ACOUSTIC CHARACTERISTICS OF JAPANESE SHORT AND LONG VOWELS: FORMANT DISPLACEMENT EFFECT REVISITED

Kakeru Yazawa and Mariko Kondo

SILS/GSICCS, Waseda University
k-yazawa@aoni.waseda.jp, mkondo@waseda.jp

ABSTRACT

In this study, we conducted a detailed acoustic analysis of Japanese short and long vowels produced in various phonetic contexts by Tokyo Japanese speakers. Japanese long vowels (/ii, ee, aa, oo, uu/) are often assumed to be simply lengthened versions of their short counterparts (/i, e, a, o, u/), but slight differences in formant frequencies have been reported. The present study shows that, while duration is undoubtedly important for distinguishing Japanese vowel length, systematic formant displacement is in fact present in all five long-short pairs. The displacement seems to occur relative to a high central position in the vowel space, which could be described as a 'neutral' position of the tongue. Analysis of formant movements further suggests that the displacement may be caused by speakers' active articulatory control to weaken vowel gestures when there is not enough time for full articulation.

Keywords: Japanese, vowel length, formant displacement, articulation, undershoot

1. INTRODUCTION

The Japanese vowel system consists of five distinct qualities, /i, e, a, o, u/, which form five short (1-mora) and long (2-mora) pairs [19]. [1] Vowel length is contrastive, e.g., /toge/ 'thorn' – /tooge/ 'mountain pass.' The short vowels contrast in both height backness: /i/ is high front [i], /e/ is mid front [e̞], /a/ is low central [ä] and /o/ is mid back [o̞]. /u/ has long been described as high back unrounded [ɯ], but a recent ultrasound study [20] found that it is in fact closer to central rounded [ʉ] or possibly even front rounded [ʏ] (in this paper, we assume that it is high advanced central rounded [ʉ̟]). The vowels thus contrast in lip rounding as well: rounded /o, u/ and unrounded /i, e, a/. Vowel reduction does not occur phonologically.

Traditionally, the long vowels have been treated as lengthened versions of their short counterparts with little or no spectral differences. Many studies on Japanese vowel acoustics have thus focused only on the short vowels [5, 7, 9, 15]. On the other hand, studies discussing the long-short contrast have typically examined how vowel duration is affected by various suprasegmental factors such as pitch accent [16] and speaking rate [10]. However, a few studies have also reported small yet systematic differences in formant frequencies between long and short vowels. One study by Hirata and Tsukada [12] found that all long vowels except /uu/ occupied a more peripheral position in the vowel space than the short vowels. This displacement effect was more pronounced for a faster speaking rate, i.e., when vowel duration was shorter. They also found that formant displacement no longer took place when vowel duration exceeded approximately 200 ms.

To account for the above displacement effect, one might refer to the vowel undershoot model [21], whereby speakers cannot reach the invariant ideal articulatory targets in durationally short vowels because the velocity of articulatory movement is limited. According to this view, phonemically short vowels in Japanese would be undershot because their duration is too short for the articulators to reach the targets. In contrast, articulatory targets are more likely to be reached in phonemically long vowels that allow more time for articulation. Consequently, according to this view long vowels should be more dispersed than short vowels in the vowel space. While this explanation seems viable, Kawahara [13] argues that the model does not fully account for the results of Hirata and Tsukada [12], because a compatible dispersion effect was not found when the speakers produced short vowels at a slower speaking rate. If the speakers were showing undershoot for short vowels, a similar dispersion effect should have been observed in the slower speaking rates because they would then have had enough time to reach the targets. Yet, this prediction was not borne out in the study [12]. Therefore, while the displacement effect is related to durational factors such as vowel length and speaking rate, there has not been a coherent explanation of why it is so.

The current study investigates spectral and temporal characteristics of Japanese short and long vowels, with a primary focus on the formant displacement effect. The current study exploits a variety of phonetic contexts because the results of Hirata and Tsukada [12] might have been confined by their use of only one consonantal context /mVmV/. This study also examines formant movements or vowel inherent spectral change (VISC), which have seldom been studied in the relevant literature (except e.g., [11]), as a potential factor related to the displacement effect.

2. METHODS

2.1. Participants

Sixteen native Japanese speakers (eight male, eight female) participated in the recording. All of them were students or graduates of universities in Japan and were aged between 21 and 30 (mean = 24.1). They had spent most of their lives in Tokyo and surrounding areas.

2.2. Data collection procedures

The recording took place in an anechoic chamber at Waseda University, using a SONY F-780 microphone with a sampling frequency of 44,100 Hz and 16-bit quantization. Participants read disyllabic nonsense words /$C_1V_1C_2V_2$/ in Japanese orthography, presented in isolated and sentence positions. V_1 was the target vowel, namely short /i, e, a, o, u/ and long /ii, ee, aa, oo, uu/. There were five combinations of C_1 and C_2 to manipulate the consonantal contexts: /bV_1p/, /dV_1t/, /gV_1k/, /zV_1s/ and /hV_1d/. V_2 was either /e/ and /o/ to minimize its influence on V_1. An example of a trial is: "*bipe – bipo – bipe to bipo ni wa i ga aru*" ('bipe – bipo – In bipe and bipo there is i'). Participants were instructed to put a pitch accent on the first syllable and to repeat a trial if they did not do so. The data collection procedure yielded 200 tokens per speaker (10 target vowels × 5 consonantal contexts × 2 following vowels × 2 phrasal conditions) giving a total of 3,200 tokens.

2.3. Acoustic analysis

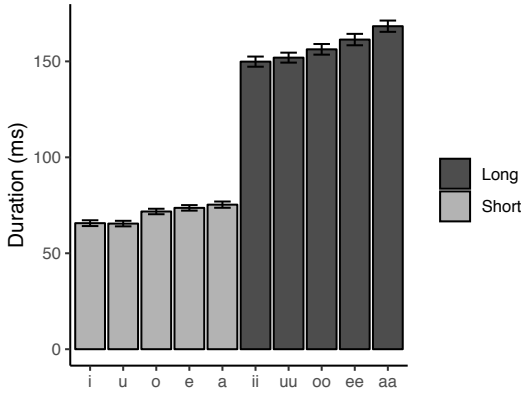
The start and end boundaries of each vowel token were first determined automatically using SPPAS [3], and then manually modified to correspond to the first and last positive zero crossings in Praat [4]. Vowel duration was measured as the time between these boundaries. The first and second formants (F1 and F2) were measured at vowel midpoints using Praat's built-in Burg algorithm. In addition, F1 and F2 values were obtained at 30 equally spaced points from the central portion of each vowel (20% to 80%) to characterize VISC [6].

3. RESULTS

3.1. Duration

Figure 1 shows the average duration of short and long vowels with ±2 standard errors. The long vowels were substantially longer than the short ones, and lower vowels tended to exhibit longer duration. To assess the statistical significance of the trends, a linear mixed effects (LME) model was fitted using *lme4* [2] and *lmerTest* [17] packages in R [22]. The model consisted of *duration* as the dependent variable, *length* ("long" or "short") and *type* ("i," "e," "a," "o," "u") as fixed effects and *participant*, *consonantal context* and *phrasal condition* as random effects. As expected, phonemically long vowels were found to be significantly longer than short ones (β = 87.17, t = 147.45, p < .001), in accordance with the traditional description of contrastive vowel length. The effect of *type* was also significant. When "a" was set as the reference, all the other vowel types were significantly shorter (ps < .001): "e" (β = -4.33), "o" (β = -7.81), "u" (β = -13.13), "i" (β = -14.06). Adding the interaction between *length* and *type* did not improve model fit according to likelihood ratio tests. Thus, vowel duration was affected by vowel height regardless of phonemic length.

Figure 1: Average duration of short and long vowels with ±2 standard errors.



3.2. Midpoint formants

Figure 2 shows the average midpoint F1 and F2 values of the short and long vowels, conditioned by gender. The long vowels occupied more peripheral positions in the vowel space than the short vowels except for /uu/, which is similar to Hirata and Tsukada [12]. To investigate the displacement effect in more detail, a series of LME models were fitted for each acoustic dimension (F1 and F2) and vowel type ("i," "e," "a," "o," "u"). Each model consisted of *formant* (either F1 or F2) as the dependent variable, *length* as the fixed effect (where "short" is the reference) and the same random effects as in the previous model. The models thus assessed the effects of *length* on vowel formants for each vowel type, while considering individual (including gender) and contextual variability.

Table 1 shows the results of the analysis. For example, the first row shows that the intercept (corresponding to mean F1 of short /i/ of all speakers) was 323.68 Hz and that the F1 of long /ii/ was estimated to be 6.85 Hz higher, which is statistically significant (t = 3.17, p < .01). As for F1, long vowels except /oo/ showed significantly larger values than short vowels. For F2, the long vowels were significantly more peripheral than their short

counterparts, apart from /uu/; F2 was higher (more forward) for /ii, ee/ and lower (more back) for /aa, oo/. Formant displacement was thus present in all vowel types in at least one acoustic dimension. Based on the direction and magnitude of the displacement, it can be inferred that the 'neutral' tongue position in Japanese is high and central (somewhere above /uu/-/u/)[2], from which all displacements occur.

Figure 2: Average midpoint F1 and F2 of short and long vowels (gray = male, black = female).

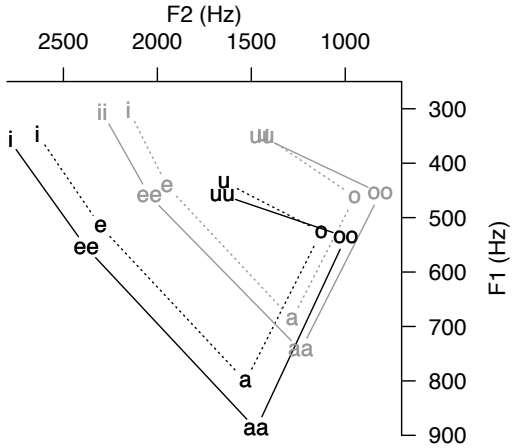


Table 1: Results of LME models run on midpoint F1 and F2 (* $p < .05$, ** $p < .01$, *** $p < .001$). Baseline = short vowels.

		Intercept	β	t
F1	/i/	323.68	6.85	3.17^{**}
	/e/	479.49	28.09	9.29^{***}
	/a/	743.83	72.34	19.71^{***}
	/o/	493.89	1.33	$0.64^{n.s.}$
	/u/	391.40	14.04	4.88^{***}
F2	/i/	2396.81	146.49	19.65^{***}
	/e/	2124.66	85.65	10.70^{***}
	/a/	1406.71	-51.08	-7.37^{***}
	/o/	1038.15	-133.46	-18.39^{***}
	/u/	1539.56	7.89	$0.78^{n.s.}$

3.3. Vowel inherent spectral change (VISC)

In order to quantify VISC, each set of the 30 measured formant values were transformed using discrete cosine transform (DCT). A DCT expresses a sequence of data points in terms of the sum of cosine functions oscillating at different frequencies. In the case of a formant trajectory, the zeroth coefficient corresponds to a straight line proportionate to the mean, the first coefficient to half a cosine (i.e., left half of a "U" shape) representing the overall direction and magnitude of change from the mean, and higher-order coefficients to fine-grained information such as trajectory curvature [18]. The following analysis focuses on the first coefficients, in which positive values represent decreasing movement and negative values represent increasing movement. The size of the values in absolute terms represents the magnitude of the movement.

A set of LME models were fitted for each acoustic dimension and vowel type in the same way as in Section 3.2, except that the dependent variable was *first DCT coefficient* for either F1 or F2. The results are presented in Table 2. The intercepts (corresponding to mean first DCT coefficients for short vowels) suggest that F1 tended to decrease for all types of short vowels. The magnitude of movement was in the order of /a/ > /o/ > /e/ > /u/ > /i/, and thus the lower vowels seem to exhibit larger upward movements. However, the effect was hindered when the vowel was phonemically long, sometimes changing the direction of movement from decreasing (i.e., closing) to increasing (opening). Turning to F2, the intercepts suggest that F2 tended to decrease for short /i, e/, but increase for short /a, o, u/. In other words, short vowels seemed to show backward movements if they were front, and forward movements if they were central or back. F2 movements for long vowels were somewhat nuanced, but there was a tendency for the forward and backward movements to be hindered or at least lessened. This was particularly evident for /ii/, where the coefficient was estimated to be very low and thus its F2 was increasing as opposed to short /i/. In sum, the results suggest that short vowels generally show converging movements toward a high and central position, which adds to the findings in Section 3.2. Not only are short vowels less peripheral than long vowels, they also move back towards the neutral position sooner.

Table 2: Results of LME models run on first DCT coefficients (* $p < .05$, ** $p < .01$, *** $p < .001$). Baseline = short vowels.

		Intercept	β	t
F1	/i/	0.97	-6.26	-5.32^{***}
	/e/	6.02	-12.80	-8.45^{***}
	/a/	19.09	-12.20	-5.71^{***}
	/o/	7.35	-9.46	-5.96^{***}
	/u/	2.88	-9.36	-6.88^{***}
F2	/i/	2.31	-18.58	-5.34^{***}
	/e/	16.00	-2.20	$-0.57^{n.s.}$
	/a/	-6.06	6.44	2.03^{*}
	/o/	-15.22	4.03	$1.20^{n.s.}$
	/u/	-11.36	5.78	$1.66^{n.s.}$

4. DISCUSSION

The detailed acoustic analysis of Japanese short and long vowels confirmed the following results: (1) lower vowels showed longer duration regardless of phonemic length; (2) formant displacement was present for all five long-short pairs, where long vowels were further away from the default position (i.e., more peripheral) than short vowels; (3) short vowels showed conversing movements toward the high central position, while long vowels resisted such movements.

All these findings are considered to have an articulatory basis. For example, Kawahara, Erickson and Suemitsu [14] used electromagnetic articulography (EMA) to investigate jaw movements in the production of the five Japanese short vowels, and found that duration and F1 were reliable acoustic correlates of the degree of jaw opening. Based on this, (1) can be attributed to lower vowels involving larger jaw movement and therefore being more time-consuming. (2) can be described as the articulatory displacement of the tongue relative to its 'neutral' position which is high and central in Japanese. Our results are slightly different from those of Hirata and Tsukada [12] who did not find a clear displacement effect for /uu/-/u/. While this could be due to the lack of different phonetic contexts in their study, it is also possible that the proximity of /uu/ and /u/ to the supposed neutral position obscured the displacement effect. As for (3), the converging movements in short vowels suggest that their gestural targets are close to the supposed neutral position. In contrast, the lower convergence exhibited by the long vowels suggest that they have more peripheral targets that are sustained. The findings (2) and (3) are related; the displacement of midpoint formants may in fact be caused by the differences in formant movements.

The original undershoot model [21] would be able to account for (1) and (2), but not (3). The model is compatible with (1) because farther articulatory targets (low > mid > high) would require more time until they are reached. The model would also explain (2) because long vowels allow more time for the articulators to reach the targets, whereas short vowels would show undershoot and therefore would be closer to the starting point, i.e., the neutral position. However, the model does not explain (3). If speakers cannot reach the targets simply because the speed of the articulators is limited, then short vowels would still show movement toward the targets (i.e, away from the neutral position), even if the magnitude of the movements would be smaller than that of long vowels. However, our result contrarily showed a consistent tendency for short vowels to move back to the neutral position in the middle of articulation. A possible reason for this is that speakers actively 'weaken' vowel gestures when there is not enough time for full articulation. Under this situation, short and long vowels would have different spatial and temporal targets: spatially, short vowel targets would be closer to the neutral position than long vowel targets; temporally, long vowel targets would be persistent while short vowel targets are not. This explanation could also account for the effect of speaking rates on formant displacement found in Hirata and Tsukada [12]. For example, a faster speaking rate would elicit more displacement because speakers would have to reduce each articulatory gesture (as there is not enough time for proper articulation) unless the intrinsic segmental duration is long enough to allow full articulation. The displacement effect would disappear at a slow speaking rate when speakers would expect enough time to fully articulate each gesture.

Finally, it is worth noting that Japanese listeners do not seem to refer to vowel type when judging whether a vowel is phonemically short or long [1]. Since vowel duration is systematically different at different heights (low > mid > high), in theory listeners could use vowel type as a secondary cue for identifying length, but they seem not to do so. This indicates that listeners have an invariant [+long] feature that is independent of vowel type, and thus our result (1) is most likely driven by articulatory constraints. Future research could test whether (2) formant displacement and (3) VISC affect the perception of vowel length in Japanese.

5. ACKNOWLEDGMENTS

This study was supported by JSPS Grant-in-Aid for Scientific Research (B) No. 15H02729 and Grant-in-Aid for JSPS Fellows No. 18J11517.

6. APPENDIX

Table 3: Average midpoint F1 and F2 as well as duration of Japanese short and long vowels.

	Male			Female		
	F1	F2	dur	F1	F2	dur
/i/	301	2154	68	346	2639	64
/e/	443	1947	76	516	2302	71
/a/	687	1283	78	801	1530	73
/o/	462	949	74	526	1127	70
/u/	348	1435	68	434	1645	63
/ii/	306	2293	147	355	2794	153
/ee/	460	2043	156	555	2378	166
/aa/	744	1237	166	889	1474	171
/oo/	455	813	153	535	996	159
/uu/	352	1442	150	459	1653	154

7. REFERENCES

[1] Arai, T., Behne, D. M., Czigler, P., Sullivan K. P. H. 1999. Perceptual cues to vowel quantity: Evidence from Swedish and Japanese. *Proc. Fonetik* 81, 8–11.

[2] Bates, D., Mächler, M., Bolker, B. M., Walker, S. C. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software* 67(1), 1–48.

[3] Bigi, B. (2015). SPPAS - Multi-lingual approaches to the automatic annotation of speech. *The Phonetician* 111–112, 54–69.

[4] Boersma, P., Weenink, D. 2018. Praat: doing phonetics by computer (version 6.0.43) http://www.praat.org/

[5] Chiba, T and Kajiyama. M. (1941). *The Vowel: Its Nature and Structure*. Tokyo: Tokyo-Kaiseikan.

[6] Elvin, J., Williams, D., Escudero, P. 2016. Dynamic acoustic properties of monophthongs and diphthongs in Western Sydney Australian English. *J. Acoust. Soc. Am.* 140(1), 576–581.

[7] Fujisaki, H., Sugito, M. 1977. Onsei no butsuriteki seisitsu [Physical characteristics of speech]. In: *Iwanami Koza Nihongo 5 On'in*, 63-106.

[8] Gick, B., Wilson, K., Koch, K., Cook, C. 2004. Language-specific articulatory settings: Evidence from inter-utterance rest position. Phonetica 61(4), 220–233

[9] Hirahara, T., Akahane-Yamada, R. 2004. Acoustic characteristics of Japanese vowels. *Proc. 18th ICA*, 3387–3290.

[10] Hirata, Y. 2004. Effects of speaking rate on the vowel length distinction in Japanese. *J. Phon*, 32(4), 565–589.

[11] Hirata, Y., Tsukada, K. 2004. The effects of speaking rates and vowel length on formant movements in Japanese. *Proc. Texas Linguistics Society Conference 2003*, 73–85.

[12] Hirata, Y., Tsukada, K. 2009. Effects of speaking rate and vowel length on formant frequency displacement in Japanese. *Phonetica* 66(3), 129–149.

[13] Kawahara, S. 2016. Japanese has syllables: A reply to Labrune. *Phonology* 33, 169–194.

[14] Kawahara, S., Erickson, D., Suemitsu, A. 2017. The phonetics of jaw displacement in Japanese vowels. *Acoustical Science and Technology* 38(2), 99–107.

[15] Keating, P. A., Huffman, M. K. 1984. Vowel variation in Japanese. *Phonetica* 41(4), 191–207.

[16] Kozasa, T. 2004. The interaction of duration and pitch in Japanese long vowels. *Proc. Barkley Linguistics Society 30*, pp. 211–222.

[17] Kuznetsova, A., Brockhoff, P. B., Christensen, R. H. B. 2017. lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software* 82(13), 1–26.

[18] Morrison, G. S. 2013. Theories of vowel inherent spectral change. In: Morrison, G. S., Assmann, P. F. (eds), *Vowel Inherent Spectral Change*. Berlin: Springer, 31–47.

[19] Nishi, K., Strange, W., Akahane-Yamada, R., Kubo, R., Trent-Brown, S. A. 2008. Acoustic and perceptual similarity of Japanese and American English vowels. *J. Acoust. Soc. Am.* 124(1), 576–588.

[20] Nogita, A., Yamane, N., Bird, S. 2013. The Japanese unrounded back vowel /ɯ/ is in fact unrounded central/front [ɯ̟ - ʏ]. *Proc. Ultrafest VI*, 39–42.

[21] Lindblom, B. 1963. Spectrographic study of vowel reduction. *J. Acoust. Soc. Am.* 35(11), 1773–1781.

[22] R Core Team. 2018. R: A language and environment for statistical computing (version 3.5.1) https://www.R-project.org/.

[1] Long vowels are phonologically considered as a sequence of identical vowels, i.e., /ii, ee, aa, oo, uu/, which are phonetically realized as [iː, eː, aː, oː, ɯ(ɯ̟)ː].

[2] Some studies suggest that the tongue rest position is language-specific [8].