A corpus-based computational analysis of high-front and -back vowel production of L1-Japanese learners of English and L1-English speakers

Martin Schweinberger, Yuki Komiya

The University of Queensland

m.schweinberger@uq.edu.au, y.komiya@uq.net.au

Abstract

This study combines acoustic phonetics with computational and applied corpus linguistics to analyze and compare the production of the monophthongal vowels /ɪ/, /iː/, /ʊ/, and /uː/ in the speech of 148 L1-Japanese learners (JPN) and 107 L1-speakers of English (ENS) based on *The International Corpus Network of Asian Learners of English* (ICNALE). The study aims to ascertain if JPN merge spectrally close vowels by calculating Bhattacharya coefficients. In addition, the study uses mixed-effects linear regression to determine if JPN compensate for the potential mergers by exaggerating durational contrasts between spectrally similar vowels. The results of the analysis confirm that JPN exhibit high degrees of overlap for both /ɪ iː/ and /ʊ uː/. Their L1-English peers, however, also exhibit substantive overlap for /ʊ uː/. With respect to duration, the analysis shows that JPN extend the duration of all vowels and exaggerate the difference between /ɪ iː/ and /ʊ uː/ to compensate for the lack of qualitative differences between short and long vowel pairs. This study represents the first corpus-based acoustic analysis of JPN vowels in spontaneous speech.

Index Terms: acoustic phonetics, vowel production, learner corpus research, Japanese learners of English

1. Introduction

While pronunciation poses a challenge for language learners, it is also the most immediate and direct display of linguistic proficiency. Listeners automatically and subconsciously categorize and infer judgments about speakers based on pronunciation [2]. In addition, pronunciation is crucial for intelligibility and is affecting real-life opportunities (jobs, partner choice, etc.).

A underlying cause for the difficulties that learners face is that languages are not independent but interact in the minds of multilingual speakers [2] which means that the L2 sound system is affected by the L1 system (and vice versa). From the perspective of JPN, English vowels are particularly challenging [3] due to

- Differences in inventory size (Japanese: 5 monophthongal vowels vs. English: app. 11 monophthongal vowels (depending on the variety of English)) [4]
- Differences in how vowels are differentiated (Japanese: duration differences versus English: formant *and* duration differences)

Formants are concentration of acoustic energy at a certain frequency [5] with the first formant (F1) and the second formants (F2) of a vowel sound inversely corresponding to the tongue height and tongue fronting during vowel production. Regarding the production of English vowels produced by JPN, it has been shown that JPN merge spectrally similar vowels (including high-front and -back vowels) [6]. Furthermore, it has been reported that JPN are very sensitive to vowel duration [7] and exaggerate duration to compensate for the relative insensitivity to formant differences [8].

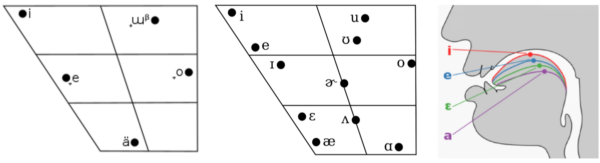

Figure 1: *left panel: vowel chart showing monophongal vowels of standard Japanese; center panel: vowel chart of Southern Californian American English; right panel: tongue position corresponding to selected front vowels.*

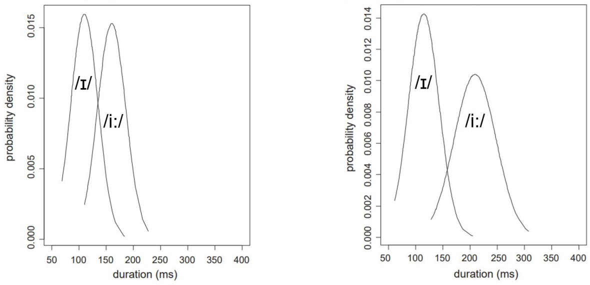

Figure 2: *left panel: durations of high front vowels produced by ENS; right panel: durations of high front vowels produced by JPN.*

Previous research on vowel production by JPN is predominantly based on read-aloud word lists or selected scripted sentences in highly controlled laboratory conditions. Hence, characteristics of the vowel production of learners in naturalistic speech environments remain largely unknown. Furthermore, previous research has relied on small samples of subjects with studies using between 8 and 15 subjects. As such, the findings provided by previous research may not warrant generalization to larger speech communities or to conversational language production in natural settings.

The present study addresses these issues and aims to provide a more detailed understanding of the following research questions:

1. Do JPN merge /iː/ and /ɪ/ as well as /uː/ and /ʊ/?
2. Do JPN exaggerate the lengths of vowels to compensate for a lack of spectral differentiation?

2. Corpus Data

The study uses data from the *International Corpus Network of Asian Learners of English* (ICNALE) [9]. The ICNALE is one of the largest publicly available multimodal learner corpora comprising more than 10,000 topic-controlled speeches and essays produced by college students in ten countries and regions in Asia as well as English native speakers. For this study, all data representing spoken monologues (spontaneous speech) collected between 2017 and 2019 from 148 JPN and 107 ENS were analyzed (this encompasses all JPN and ENS speakers in the data that produced relevant tokens and, in the case of ENS, spoke American English).

Every speaker contributed two one-minute recording to the spoken monologues' component of ICNALE. Speech samples were recorded on mobile devices or personal computers using in-built microphones resulting in a highly variable quality of recordings.

3. Data Processing

Data processing started with using Web-MAUS [10] to (force) align the audio files and transcriptions provided by ICNALE into Praat TextGrids (the forced alignment used both US and British models). All subsequent steps of the analysis were performed using R Version 4.2 [11] in RStudio [12].

The first to third formants of all vowels as well as vowel durations were extracted using rPraat [13], wrassp [14], and tidyverse [15]. The algorithm targeted a range between three and 7 formants for each vowel resulting in five formant values for each of the first to third formants in each vowel. The optimal formant values out of these five options were determined based on the minimal Euclidean distance to standard American English vowel formants based on [16] and standard southern British English based on [17]. Based on the Euclidean distance, alternative options and the less well fitting target variety were removed from the analysis so that the data set contained only one observation for each vowel (see Table 1)

Table 1: *Overview of the semi-processed data set.*

Type	Speakers	/ɪ/	/iː/	/ʊ/	/uː/	Total
ENS	132	1,077	1,182	348	513	3,120
JPN	149	2,779	1,084	608	623	5,094
Total	281	3,856	2,266	956	1,136	8,214

In a next step, socio-demographic information about the speakers (speaker type, age, gender, English proficiency) was added to the data and ENS not from North America were removed. All further analyses continued with standard American English as target variety. Table 2 provides an overview of the socio-demographics of the speakers in the data and Table 3 shows the proficiency levels among the L1 Japanese learners.

Next, vowels were normalized using a z-transformation after grouping the data by speaker type (ENS vs JPN) and gender. After this normalization procedure, all vowels not representing /iː/, /ɪ/, /uː/, and /ʊ/ were removed from the analysis.

Then, multi-syllabic words or words containing more than 9 characters were removed to better control for variability caused by the phonetic and phonological environments in which the vowels were produced. Only words were retained which had a /CV(C)/ syllable structure (e.g., *get*, *gut*, *hit*, *shit*, *due*, *we*, *see*).

To account for the low quality of audio recordings and to remove outliers and inaccuracies, kernel density estimation was applied to the z-transformed first and second formats. All vowels having density values in the lower quartile of first and second formats were removed. The final data set is summarized in Table 4.

Table 2: *Overview of information about the speakers represented in the data.*

Type	Gender	18-29	30-39	40-49	50+	Total
ENS	female	25	5	1	4	35
	male	30	27	9	6	72
JPN	female	63	1	0	0	64
	male	82	1	1	0	84
Total		200	34	11	10	255

Table 3: *Overview of proficiency levels among JPN speakers in the data.*

Gender	A2	B1	B2	Total
female	14	38	12	64
male	16	50	18	84
Total	30	88	30	148

Table 4: *Overview of the final data set (percentage of retained observations compared to semi-processed data in brackets).*

Type	Speakers	/ɪ/	/iː/	/ʊ/	/uː/	Total
ENS	107	560	785	159	311	1,815
	(81.0)	(52.0)	(66.4)	(45.7)	(60.6)	(58.2)
JPN	148	917	481	155	231	1,721
	(99.3)	(33.0)	(44.4)	(25.5)	(37.1)	(33.6)
Total	255	1,477	1,203	314	542	3,536
	(90.7)	(38.3)	(53.1)	(32.8)	(47.7)	(43.0)

4. Statistical Analysis

The statistical analysis made use of two procedures

- Bhattacharya coefficients: to assess potential spectral mergers of /iː/ and /ɪ/ as well as /uː/ and /ʊ/
- Mixed-effects linear regression: to assess if JPN exaggerate the length of vowels to compensate for a potential lack of spectral differentiation

Bhattacharya coefficients are suited to assess vowel mergers as this coefficient represents a measure of overlap of scatter clouds with 1 representing perfect overlap and 0 representing zero overlap.

Mixed-effects linear regression modeling was performed using the lme4 [18] and the sjPlot package [19] with a step-wise step-up model fitting procedure. The regression analysis evaluated the effect of the following variables and their two-way interactions. If models exhibited substantial multicollinearity (variance inflation factors ≤ 5, the model was considered not trustworthy). Table 5 details the variables that were tested during the statistical modeling and provides information about their scaling as well as how they were operationalized.

Proficiency could not be included into the model and used as a predictor as no proficiency information was available for ENS. Including proficiency as a predictor would have led to the

Table 5: *Variables included in the mixed-effects regression modelling (ran. eff. = random effect, ind. var. = independent variable, cat. = categorical scaling, nom. = nominal scaling).*

Variable	Type	Scale	Levels / Description
duration	dep. var.	num.	duration of vowel in ms
speaker	ran. eff.	cat.	id of speaker
word	ran. eff.	cat.	word, e.g., *good*, *foot*,*he*, *hit*
vowel	ind. var.	cat.	vowel sound: /iː/, /ɪ/, /uː/, /ʊ/
type	ind. var.	nom.	vowel produced by ENS or JPN
status	ind. var.	nom.	status of word (gramm. vs lex.)
gender	ind. var.	nom.	female vs. male (self-reported)
age	ind. var.	num.	age of speaker in years

automatic exclusion of all ENS data points or it would have resulted in ill-fit models due to the absence of variability in proficiency among ENS.

5. Results

The following reports on the findings of the analysis separated by research question the statistical analyses have addressed.

5.1. Vowel mergers

Regarding the potential merger between /iː/ and /ɪ/ among JPN and ENS speaker, the high values of the Bhattacharya coefficient confirmed the expected merger of spectrally similar high-front vowels among JPN but not among ENS:

- JPN Bhattacharya coefficient (/iː/, /ɪ/): .870
- ENS Bhattacharya coefficient (/iː/, /ɪ/): .668

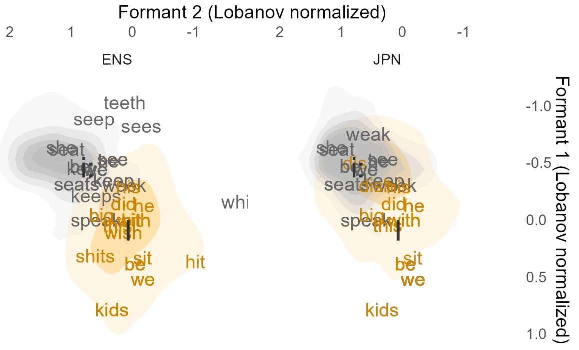

Figure 3: *left panel: overlap of high-front vowels among JPN; right panel: overlap of high-front vowels among ENS.*

Regarding the potential merger between /uː/ and /ʊ/ among JPN and ENS speaker, the high values of the Bhattacharya coefficient confirmed the expected merger of spectrally similar high-back vowels for JPN but also confirmed a merger of spectrally similar high-back vowels among ENS:

- JPN Bhattacharya coefficient (/uː/, /ʊ/): .881
- ENS Bhattacharya coefficient (/uː/, /ʊ/): .929

Thus, the analysis confirmed the expected mergers of spectrally similar high-front and -back vowels among JPN but also a merger of high-back vowels among ENS.

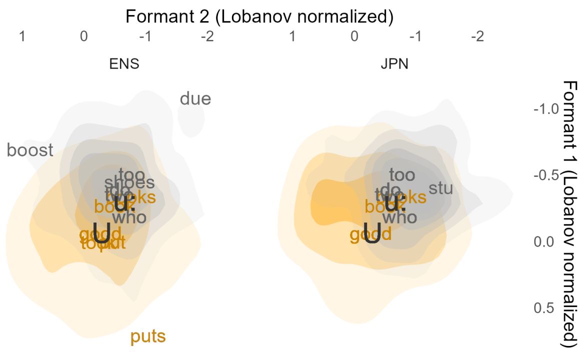

Figure 4: *left panel: overlap of high-back vowels among JPN; right panel: overlap of high-back vowels among ENS.*

5.2. Vowel duration

The regression modeling arrived at a final minimal adequate model with a notable explanatory capacity accounting for 17.9 percent of the overall variability in the data. The data confirmed speaker type, vowel, and gender as significant main effects. More importantly, as this directly addresses the second research question and in addition to an interaction between gender and speaker type, the model reports significant interactions between speaker type and vowels. These latter interactions confirm that JPN exaggerate all vowel durations compared with the vowel durations of ENS.

Table 6: *Results of the mixed-effects regression modelling.*

	Final minimal adequate model		
Predictors	Est.	CI	p
(Intercept)	0.06	0.06 - 0.07	<0.001
type$_{JPN}$	0.03	0.02, 0.03	<0.001
label$_{i:}$	0.01	0.01, 0.02	<0.001
label$_{U}$	-0.00	-0.02, 0.01	0.587
label$_{U:}$	0.03	0.02, 0.04	<0.001
gender$_{male}$	0.01	0.00, 0.01	0.117
type$_{JPN}$ * label$_{i:}$	0.01	0.01, 0.02	0.002
type$_{JPN}$ * label$_{U}$	0.01	-0.01, 0.02	0.220
type$_{JPN}$ * label$_{U:}$	0.02	0.01, 0.03	<0.001
type$_{JPN}$ * gender$_{male}$	-0.02	-0.03, -0.01	0.003
Random effects			
ICC		0.08	
N		255 speakers 44 words	
Observations		3536	
Mar. R^2 / Cond. R^2		0.112 / 0.179	

The results show that JPN extend or exaggerate all high-front and back vowel durations and not just long vowels. The effect plot shown in Figure 5 furthermore shows that JPN exaggerate the duration difference of both /iː/ and /ɪ/ as well as /uː/ and /ʊ/. This is confirmed by Figure 6 as the differences between durations are notably higher for both vowel pairs among JPN compared with ENS.

The results of the statistical analysis thus confirm that JPN merge spectrally similar vowels and appear to compensate the lack of a differentiation by exaggerating both the durations of high-front and -back vowels and by exaggerating the durational

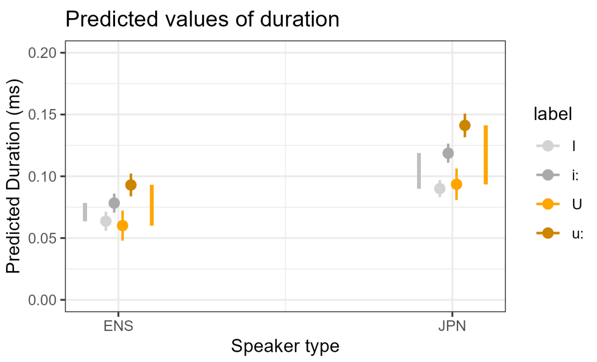

Figure 5: *Predicted duration values based on the final minimal adequate model by vowel and speaker type. (Gray and orange lines show the difference between short and long vowel durations within speaker groups)*

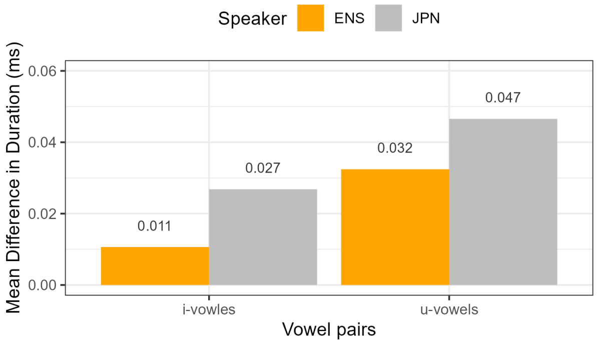

Figure 6: *Duration differences by speaker type (orange: ENS, gray: JPN) and vowel pair. The bars show the difference in duration between /iː/ and /ɪ/ (left) and between /uː/ and /ʊ/ (right).*

differences between long and short vowels (see Figure 6).

6. Discussion

The findings presented here confirm previous research which reported the tendency to merge spectrally close vowels produced by JPN in lab settings (see [20], [21]). Also in alignment with previous research [22] is that JPN exaggerated durational contrasts between spectrally similar vowels which, again, had been reported for JPN in lab settings.

The findings presented here also offer unique insights in that the study extends previous research to natural settings and substantially expands the empirical basis of existing research. In addition, the data analyzed here suggests a merger of high-back vowels (/uː/ and /ʊ/) among ENS in spontaneous speech.

One noteworthy limitation of the present study relates to the variable quality of the recordings which can be considered not only substandard but relatively poor for at least a subsection of the minute-long recordings that make up the spoken monologue component of the ICNALE. The reason for the low quality of the audio recordings is that the audio data were recorded predominantly using in-built microphones of mobile devices. While the quality of the audio data could, at least in part, be compensated using statistical procedures (kernel density estimation) which reduced the existing noise to a certain extent, such means are ultimately limited and unfit to fully remedy data quality. Another limitation consists in the fact that, given the variability and distributional characteristics of spontaneous speech, it is difficult to control the semantic and phonological environments of vowels, which, however, affect vowel production and thus formant values [23].

The advantages of the present study are that it has produced insights into vowel production by JPN in spontaneous speech which is under-explored even in learner corpus research. Also, the study is among the first to study JPN vowel production in natural settings which allows to generalize findings to real-life learner speech. Finally, and despite its limitations, the fact that the poor quality of the data could be compensated using advanced methods enables to extend the methods presented here to further automated corpus-based investigation on larger and more diverse samples of learner speech.

7. Conclusions

The present study represents one of the first large-scale, corpus-based studies of ESL vowel production in natural speech extending previous research by substantively extending the data base in terms of both the number of speakers and observations. In addition, the application of kernel density estimation to mitigate low quality audio data is promising but requires additional investigation and comparisons against gold standard data sets as well as manually annotated data to determine to what this method can compensate for poor data quality. Potential follow-up studies could zoom in on perception and intelligibility to investigate the auditory and cognitive implications of the acoustic effects presented here. Finally, the present study can be a prototype that can easily be extended to other learner varieties and multi-modal data sources.

8. References

[1] Gilakjani, A. P., and Ahmadi, M. R., "Why is pronunciation so difficult to learn?", English Language Teaching, 4(3), 74-83, 2011.

[2] Flege, J. E., "Second-language speech learning: theory, findings, and problems", in W. Strange [Ed.], Speech perception and linguistic experience: Issues in cross-linguistic research, 233-277, York Press, 1995.

[3] Franklin, A. D. and Stoel-Gammon, C., "Using multiple measures to document change in English vowels produced by Japanese, Korean, and Spanish speakers: The case for goodness and intelligibility", American Journal of Speech-Language Pathology, 23(4), 625-640, 2014.

[4] Homma, Y., "Acoustic phonetics in English and Japanese", Yamaguchi Shoten, 1992.

[5] Ladefoged, P., and Johnson, K., "A Course in Phonetics", Cengage, 2014.

[6] Ingram, J. C. L., and Park, S. G., "Cross-language vowel perception and production by Japanese and Korean learners of English", Journal of Phonetics, 25(3), 343-370, 1997.

[7] Kato, H., Tajima, K., Akahane-Yamada, R., "Native and non-native perception of phonemic length contrasts in Japanese", The Journal of the Acoustical Society of America, 110(5), 2686, 2001.

[8] Morrison, G. S., "Japanese listeners' use of duration cues in the identification of English high front vowels", in Larson, J. and Paster, M. [Eds.], Proceedings of the 28th annual meeting of the Berkeley Linguistics Society, 189–200, Berkeley Linguistics Society, 2002.

[9] Ishikawa, S., "Design of the ICNALE Spoken: A new database for multi-modal contrastive interlanguage analysis", Learner Corpus Studies in Asia and the World, 2, 63-76, 2014.

[10] Kisler, Th., Reichel, U. D. and Schiel, F. “Multilingual processing of speech via web services”, Computer Speech and Language 45: 326–347, 2017.

[11] R Core Team, “R: A language and environment for statistical computing”, R Foundation for Statistical Computing, Vienna, Austria. (https://www.R-project.org), 2022.

[12] RStudio Team, “RStudio: Integrated Development Environment for R”, RStudio PBC, Boston, MA (http://www.rstudio.com), 2022.

[13] Bořil, T., Skarnitzl, R., “Tools rPraat and mPraat”, in P. Sojka, A. Horák, I. Kopeček, and K. Pala [Eds.], Text, Speech, and Dialogue, 367-374, Springer International Publishing, 2016.

[14] Bombien, L., Winkelmann, R., and Scheffers, M., “wrassp: an R wrapper to the ASSP Library”, R package version 1.0.1, 2021.

[15] Wickham, H., et al., “Welcome to the tidyverse”, Journal of Open Source Software, 4, 43, 1686, 2019.

[16] Yang, B., “A comparative study of American English and Korean vowels produced by male and female speakers”, Journal of Phonetics, 24, 245–261, 1996.

[17] Deterding, D., “The Formants of monophthong vowels in standard southern British English pronunciation”, Journal of the International Phonetic Association, 27, 1-2, 47-55, 2009.

[18] Bates, D., Maechler, M., Bolker, B., and Walker, S., “Fitting linear mixed-effects models using lme4”,Journal of Statistical Software, 67, 1, 1-48, 2015.

[19] Lüdecke, D., “sjPlot: Data visualization for statistics in social science”, R package version 2.8.10, 2021.

[20] Ueyama, M., “Duration and quality in the production of the vowel length contrast in L2 English and L2 Japanese”, in M. J. Solé., D. Recasens., and J. Romero. [Eds.], 15th International Congress of Phonetic Sciences, 1509-1512, Universitat Autònoma de Barcelona, 2003.

[21] Tsukada, K., “Native vs non-native production of English vowels in spontaneous speech: An acoustic phonetic study”, in P. Dalsgaard, B. Lindberg, and H. Benner [Eds.], Proceedings of the 7th European Conference on Speech Communication and Technology (Eurospeech 2001), 305-308, 2001.

[22] Tsukada, K., “Durational characteristics of English vowels produced by Japanese and Thai second language (L2) learners”, Australian Journal of Linguistics, 29(2), 287-299, 2009.

[23] Visceglia, T., Chiu-Yu, T., Kondo, M., Meng, H., and Sagisaka, Y., “Phonetic aspects of content design in AESOP (Asian English Speech cOrpus Project)”, Oriental COCOSDA International Conference on Speech Database and Assessments, 60-65, 2009.