/ stimuli in voiced contexts and does not preclude a value higher than the duration of the English /i/ stimuli in voiceless contexts and /ɪ/ stimuli in voiced and voiceless contexts.

Direct evidence for the spectral assimilation of American English vowels to Japanese vowel categories is provided by Strange et al (1998). English /i/ and /ɪ/ in /hVbɑ/ context were assimilated to Japanese high front vowels at rates of 99% and 59% respectively. The remaining 41% of the English /ɪ/ stimuli were assimilated to Japanese mid front vowels. The same pattern, although smaller in magnitude, was found in the present study, /ɪ/ stimuli in the velar-alveolar contexts were identified as /ɛ/ at a rate of 10%. However, /ɪ/ stimuli in the velar-alveolar contexts were not identified with /ɛ/ (except by at a negligible rate of 1%). Combining the consonant contexts in Strange et al. (1998) with those of the present study it appears that the likelihood of /ɪ/ being assimilated to a mid front Japanese vowel is greater before bilabial plosives than before alveolar plosives, and least before velar plosives. Such a pattern could be the result of the acoustic differences in formant transitions leading into bilabial, alveolar, and velar plosives, or may be the result of differences in auditory or higher order perception. However, it should be noted that the majority of /e/ assimilations in Strange et al (1998) were in response to the productions of only one of the four speakers in one of the two presentation conditions (isolated disyllabic words as opposed to words in a carrier sentence). Further investigation would be needed to establish whether this is a stable pattern and whether it is due to acoustic or auditory factors.

Strange et al (1998) also measured temporal assimilation patterns, both for when the vowel was presented in a disyllabic word in isolation, and for when the word was presented within a carrier sentence. The modal response to English /ɪ/ was short Japanese /i/ in both sentence and disyllable condition, assimilated at rates of 58% and 77% respectively. The modal response to English /i/ was 83% long Japanese /iː/ in sentence condition, and 59% short Japanese /i/ in disyllable condition. Whilst acoustic measures revealed that some variation in assimilation patterns could be assigned to duration differences across speakers and conditions, not all variation could be assigned to absolute or relative duration differences. Strange et al (1998) concluded that the larger rhythmic structure of the sentence had an effect on the temporal assimilation pattern.

Strange et al's (1998) isolated disyllable /hVbɑ/ was the closest match to the present study's voiced /d _ g/ and /g _ d/ contexts. Comparing these contexts, English /i/ is hypothesised to be assimilated to Japanese long /iː/ at a rate of 95% in the present study and is assimilated at a rate of 59% in Strange et al (1998). English /ɪ/ is hypothesised to be assimilated to Japanese short /i/ at a rate of 92% in the present study and is assimilated at a rate of 77% in Strange et al (1998). The higher assimilation rates in the present study may be due to either greater relative differences in duration between the /i/ and /ɪ/ stimuli used, or the absolute durations of the stimuli being further, in opposite directions, from the duration threshold used by the Japanese listeners for isolated word context.

Ingram & Park (1997) had Japanese and Korean learners of English identify Australian English front vowels in isolated /hVd/ words. Listeners were asked to identify the stimuli both in terms of English vowels and in terms of the vowels of their first language. Words were recorded by two male speakers and although there was little interspeaker difference in the vowels' spectral properties, there were considerable interspeaker duration differences. The Korean listeners were influenced by the absolute duration of individual vowels, both in the native language and the English category identification tests. The Japanese listeners, however, consistently identified the stimuli irrespective of the interspeaker duration differences. English /i/ was 100% identified as Japanese long /iː/ and 99% identified as English /i/, English /ɪ/ was 91% identified as Japanese short /i/ and 99% identified as English /ɪ/. Ingram & Park (1997) hypothesised that because of their experience with phonemic duration differences in their native language, the Japanese listeners were able to adjust for the interspeaker duration differences. If this hypothesis is correct, it raises the question of why the Japanese listeners in the present study were not able to adjust for duration differences due to phonological context differences. Since the stimuli in Ingram & Park (1997) were presented in isolated words, it would not be possible for the listeners to use cues from the larger rhythmic structure of a sentence to adjust for interspeaker speaking rate differences (the larger rhythmic structure was identified as important in Strange et al, 1998). Examination of the duration of Ingram & Park's (1997) stimuli reveal that the longest /ɪ/ token (from the speaker who might be designated the slowest talker) was 165ms and the shortest /i/ token (from the speaker who might be designated the fastest talker) was 255ms long (values estimated from figure 2 of Ingram & Park, 1997). If the duration threshold for long versus

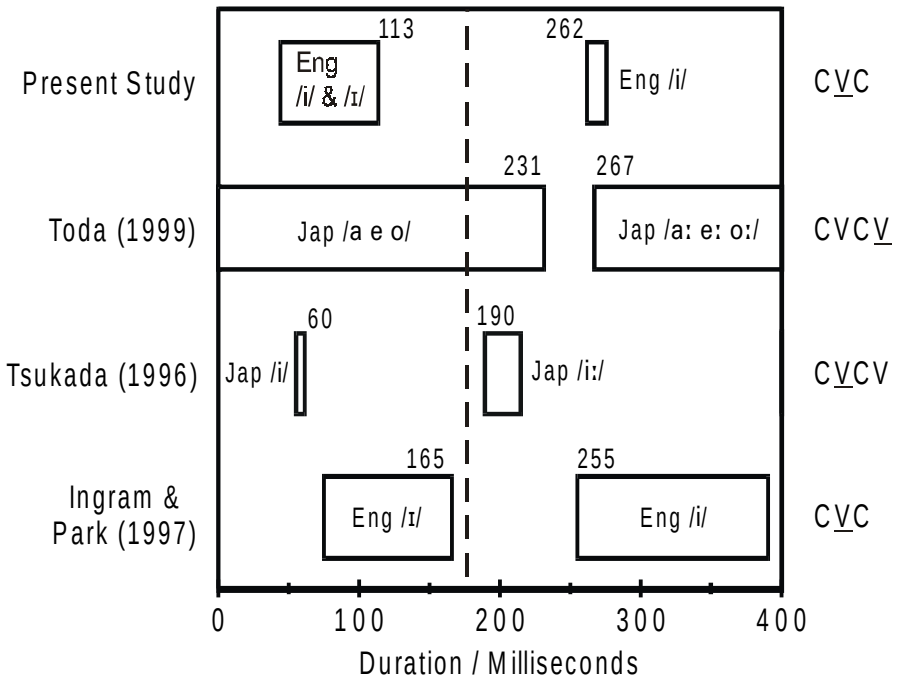

Figure 3. A comparison of vowel duration data from four studies. The dotted line represents the position of a vowel duration threshold which would be in accordance with the data from the present study, Tsukada (1996), and Ingram & Park (1997). The underlined V to the right of the figure indicate the context of the vowels measured in each study. The numbers next to the boxes indicate the upper and lower duration bounds observed for short and long vowels respectively (the numbers for Tsukada, 1996, represent mean values). The data from Toda (1999) represent word final vowels and the numbers represent the upper and lower bounds of observed threshold duration.

short Japanese vowels in isolated word context lies between 165ms and 255ms, then Ingram & Park's (1997) results can be accounted for without hypothesising that the listeners were able to normalise for speaker dependent duration differences.

A comparison of data from the studies above relating to Japanese vowel duration thresholds is presented in figure 3. The data from the present study, Tsukada (1996), and Ingram & Park (1997) suggest a duration threshold of around 175ms for Japanese short versus long /i/ in CVC isolated word contexts (this threshold value is indicated by the dotted line in figure 4). This is compatible with the hypothesis that Japanese listeners primarily use the Japanese long-short vowel duration threshold to distinguish English /i/ and /ɪ/.

Although duration is hypothesised to be the primary cue used by Japanese listeners, some use of spectral information is also apparent. Whilst the listeners in both Strange et al (1998) and Ingram & Park (1997) identified English /i/ and /ɪ/as Japanese /i/, they rated English /ɪ/ as less prototypical of the Japanese category than English /i/. Also the listeners in Ingram & Park (1997) identified English /ɪ/ tokens as English /ɪ/ at a rate of 99%, higher than the 91% rate at which they assimilated these tokens to short Japanese /i/. Some use of spectral information in the present study would account for the

fact that /i/ in voiceless context was correctly identified as /i/ at a rate of 27% rather than being identified with /ɪ/ at a rate of 100% as would be predicted from durational properties alone. This would be consistent with category-goodness assimilation, with English /ɪ/ being a poorer example of Japanese short /i/ than English /i/.

The hypothesis developed here is that English /i/ in voiced context is assimilated to Japanese long /i:/, and that English /i/ in voiceless context and /ɪ/ in both voiced and voiceless context are assimilated to Japanese short /i/. This represents a duration-based two-category assimilation pattern. Within this two-category assimilation pattern there appears to be a secondary spectrally-based category-goodness assimilation pattern where English /ɪ/ is a poorer example of Japanese short /i/ than is English /i/ in voiceless contexts. The two-category assimilation is hypothesised to be due to the transfer of the phonemic long-short duration contrast in Japanese. In order to ascertain whether the results found here are likely to be due to a transfer effect from Japanese, rather than due to a general second language learning strategy resulting from Bohn's (1995) desensitisation hypothesis, a comparison must be made between the identification pattern of the Japanese listeners and the identification pattern of listeners whose first language has a similar vowel inventory in terms of quality but has no phonemic duration distinction. A second experiment was therefore conducted using Spanish listeners.

3 Experiment 2 - Spanish listeners

3.1 Spanish vowels

The following description covers peninsular Spanish as spoken in central and northern Spain. Spanish has five monophthongs /i e a o u/ with back vowels realised as rounded and other vowels realised as non-rounded. Spanish does not have phonemic vowel duration as in Japanese, nor does it use differences in vowel duration to signal voicing contrasts in following consonants as in English.

3.2 Methodology

3.2.1 Participants

The Spanish participants consisted of 14 second-cycle (third and fourth year undergraduate) English majors at the University of the Basque Country in Vitoria-Gasteiz, and 7 upper intermediate level students in a business English school in the same city (a total of 21 participants). The university students were all enrolled in an English-Spanish translation course and had taken an English pronunciation course at an earlier stage in their studies. The researcher, who knew the Spanish university students and taught the Spanish business school students and a subgroup of the Japanese interpreting-translating school students, judged the English speaking ability and overall intelligibility of accent of the Spanish students to be comparable with that of the Japanese students.

3.2.2 Stimuli, procedure, and analysis

The stimuli, answer booklet, experimental procedure and statistical analysis were the same as used with the Japanese participants. The stimuli were played directly from the original Mini Disc recording using the Sony MZS-R5ST Mini Disc system and Sony SRS-A37 speakers.

3.3 Results

Confusion matrices for vowel identification by Spanish listeners are presented in table 3. The /i/ stimuli in the /t _ k/ context are identified 50% as /i/ and 50% as /ɪ/. In the /k _ t/ context the Spanish listeners identified /i/ with /i/ at a rate of 36% and with /ɪ/ at a rate of 64%. A one-sample two-tailed t-test reveals that there is no reason to reject, at an alpha level of .05, the hypothesis that the 36%-64% result from the /k _ t/ context is a sample from a population with a true identification rate of 50%-50% [$t(41) = 1.909$, $p = .063$]. (In statistical tests /i/ responses were coded as +1 and /ɪ/ responses coded as -1.) These results suggest that the Spanish listeners are unable to distinguish the English /i/ and /ɪ/ response categories, and so assign the /i/ stimuli in voiceless contexts to one or the other category at random.

This result is quite different to that from the native English listeners who had 100% correct identification and to that of the Japanese listeners who more strongly favoured the /ɪ/ response (at a rate of 70%). A two-way ANOVA comparing the identification of /i/ in /t _ k/ and /k _ t/ contexts for Spanish and Japanese listeners revealed that, at an alpha level of .10, there was a significant effect for first language group [$F(1, 282) = 6.784$, $p = .0097$] but not for consonant context [$F(1, 282) = 2.204$, $p = .139$] or for the interaction between the two [$F(1, 282) = .139$, $p = .369$]. There is

Table 3. Confusion matrices for identification of stimuli by Spanish listeners. Numbers indicate the percentage of particular responses out of the total of responses for particular stimulus types. Percentages are based on the total number of responses per row, 42 if no responses are missing, missing responses are not included in the percentage calculations. Percentage values are rounded to the nearest integer, row totals may not be 100 due to rounding errors. A reported percentage value of 2 represents a single identification response. Only non-low front vowel responses are included in the matrices, Spanish listeners did not use other responses. Numbers in boxes correspond to correct identifications of the stimuli.

Spanish Listeners

/t _ k/

stimuli \ response	i	ɪ	e	ɛ
i	50	50		
ɪ	14	83		2

/k _ t/

stimuli \ response	i	ɪ	e	ɛ
i	36	64		
ɪ	21	74	2	2

/d _ g/

stimuli \ response	i	ɪ	e	ɛ
i	81	19		
ɪ	2	98		

/g _ d/

stimuli \ response	i	ɪ	e	ɛ
i	88	12		
ɪ	7	90		2

therefore evidence to suggest that the Spanish listeners' identification of /i/ in voiceless contexts represents a random selection of /i/ and /ɪ/ responses, and that this differs from the Japanese listeners' identification which is biassed towards /ɪ/ responses.

In voiced contexts the Spanish listeners' correct identification rates for /i/ are relatively high, mean 85%, although lower than the native English and Japanese listeners' scores, mean 97% and 95% respectively. The only incorrect response chosen by the Spanish listeners is /ɪ/.

The Spanish listeners' correct response rates for /ɪ/ in voiced contexts are very high and similar to the correct response rates for the native English and Spanish listeners, mean 94% compared to 97% and 92% respectively. In voiceless contexts, however, the Spanish listeners' correct identification rates are lower, mean 79% compared to 99% and 93% respectively. The Spanish listeners' modal incorrect response is /i/ in all cases.

3.4 Discussion

On a phonemic level, it might be expected that English /i/ and /ɪ/ would both be assimilated to Spanish /i/. On a phonetic level, Flege (1991b) compared spectral data from English and Spanish vowels and predicted that whilst English /i/ would be assimilated to Spanish /i/, English /ɪ/ would be assimilated to Spanish /e/ and /i/ since English /ɪ/ overlapped the vowel space of Spanish /e/. This prediction was born out by the results of Flege's (1991b) experiments which asked American Spanish listeners to identify American English vowels in terms of Spanish categories; whilst English /ɪ/ was primarily identified with Spanish /i/ it was also identified with Spanish /e/ to a significant extent. This may have lead to a prediction that the Spanish listeners in the present study would identify English /ɪ/ with English /e/ or /ɛ/ at a relatively high rate. This is clearly not the case for the Spanish listeners (see table 3), although it might be possible to employ the same line of reasoning to account for the /ɪ/ stimuli identified as /ɛ/ by the Japanese listeners. A possible reason for the difference in results between the present study and Flege (1992) is that all the /ɪ/ stimuli in the present study were higher in the vowel space (F1 range 296-385Hz.) than the /ɪ/ stimuli spoken by male speakers in Flege's study (F1 range 389-471Hz.).

In the present study, there is evidence that the Spanish listeners have established separate categories for the two English vowels and that both spectral properties and duration appear to be relevant criteria. This is likely to be the result of a category-goodness assimilation where some stimuli are perceived as better examples of Spanish /i/ than others. Correct identification scores for both /i/ and /ɪ/ are highest in the voiced contexts, this may be due to a greater difference between the duration of /i/ and /ɪ/ (see figure 2) or to a greater ability to extract spectral information from longer stimuli. The fact that correct identification scores for /ɪ/ are lowest in the voiceless contexts, where the duration of the stimuli are shortest, suggests that this vowel is not identified using duration alone (this applies both for absolute and categorical duration). This analysis would be in accordance with the findings of Bohn (1995) and Flege, Bohn & Jang (1997)

that Spanish listeners gave approximately equally weight to spectral and durational cues when identifying members of American English “beat” - “bit” continua.

Statistical analysis suggest that the Spanish listeners selected /i/ and /ɪ/ responses at random when presented with English /i/ stimuli in voiceless contexts. Assimilation patterns in voiceless contexts are therefore single-category rather than category-goodness. Thus these stimuli must be both spectrally and durationally ambiguous relative to the identification criteria used by the Spanish listeners. From examination of the spectral properties of the English /i/ and /ɪ/ stimuli (see figure 1), it appears that there is actually greater separation between /i/ and /ɪ/ in voiceless contexts than in voiced contexts. This leads to the conclusion that duration must be a relevant factor. The relative duration difference between the two vowels in voiceless contexts is much less than in voiced contexts (see figure 2) corresponding to the lower correct identification rate. The relative and absolute duration difference between English /i/ and /ɪ/ is greater in voiced contexts than in voiceless contexts, and the Spanish listeners’ ability to distinguish the vowels is greater in voiced contexts. It may be that the Spanish listeners are using duration to distinguish the two English vowels but that, unlike the Japanese listeners, they have developed a contrast based on duration differences which is sensitive to voiced versus voiceless consonant contexts.

4 General Discussion

The major hypothesis developed as a result of the present study is that Japanese listeners distinguish English /i/ and /ɪ/ primarily using the same duration criteria used to distinguish Japanese long /i:/ versus short /i/. Although English tense-lax pairs do differ in terms of duration, this cue is only of secondary importance for native English listeners, who rely more on spectral cues (Bohn & Flege, 1990; Bohn, 1995; Bohn & Flege, 1996; and Flege, Bohn & Jang, 1997). Since the duration of English vowels is affected by the voicing of the following vowel, this may affect the assimilation of English /i/ and /ɪ/ to Japanese categories. In particular, English /i/ before a voiced consonant will tend to be long enough to be assimilated to Japanese long /i:/, but English /i/ before a voiceless consonant and English /ɪ/ before either a voiced or voiceless consonant will tend to be short enough to be assimilated to Japanese short /i/. Under such an assimilation pattern, Japanese learners of English would not be expected to develop new perceptual categories for English /i/ and /ɪ/. Lacking new English categories, Japanese listeners would continue to rely on the existing Japanese categories, long /i:/ and short /i/, in order to perceive the difference between English /i/ and /ɪ/. It is therefore assumed that when Japanese learners of English identify a vowel as English /i/ they are indicating a vowel which has been assimilated to Japanese long /i:/, and that when they identify a vowel as English /ɪ/ they are indicating a vowel which has been assimilated to Japanese short /i/. Japanese listeners do appear to make some use of spectral information in distinguishing English /i/ and /ɪ/, hence a small number of English /ɪ/ stimuli may be assimilated to a Japanese short vowel

but identified with English /i/. However, duration appears to overwhelmingly be the primary cue used by Japanese listeners.

Spanish listeners also use duration and spectral information to distinguish English /i/ and /ɪ/, however, they do not use duration in the same way as Japanese listeners. It is not clear from the present study whether Spanish listeners primarily use spectral or durational cues. Even if durational cues are primary, it appears that Spanish listeners use relative rather than categorical durational criteria. The Spanish listeners do seem, to some extent, to be able to adapt their criteria to allow for differences in vowel duration due to the voicing condition of the adjacent consonants. It is therefore claimed that the English /i/ and /ɪ/ vowel identification pattern observed for the Japanese listeners in the present study is due to the transfer of the categorical long-short vowel distinction from their first language.

The influence of consonant voicing context on non-native perception of English vowels has not previously been observed since earlier studies have tended to confine themselves to a single consonant context, for example Bohn (1995) and Flege, Bohn & Jang (1997) used a /b _ t/ context, Ingram & Park (1997) used a /h _ d/, and Strange et al (1998) used a /h _ bɑ/ context.

One caveat in the interpretation of the identification patterns in the present study is that since the responses *kit* and *dig* correspond to common real words, the listeners may have favoured these responses over *keet* and *deeg* which do not correspond to real words. No such effect would be possible between *gid* and *geed* which neither of which correspond to real words, an effect is also unlikely between *tik* and *teak* since both correspond to real words which are probably equally likely to be unfamiliar to the Spanish and Japanese listeners. This may account for the slightly higher rate of /ɪ/ identification in /k _ t/ and /d _ g/ contexts for both the Spanish and Japanese listeners. However, any such bias does not appear to have obscured effects due to duration differences which form the basis of the major hypothesis developed here.

5 Future research

Further research is planned in order to obtain evidence that will support or disprove the above hypothesis. In order to ensure that the properties of the vowel stimuli match the vowels of the particular English dialect which the non-native listeners are learning, the research will be conducted using students who are resident in an English speaking country. The non-native speakers will be tested shortly following their arrival and again several months later. This will provide data which may be compared with the claims of the SLM concerning the development of second language speech sound perception and production. Perceptual data will be gathered using synthetic speech continua. For English /i/ and /ɪ/ continua, the properties varied will be the first two formants, the duration of the vowel, the duration of the closure of the following plosive, and the speaking rate of the carrier sentence. The words used will be “bit”, “beat”, “bid”, and “bead”, all real English words. This will allow investigation of the proposal by Strange et al (1998) that the larger metrical structure of the sentence will affect the Japanese listeners use of categorical duration. It will also allow an investigation into the effects of

vowel-plosive duration ratio on both vowel identification and on the perception of plosive voicing. It will be possible to compare this data with the results of Crowther & Mann's (1992) investigation of the influence of vowel duration on Japanese listeners' perception of voicing in post-vocalic English plosives. In order to test whether the Japanese listener's durational perception of the English continua matches their perception of Japanese long versus short vowels, they will be asked to identify the English continua in terms of Japanese words as well as English words, and they will also be tested on similar Japanese long /i:/ - short /i/ continua. The properties varied in the Japanese continua will be the duration of the vowel, the duration of the closure of the following plosive, and the speaking rate of the carrier sentence. Whereas phonation is not expected during the closure of either the phonemically voiced or voiceless English plosives, phonation is expected during the closure of the Japanese voiced plosives. The English continua will therefore be synthesised with no phonation, whereas two parallel Japanese continua will be synthesised, one with phonation and the other without. Production data for Japanese and English will also be gathered in order to compare the acoustic properties of the Japanese participants' productions with their perception. In order to ascertain whether the results from the Japanese participants are due to the Japanese phonemic length distinction or to a general second language learning strategy, Spanish speaking participants will also be tested on the English continua, and English and Spanish production data will be collected from them. A control group of native English participants will also provide production data and be tested on the English continua.

Acknowledgments

Thanks to staff and students at Inter School, Tokyo, the University of the Basque Country, McDonnell Language Services, and Vitoria-Gasteiz Chamber of Commerce. Thanks also to M. J. Munro, O.-S. Bohn, and M. Fourakis for comments on an earlier version of this paper.

gsm@alumni.sfu.ca
Correspondence concerning this paper should be addressed to ~~geoff@japan.co.jp~~

References

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* (Vol. 9, pp. 171-204). Baltimore: 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.

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). Baltimore, York Press.

Bohn, O.-S., & Flege, J. E. 1990. Interlingual identification and the role of foreign language experience in L2 vowel perception. *Applied Psycholinguistics, 11*, 303-328.

Bohn, O.-S., & Flege, J. E. 1996. Perception and production of a new vowel category by adult language learners. In A. James & J. Leather (Eds.), *Second language speech: Structure and process* (pp. 53-73). Berlin: Mouton de Gruyter.

Crowther, C., & Mann, V. 1992. Native language factors affecting use of vocalic cues to final consonant voicing in English. *Journal of the Acoustical Society of America, 92*, 711-722.

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

Flege, J. E. 1987. The production of "new" and "similar" phonemes in a foreign language: Evidence for the effect of equivalence classification. *Journal of Phonetics, 15*, 47-65.

Flege, J. E. 1991a. Perception and production: The relevance of phonetic input to L2 phonological learning. In Heubner, T. & Ferguson, C. (Eds.), *Crosscurrents in second language acquisition and linguistic theory*. (pp. 249-284). Philadelphia: John Benjamins.

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

Flege, J. E. 1992. Speech learning in a second language. In C. Ferguson, L. Menn, & C. Stoel-Gammon (eds.), *Phonological development: Models, research, and application* (pp. 565-604). Timonium, MD: York Press.

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). Baltimore: York Press.

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.

Ingram, 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.

Kluender, K. Diel, R., & Wright, B. 1988. Vowel-length differences before voiced and voiceless consonants: An auditory explanation. *Journal of Phonetics, 16*, 153-169.

Kuhl, P. K. 1991. Human adults and human infants show a "perceptual magnet effect" for the prototypes of speech categories, monkeys do not. *Perception and Psychophysics, 50*, 93-107.

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). Baltimore: York Press.

Kuhl, P. K., Williams, K. A., Lacerda, F., Stevens, K. N., & Lindblom, B. 1992. Linguistic experience alters phonetic perception in infants by 6 months of age. *Science, 255*, 606-608.

Morrison, G. S. 2001. Perception of English vowels by native English, Japanese, and Spanish listeners. In preparation.

Shibatani, M. 1990. *The languages of Japan*. Cambridge, UK: Cambridge University Press.

SIL. 1998. Speech Analyser 1.06a [software]. Waxhaw, NC: Author. Available http//:www.sil.org

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

Toda, T. K. 1999. Development of speech discrimination by learners of Japanese as a second language: Remarks from a longitudinal study. In P. Robinson (Ed.), *Representation and Process: Proceedings of the 3rd Pacific Second Language Research Forum: Vol. 1* (pp 207-233). Tokyo: Pacific Second Language Research Forum.

Tsukada, K. 1996. Acoustic analysis of Japanese-accented vowels in English. In P. McCormick & A. Russell (Eds.), *Proceedings of the 6th Australian International Conference on Speech Science and Technology* (pp. 373-378). Canberra: Australian Speech Science and Technology Association.

Wells, J. C. 1982. *Accents of English 2: The British Isles*. Cambridge, UK: Cambridge University Press.