

Acoustic Properties of Japanese and English Vowels: Effects of Phonetic and Prosodic Context

MIWAKO HISAGI, KANAE NISHI, AND WINIFRED STRANGE
The City University of New York-Graduate Center

1. Introduction

Adult learners of foreign languages have considerable difficulty learning to perceive and produce some, but not all, non-native vowels (e.g., Nishi et al., 1998). According to Flege's Speech Learning Model (Flege, 1995) non-native vowels that are "phonetically similar" to native vowels will be perceptually assimilated to native categories, resulting in interlanguage (IL) representations which subsume both L1 and L2 vowels. This will lead to accented production of the L2 vowels, and to perceptual confusions if two L2 vowels are assimilated to the same L1 category. In contrast, non-native vowels which are highly dissimilar phonetically from L1 vowels will (with L2 experience) come to be represented separately from all L1 categories in IL, and may be produced as more native-like. In predicting these patterns of perceptual assimilation, many previous studies have inferred phonetic similarity/dissimilarity from cross-language comparisons of the acoustic structure of vowels produced in only one or a few phonetic and prosodic contexts (often in CV or CVC syllables produced in lists). Previous re-

search has documented large differences in the phonetic realization of American English (AE) vowels as a function of phonetic context (Hillenbrand et al., 2001) and prosodic context (Fourakis, 1991; Strange et al., submitted). Less research has been published on the allophonic and prosodic variation of Japanese (J) vowels. Thus, the goal of the present study was to compare the allophonic and prosodic variation in spectral and temporal structure of J and AE vowels, using acoustical analysis of corpora in which the phonetic and prosodic context was varied systematically. To the extent that the type and amount of phonetic variation *differs* across the two languages, we would expect that cross-language perceptual similarity might also vary with contextual variables.

The J vowel inventory includes five spectrally distinctive pairs of vowels that are temporally differentiated (e.g., 1-mora /i/ vs 2-mora /ii/), whereas AE has eleven spectrally distinctive (non-rhotic) vowels: seven phonetically long and four phonetically short vowels (Peterson & Lehiste, 1960). In the present study, vowels were produced in utterances which varied in (a) speaking style (citation-form lists of disyllables vs multisyllabic nonsense words imbedded in a carrier sentence), (b) phonetic context (labial vs alveolar preceding/following consonants), (c) prosodic context (narrow Focus vs Postfocus sentence position) and (d) speaking rate (normal vs rapid sentences). Cross-language comparisons were then conducted to determine how each of these variables influenced the spectral structure (formant frequencies at syllable midpoint) and the temporal structure (syllable duration) of the target syllables. The following comparisons will be presented in this paper:

- 1) Consonantal Context: vowels in b-p vs d-t context produced in multisyllabic nonsense words medially in a sentence produced at normal speaking rate with instructions not to emphasize the target word (no focus).
- 2) Sentence Prominence: vowels in d-t context produced in multisyllabic nonsense words imbedded in sentences in which the target word was in narrow focus position vs following the word with narrow focus (postfocus).
- 3) Speaking Rate: vowels in d-t context produced in multisyllabic words imbedded in sentences produced with the speakers' self-selected normal conversational rate vs at a rapid rate.

Because of limitations of space, the data for the citation-form disyllables will not be presented.

2. Methods

2.1 Subjects

Four Japanese and four Americans participated as speakers in this study. The Japanese participants (2 males, 2 females) were native speakers of Kansai dialect, and ranged in age from 18 to 24 years old (mean=21 years). American participants (2 males, 2 females) were native speakers of New

York dialect and ranged in age from 34 to 41 years old (mean=37 years). All Japanese had never lived outside of the Kansai area in Japan and had not had extensive exposure to spoken English outside of Japan. American participants were not fluent in any other language. All subjects were paid for their participation.

2.2 Stimulus materials

The target vowels for Japanese were the five long vowels, /aa, ee, ii, oo, ∞ / and the five short vowels, /a, e, i, o, ∞ /; for American English the vowels included seven long vowels, /i, e, æ, A, □, o, u/ and four short vowels, /I, ε, ϕ, Y/. For AE, the carrier sentence was 'I said five gaCVCa this time.' For J, the carrier sentence was 'Kono hon-ga CVCa desu ne'.

2.3 Procedure

Instructions given for each condition were as follows: (1) Normal: "Please speak as if you are talking to a native speaker; don't speak slowly and don't emphasize the nonsense word," (2) Focus: "Please speak as if you are answering a question, 'Did you say five gaduta?' when your answer is 'gadita'," (3) Postfocus: "Please speak as if you are answering a question, 'Did you say three gaduta?' when your answer is 'five'," and (4) Rapid: "Please speak as rapidly as possible without leaving any sounds out." Speakers were provided a reading list in which each of the utterances was preceded by a number indicating the target vowel identity. Speakers were instructed to pause briefly between the number and the sentence. Utterances were blocked by phonetic/prosodic condition (normal b-p, normal d-t, Focus d-t, Postfocus d-t, rapid d-t) to establish a consistent rhythm and prosody across utterances containing all vowels within each condition. All vowels were produced four times in each condition with conditions randomized across the four speakers of each language. Within blocks, vowels were randomly ordered in four different sequences. Any clearly mispronounced or disfluent utterances were either self-corrected by the speaker or the experimenter asked for a repetition at the end of each block.

Three J speakers were recorded in an anechoic chamber at ATR Human Information Science Laboratories, using a condenser microphone (SONY ECM-77) connected to a DAT recorder (SONY PCM-2500A, B). One J speaker and all AE speakers were recorded in an IAC chamber in the Speech Acoustic and Perception Laboratory at the City University of New York-Graduate Center, using a dynamic microphone (SHURE SM-48) fed to a Preamplifier (Earthworks Lab 101) and then into a PC computer through a soundcard. Stimuli were directly transferred to computer files (22.05 kHz sampling rate, 16 bit quantization) using "SoundForge 4.5 (Sonic Foundry, inc)" software.

Acoustic analysis was performed on a PC computer, using customized programs written in MATLAB. First, time-synchronized waveforms and wideband spectrograms (1024 points, 6 dB pre-emphasis) of an utterance were displayed. The experimenter located the onset and offset of the target syllable by hand. Onset was defined as the initial consonant release burst (the beginning of opening of the vocal tract) and offset was defined as the beginning of silence associated with the beginning of following stop consonant closure. The program then displayed three spectral sections located at 25%, 50%, and 75% durational points in each syllable in which formant center frequencies (derived from LPC analysis – 24 coefficients autocorrelation) were superimposed as lines on the narrow band FFT spectrum (512 pts). In most cases, the LPC analysis selected formant frequencies corresponded closely with spectral peaks in the FFT spectrum. In those cases where the LPC analysis generated spurious peaks or indicated missing or merged formants, values were estimated from the FFT spectra by manual placement of a cursor at the spectral peak. Data from the 50% point are presented in this paper. Average fundamental frequency (voice pitch) was also calculated (autocorrelation method) for the middle half of each syllable (between 25% and 75% points).

3. Results

Results of acoustic comparisons across the three contextual variables are presented below. Changes in vowel quality (formant frequencies at syllable midpoint) are shown in figures which plot the mean Formant 1/Formant 2 (F1/F2) values (in Barks) averaged over tokens and speakers (16 tokens/vowel); coordinates are arranged to display vowel locations in the traditional vowel quadrilateral with high vowels above, low vowels below and front vowels to the left. Long and short vowels for each language are shown separately. Changes in vowel quantity (syllable duration) are also reported in tables for long and short vowels separately for each language, averaged over tokens and speakers. Finally, changes in fundamental frequency (f0) are reported as a function of the prosodic variables of sentence prominence and speaking rate.

3.1 Effects of Phonetic Context: bVp vs dVt

3.1.1 Vowel Quality: Figure 1 shows the combined data of males and females. As Figure 1 shows, differences in the average F1/F2 values as a function of consonantal context differed across languages and across long and short vowels in each language. J 2-mora vowels reflected the smallest average change overall; F2 values (corresponding to vowel backness) changed only slightly for the back vowels /aa, oo, ∞∞/ (mean F2 change = 109 Hz, 124 Hz, 126 Hz, respectively), while front vowels remained stable. In contrast, J 1-mora back vowels showed large increases in average F2

frequencies (i.e. greater fronting) in alveolar contexts relative to labial contexts ($a=344$ Hz, $o=355$ Hz, $\infty=377$ Hz), and F1 values also decreased for /a/ (raised). Thus, 1-mora and 2-mora back J vowels differed more in spectral structure in alveolar contexts than in labial context and the 1-mora vowels were less differentiated in vowel space in alveolar contexts. The AE vowels showed a somewhat different pattern. For the long vowels, only /u, o/ showed significant fronting (F2 increases of 586 Hz and 210 Hz, respectively), while F2 values for the remaining back vowels /A, ɔ/ and the three front vowels /i, e, ɪ/ were quite stable across phonetic contexts. All four short AE vowels showed increased F2 values in alveolar context relative to labial context, although the average change was greater for back vowels ($Y=489$ Hz; $\varphi=242$ Hz) than for front vowels ($I=155$ Hz; $\varepsilon=170$ Hz). These changes resulted in differences in the *relative* locations of AE vowels in the vowel space, with so-called back rounded vowels /u, Y, o/ much closer to front vowels in alveolar context than in labial context. That is, back rounded high and mid vowels were allophonically produced as front rounded vowels in alveolar context.

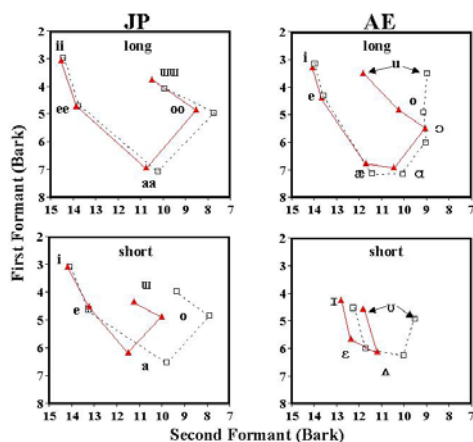


Figure 1. Vowel space for J vowels (left) and AE vowels (right) in /bVp/ (open squares) vs /dVt/ (closed triangles) contexts

3.1.2 Vowel Quantity: Table 1 displays the average syllable durations for J and AE vowels in labial and alveolar contexts. Both long and short vowels in both languages were longer in /dVt/ context than in /bVp/ context. Larger differences in AE may be due to the final /t/ being produced as a (voiced) flap in intervocalic context. Note that the ratio of 2-mora to 1-mora vowels in Japanese remained quite constant across contexts, whereas the long/short vowel ratio for AE vowels was reduced in alveolar context.

Table 1. Average durations (in ms) for long and short Japanese and American English vowels in labial and alveolar consonantal context.

JAPANESE	Labial (LB)	Alveolar (AV)	Ratio (LB/AV)
long vowels (L)	142	159	0.89
short vowels (S)	75	86	0.87
Ratio (L/S)	1.89	1.85	

AMERICAN ENGLISH	Labial (LB)	Alveolar (AV)	Ratio (LB/AV)
long vowels (L)	125	158	0.79
short vowels (S)	88	124	0.71
Ratio (L/S)	1.42	1.27	

3.2 Effects of Sentence Prominence: Focus vs Postfocus

3.2.1 Vowel Quality: Figure 2 displays the F1/F2 vowel spaces for J and AE vowels, again averaged across tokens and speakers. When the nonsense words containing the target vowels were in Focus position, vowels were, on average, more differentiated in spectral structure in both languages. That is, F1 values for low vowels tended to be higher (vowels were lower) and F2 values for front and back vowels differed more. Note, however, that even in the Focus condition, AE high and mid back vowels were allophonically fronted in this alveolar context. Thus, sentence prosody did not, in general, appear to have a large effect on spectral structure for either J or AE vowels, at least in this one consonantal context.

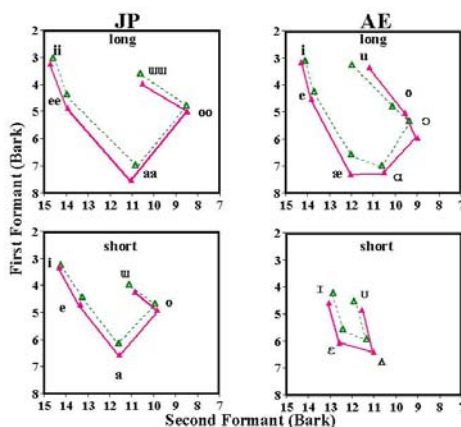


Figure 2. Vowel space for J vowels (left) and AE vowels (right): Focus (solid lines) vs Postfocus (dashed lines) in /dVt/ contexts

3.2.2 Vowel Quantity: Table 2 presents the average durations of long and short J and AE vowels and the ratio of long to short vowels. In the

righthand column, the duration ratios for the Focus relative to the Postfocus condition are also given. As these data show, J short vowels did not show a change with sentence prominence, while J long vowels were 12% longer in the Focus condition. Thus, the ratio of 2-mora to 1-mora vowels was reduced in the Postfocus prosodic condition, relative to the Focus position, and to the Neutral Focus position reported in Table 1 above (1.85 for alveolar context). Syllable durations for nonsense words in Focus position in AE showed more similar lengthening of long and short vowels relative to Postfocus position (17% and 12% respectively) and to Neutral Focus position shown in Table 1 (12% and 6%, respectively). This resulted in the long/short vowel ratio remaining quite stable over all three conditions in which vowels were preceded and followed by alveolar consonants. Note, however, that even in Postfocus contexts, J 2-mora and 1-mora vowels differed significantly more in relative duration than did AE long and short vowels.

Table 2. Average durations (in ms) for long and short Japanese and American English vowels in Focus and Postfocus context.

JAPANESE	Focus (F)	Postfocus (PF)	Ratio (F/PF)
long vowels (L)	164	147	1.12
short vowels (S)	87	86	1.01
Ratio (L/S)	1.89	1.71	

AMERICAN ENGLISH	Focus (F)	Postfocus (PF)	Ratio (F/PF)
long vowels (L)	177	151	1.17
short vowels (S)	132	118	1.12
Ratio (L/S)	1.34	1.28	

3.2.3 Voice Pitch (f₀): For comparison of voice pitch (f₀), the male and female data were separated because of gender differences in average voice pitch. Table 3 presents the average data for J and AE males and females. The f₀ values represent the pitch of the target vowels, averaged over all ten vowels for J utterances and all eleven vowels for AE utterances. For all groups, the average f₀ was greater for syllables in Focus position than in the Postfocus position. However, Japanese males showed somewhat smaller changes in voice pitch than AE males and J and AE females.

Table 3. Average f₀ (in Hz) for Japanese and American English vowels in Focus and Postfocus context.

JAPANESE	Focus (F)	Postfocus (PF)	Ratio (F/PF)
Male	176	150	1.17
Female	265	213	1.24
AMERICAN ENGLISH	Focus (F)	Postfocus (PF)	Ratio (F/PF)
Male	159	105	1.51
Female	237	174	1.36

3.3 Effects of Speaking Rate: Normal vs Rapid

3.3.1 Vowel Quality: Figure 3 presents the vowel spaces for J and AE vowels produced in sentences at the speakers' self-selected normal rate of speech vs at a rapid rate. As the figure shows, the spectral structure of both long and short vowels in both languages remained very similar, on average, in the two rate conditions. There was some centralization of AE /ɪ/, ø/ and J /a, o, oo/ in the rapid condition, while J /aa/ showed the opposite effect (higher F1 in rapid condition). Thus, in alveolar context at least, most J and AE vowels spoken rapidly in sentence length utterances had very similar spectral structure to those produced at slower rates (and more prominent sentence positions) in the same multisyllabic nonsense word context.

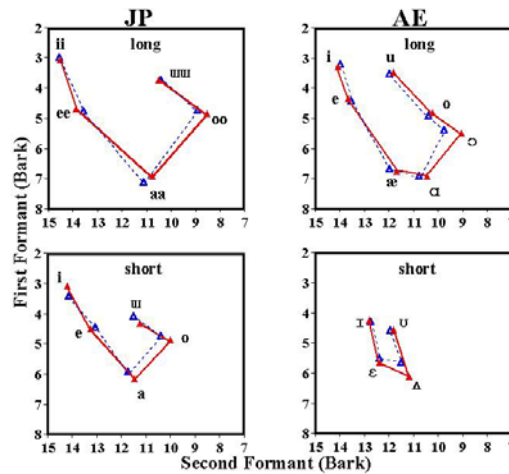


Figure 3. Vowel space for J vowels (left) and AE vowels (right): normal (solid lines) vs rapid (dashed lines) in /dVt/ contexts

3.3.2 Vowel Quantity: Table 4 presents the average duration data for the normal rate sentences (also shown in Table 1) and the rapid rate sentences, separately for long and short vowels in J and AE languages. As before,

durations of syllables containing J 1-mora vowels did not decrease in the rapid condition, relative to those in the normal rate (Neutral Focus) condition, nor indeed relative to those produced in the Focus condition. In contrast, the 2-mora J vowels were, on average, 28% shorter in the rapid condition. Thus, the 2-mora/1-mora duration ratio decreased considerably in the rapid utterances (1.51). In contrast, both long and short AE vowels decreased in duration proportionally (21% and 22%, respectively) such that the long/short duration ratio remained stable across speaking rates. Note that the vowel length differences in rapidly produced J utterances were still greater than they were in AE utterances.

Table 4. Average durations (in ms) for long and short Japanese and American English vowels in normal and rapid context.

JAPANESE	Normal (N)	Rapid (R)	Ratio(N/R)
long vowels (L)	159	124	1.28
short vowels (S)	86	82	1.05
Ratio (L/S)	1.85	1.51	

AMERICAN ENGLISH	Normal (N)	Rapid (R)	Ratio(N/R)
long vowels (L)	158	131	1.21
short vowels (S)	124	102	1.22
Ratio (L/S)	1.27	1.28	

3.3.3 Voice Pitch (f₀): Table 5 gives the average f₀ values in normal and rapid conditions, summed over all ten J vowels and all eleven AE vowels, respectively. There were no consistent differences in f₀ as a function of speaking rate for the Japanese utterances. For AE speakers' utterances, f₀ increased in the rapid condition for both males and females. One J female speaker had a very low f₀ which did not increase for rapid utterances. Future studies with more speakers are needed to explore if there are cross-language differences in f₀ as a function of speaking rate.

Table 5. Average f₀ (in Hz) for Japanese and American English vowels in normal and rapid context.

JAPANESE	Normal (N)	Rapid (R)	Ratio (N/R)
Male	143	146	0.98
Female	253	226	1.12

AMERICAN ENGLISH	Normal (N)	Rapid (R)	Ratio (N/R)
Male	119	138	0.86
Female	180	187	0.96

4. Discussion

4.1 Summary of Effects of Consonantal Context

Both languages showed acoustic fronting of back vowels in /dVt/ context (higher second formants) and vowels were longer in /dVt/ context. The overall shape of the vowel space (i.e. relative locations of vowels) remained similar across contexts for J vowels; however, the space tended to shrink for 1-mora vowels in alveolar context, due to coarticulatory fronting of back vowels and raising of the low vowel. In contrast, the shape of the vowel space changed across contexts for AE vowels, such that high and mid back vowels were fronted relative to mid low and low back vowels in alveolar context. This resulted in considerable acoustic overlap of instances of front and back short vowels in that context. Finally, syllable durations showed greater change for AE than for J vowels. For the latter, the ratio of 2-mora to 1-mora vowels remained relatively constant.

4.2 Summary of Effects of Sentence Focus

The spectral variation of vowels in sentence Focus vs Postfocus context was similar (and relatively small) across languages. In both languages, sentence prominence was also indicated by increases in syllable duration and fundamental frequency. However, in order to maintain the length contrast between 2-mora and 1-mora vowels in Japanese, sentence focus for nonsense words containing J short vowels was achieved by increased pitch instead of increased vocalic duration. For AE vowels, duration differences between long and short vowels were enhanced in Focus context, relative to Neutral and Postfocus context. However, long/short vowel ratios remained considerably smaller for AE vowel contrasts than for J vowel contrasts in both prosodic contexts.

4.3 Summary of Effects of Speaking Rate

Spectral variation as a function of speaking rate was similar across languages, with a few vowels showing a small amount of centralization, while others were either the same or actually more differentiated in rapid condition. As expected, syllable durations decreased for both J and AE rapid utterances. However, in Japanese, only long vowels were shorter, on aver-

age, while short vowels did not change. Thus, the ratio of 2-mora to 1-mora vowels decreased in the rapid condition for J vowels, while in AE utterances, the long-short vowel ratio remained stable. However, even in the context of very rapid speech, vowel length was more distinctive for J vowels than for AE vowels, as would be expected given the phonological status of vowel length in Japanese. Finally, results of changes in voice pitch were not conclusive, given individual differences in patterns of change across speakers.

4.4 Implications

These within- and cross-language similarities and differences in the effects of prosodic and phonetic context on vowel acoustics reflect both language-universal constraints and language-specific “rules” for changes in vowel quality and quantity. Raising awareness of these cross-linguistic differences may help second language learners of Japanese to improve their perception and production.

References

- Flege, J.E. (1995). Second language speech learning: Theory, findings, and problems. In *Speech Perception and Linguistic Experience: Issues in Cross-language Research* (W. Strange, editor), pp.233-277. Timonium, MD: York Press.
- Fourakis, M. (1991). Tempo, stress, and vowel reduction in American English. *Journal of the Acoustical Society of America*, *90*, 1816-1827.
- Hillenbrand, J.M., Clark, M.J., & Nearey, T.M. (2001). Effects of consonant environment on vowel formant patterns, *Journal of the Acoustical Society of America*, *109*, 748-763.
- Nishi, K., Strange, W., Akahane-Yamada, R., Trent, S.A., and Thornton, D.H. (1998). Perceptual assimilation of Japanese vowels by American English listeners: effects of speaking style. Paper presented at the 136th Meeting of the Acoustic Society of America, Norfolk, VA.
- Peterson, G. E., & Lehiste, I. (1960). Duration of syllable nuclei in English, *Journal of the Acoustical Society of America*, *97*, 1286-1296.
- Strange, W., Bohn, O., Trent, S. and Nishi, K. (submitted). Acoustic and perceptual similarity of North German and American English vowels. *Journal of the Acoustical Society of America*.