A Study on the Differences in Vowel Duration in Japanese in Natural and Written Speech: An Empirical Analysis Based on a Corpus

Peng Liu
Foreign Studies School, Liaoning University Of International Business and Economics, Dalian 116052, Liaoning, China
Corresponding author: Peng Liu (e-mail: 13516051062@163.com)

ABSTRACT Japanese vowel duration is a key phonetic feature for distinguishing lexical meaning and stylistic register, but its realization patterns across different speech contexts remain insufficiently quantified. This study constructs a bimodal corpus involving 20 native Japanese speakers, including 50 minutes of spontaneous conversation and 30 minutes of news text reading. Using Praat software, duration data for the five Japanese vowels /a, i, u, e, o/ are annotated, producing 8,347 valid samples. The analysis first extracts absolute vowel duration within syllable boundaries, then calculates the long-to-short ratio, applies mixed-effects modeling to control individual speaker differences, and uses independent-sample t-tests to verify inter-group significance. The results show that word-final vowels in conversational contexts are shortened by 28.4% compared with written-language reading, with /u/ showing the most significant shortening rate of 31.7%. Prosodic boundary location and speech-rate variation are the main factors affecting duration differences. The findings provide acoustic signal evidence for natural speech training, Japanese prosody instruction, and prosodic modeling in speech synthesis systems.

INDEX TERMS vowel duration, corpus analysis, japanese phonetics, prosodic modeling, acoustic signal processing

I. INTRODUCTION

THE duration of the Japanese vowel as a significant method of contrasting phonemes is directly connected to the semantic distinguishing role of the words. The research on traditional phonetics is largely concerned with the pronunciation of individual words in a laboratory setting, where ideal vowel duration values are observed under strictly controlled laboratory conditions. Nevertheless, the pronunciation strategy of the speaker is limited by various factors in the actual communicative situations including speech rate, emotion, and formality of the topic, which leads to the complicated patterns of variation of the duration of the vowels. Natural conversation and written reading are two common extremes of spoken language production. The former is concerned with communicative efficiency whereas the latter is concerned with phonetic standardization. It is still unclear as to whether the systematic disparities in the vowel duration realization mechanisms under the two styles are quantitatively verified by the use of big volume of corpora [1,2].

The current literature on Japanese vowel duration is mostly concerned with the differences at the level of a single linguistic context, including the aspects of moderation of the stress position in duration and coarticulation of the consonant contexts. Although the cross-speech comparative analysis has been examined by some scholars, the sample is low and there is no single measuring standard. Others employ reading materials to replicate the situation of conversation, but the pre-determined texts are not able to recreate the dynamic aspects of prosodic pauses and turn taking in natural communication. Even though the corpus-based research method has seen advancement in studying the prosodic features of other languages like English and Chinese, its use in studying the Japanese vowel duration has not been implemented fully. As statistical devices to correct individual differences, mixed-effects models can isolate stylistic effects and speaker effects in nested data structures, which has methodological backing of cross-contextual comparisons.

II. RELATED WORK

The phonetic analysis of vowel length in Japanese has a fairly well-developed theoretical framework. Experiment phonetics have also established the duration ratio of long and short vowels in standard pronunciation to be at par with two to one. Initial research mostly relied on isolated word reading, which involved the absolute length of time of the five vowels in various syllable positions through manipulation of the consonant environment and stress patterns. These researches established that the vowels at the onset of a word are far longer than those in the middle and at the end of a word because of the effect of the enhancement of prosodic boundaries. Despite the abundance of acoustic parameters that has been gathered by traditional studies methods, the reproducibility of the results has been constrained by the fact that the experimental materials used are artificial [3,4]. Newly, however, as corpus phonetics have emerged, scholars are starting to be interested in the patterns of duration variation in the actual communicative situation, and most of them concentrate on the analysis of a single syllable. It has been confirmed that the effect of stylistic variations on prosodic aspects is evident in more than one language. In a formal style, the accent is generally more pronounced, speech slower and prosodic patterns more regular. The pitch contours and pause patterns are the main areas of the Japanese studies, and the fundamental frequency variation between news broadcasts and normal conversations has been found to be very different between them [5,6]. The prosodic parameter of vowel length is not often studied alone in cross-stylistic studies. Other experiments artificially produce spoken language through the reading of dialogue scripts, however, this procedure does not reproduce the effect of hesitation, repetitions, and self-corrections on duration in spontaneous speech. Although it is generally believed that speech rate is a major cause of variation in duration, little is known about the interaction process between the speech rate and the prosodic boundaries. Mixed-effects models have emerged as a widely used analytical instrument of English prosodic studies, and can simultaneously address both the effect of fixed and random effects, isolating individual variation and systematic variation [7,8].

III. METHODS

A. CORPUS CONSTRUCTION AND DATA COLLECTION

In this research, 20 native Japanese speakers (10 men and 10 women) were used as participants in a recording experiment (age 22-35 years). The natural conversation data were collected by a paired conversation task. Pairs of the participants were free to discuss a given topic in a soundproofed room over a period of 15 minutes. Issues discussed the daily life, work experience, and hobbies. The stereo audio was simultaneously recorded at 48kHz/16bit with a ZOOM H6 portable recorder. In the case of written language data, 30 NHK news articles of average length (150 characters) were picked. Participants needed to read them aloud in a normal broadcast speed, among which the article was recorded thrice and the best version was selected. The background noise in the recording setup was also maintained to less than 35dB and the distance of the microphone remained at 25cm.

The Julius speech recognition engine was applied for automatic syllable boundary detection, achieving a 92.3% temporal alignment accuracy compared to a gold-standard manual subset (100 samples) within a ±20ms tolerance window. In the manual proofreading phase, two trained annotators independently corrected these boundaries using Praat software. Inter-annotator agreement for boundary placement was assessed using the Intraclass Correlation Coefficient (ICC), yielding a value of 0.92, while the vowel category labeling achieved a Kappa coefficient of 0.89. In the manual proofreading phase, the positions of boundaries were verified manually by two trained annotators on Praat software and conflict situations were handled by third party adjudication. The last corpus has 4,891 vowel samples in the dialogue modality and 3,456 samples in the written language modality and a total of 83 minutes. The samples of each vowel capture its type of phonemes and the position in a syllable, phonemes surrounding it, stress and rate of speech, creating a structured dataset of 12 dimensions [9,10].

B. VOWEL DURATION ANNOTATION AND EXTRACTION PROCESS

The assessment of the vowel duration used the hierarchical annotation system of Praat 6.3.10 software TextGrid. Vowel boundaries were outlined with the help of both broadband spectrogram (5ms window) and waveform graph by the annotators. The starting point of the surge of the energy was based on the position of the energy burst at the second formant F2 and the end point was based on the diminution of the energy at the F2 to emergence of the neighboring consonant spectrum attributes. In VV sequences between vowels, the boundary was determined at the cross over point of the F2 curves of the two vowels. The marking of long vowels was based on two perceptual-acoustic principles of verification: once the contrast between long and short vowels was established after the auditory testing, the duration of the vowel segment was measured to define whether it was more than a threshold of 1.5 times the average of short vowels in this position. Praat script will automatically obtain the value of length of each marked interval and also read out the pitch profile (fundamental frequency F0 sampling point, at 10ms intervals) and the frequencies of the first three formants. Three levels of screening were implemented for quality control: segments identified as vowels shorter than 30ms were excluded based on the established boundary detection limits of the Julius engine and to avoid acoustic artifacts, a threshold consistent with standard practices in Japanese corpus phonetics to distinguish vocalic nuclei from transitional glitches. While high vowel devoicing can lead to extreme compression, these samples were handled separately to ensure that the primary duration analysis reflects measurable periodic energy, creamy voice parts with F0 jumps over 50Hz/10ms, and samples with unusual formant bandwidth are eliminated. Two labelers performed a consistency test on 500 randomly selected samples, with a Kappa coefficient of 0.89. The extracted raw duration data is imported into the R language environment, and z-score normalization is used to eliminate individual differences in pronunciation speed, generating normalized duration values as input variables for statistical analysis.

C. CALCULATION METHOD FOR THE RATIO OF LONG TO SHORT NOTES

The long and short vowel contrast ratio calculation is based on the minimum pair database. 247 pairs of words with only different vowel lengths are extracted from the corpus. Each vowel position is calculated separately to avoid the systematic deviation of the word beginning, middle and end positions affecting the reliability of the results.

The choice of paired samples is based on a strict principle of environmental consistency: the previous and successive phonemes should be placed in the same environment, the stress pattern needs to have a flat or undulating correspondence, and the difference in the speech rate should be regulated within 0.5 syllables per second. The analysis of coarticulation of the speech corpus that leads to vowel centralization is observed by two-dimensional formant space clustering analysis. Sample of more than two standard deviations different with the centroid of the prototype vowel are marked separately and contrast values are analyzed statistically, however, they are not removed in the dataset to maintain the actual speech characteristics.

The mean contrast ratios of each vowel category were mean of both spoken and written modalities and the range of the confidence interval was estimated through a resampling method which was repeated 10,000 times. The /i/ and /u/ vowels are highly influenced by the high vowel weakening rule and this necessitates some further stratified calculations between spoken and silent environments. The Welch modified test was used to determine the significance of the differences between the two groups and the strength of the effects was measured using Cohen standardized difference index, which measures the strength of the effects as being strong having a value greater than 0.8.

D. MIXED-EFFECTS MODEL AND STATISTICAL TEST

The lme4 package in R was utilized to construct the mixed-effects model. The model specified vowel duration as the response variable, with register, vowel type, and syllable position as fixed effects. To account for nonsystematic variance, individual speakers (n=20) and unique lexical items (word types, n=412) were included as random intercepts. The model was fitted using the BOBYQA optimizer with maximum likelihood estimation. Convergence was confirmed by a gradient tolerance of less than 0.001, and the model was checked for singularity using the 'isSingular' function to ensure the random effects structure was supported by the data volume.

Interaction effect test involved interaction second-order effect of register and vowel type interaction and interaction third-order effect of register and vowel type and position of syllable. The Akaike information criterion and the Bayesian information criterion were used to assess the models that were compared and the best model among the nested models was selected by means of the likelihood ratio test. Residual diagnostics revealed that there was a good fit to the normality QQ plot, and the homogeneity of variance was checked by the Levene test.

The independent samples t-tests were employed to confirm the overall difference between the duration of vowels in spoken and written language. The normality of the data distribution was tested using the Shapiro-Wilk test before the test and the homogeneity of variance was tested using an F-test. Same subject t-tests were used in comparison of the time spent in both language styles. Besides d used by Cohen, delta by Glass was applied to deal with unequal effects of the variance in the calculation of the effect size. To prevent inflation of the family error rate, Bonferonni adjustment had been applied to the significance level of numerous comparisons.

IV. RESULTS AND DISCUSSION

A. COMPARATIVE ANALYSIS OF VOWEL DURATION IN NATURAL DIALOGUE AND WRITTEN LANGUAGE

Mixed-effects model analysis revealed a significant main effect of register type on vowel duration, with the duration reduction phenomenon in the dialogue modality being consistent across all participants. The goodness-of-fit index showed a conditional R^2 of 0.78, indicating that fixed and random effects together explained 78% of the duration variation. The interaction effect between register and vowel type was significant, with an F-value of 15.32, suggesting differences in the sensitivity of different vowels to register changes. Residual analysis found no systematic deviations, validating the model's hypotheses. Paired t-tests confirmed the duration compression trend in the dialogue context, with an effect size of Cohen's d of 1.24, falling within the strong effect category.

Table 1 shows that the duration compression effect at the word-final position is the most significant, with the final vowel shortened by 28.4% in conversational contexts compared to written language, exceeding the overall average reduction. The difference at the word-initial position is relatively mild, with a reduction rate of 21.3%, indicating that the prosodic boundary reinforcement effect is more pronounced in written language reading. The non-overlapping confidence intervals confirm the robustness of the differences between groups. Standard deviation data show that the duration variability in the conversational modality is smaller than that in written language, reflecting a trend towards simpler duration control in natural communication. Differences in t-values at different positions indicate that syllable position modulates the intensity of the register effect; the word-final position, as a sensitive area for prosodic weakening, experiences greater duration reduction pressure when the speech rate increases.

TABLE 1. Comparative Analysis Results

Register	Average duration of vowels (ms)	Standard deviation (ms)	95% confidence interval	Sample size	t-value
Natural Dialogue	78	18.3	[75.2, 80.8]	4891	-18.76
Written language	102	22.7	[98.5, 105.5]	3456	-
Difference rate	-23.4%	-	$[-25.8\%, -21.0\%]$	-	$p<0.001$
Dialogue at the beginning of words	85	19.1	[81.8, 88.2]	1634	-12.43
Preposition of words in written language	108	23.5	[103.9, 112.1]	1152	-
End-of-word dialogue	68	16.8	[65.3, 70.7]	1589	-21.35
Word ending position (written language)	95	21.2	[91.7, 98.3]	1098	-

B. DIFFERENTIATED REPRESENTATIONAL CHARACTERISTICS OF THE FIVE TYPES OF VOWELS

To compare the five vowel categories of the research, it is necessary to control the environmental variables of phonemes and use the minimum pair of contrasting samples with CVV structure in order to have the same benchmarks of comparison. Some subsets were created during the preprocessing of data, and each of them included a pair of samples of the spoken or written language based on the type of a vowel. The devoicing of high vowels /i/ and /u/ frequently occurs in environments following voiceless consonants; hence, their verbal samples in the environment were isolated to eliminate the effect of the occlusion segments on the duration measurement. The front and back nasal environmental samples of the /e/ and /o/ vowels were coded since nasalization synergy can increase significantly the vowel time by about 8-12 milliseconds. The phoneme with the largest opening (/a/ vowel) was used as a benchmark to confirm the relative change of other vowels as it has the most stability in duration.

Table 2 shows that the shortening rate of the /u/ vowel reached 31.7%, significantly higher than other vowel types. This is consistent with the phonological pattern that high back vowels in Japanese are prone to weakening during rapid pronunciation. The long/short contrast ratio of the /u/ vowel decreased to 1.54 in the dialogue, indicating that stylistic shifts had the most significant impact on its ability to maintain long vowels. The shortening rate of the /i/ vowel was 27.6%, ranking second, reflecting that high vowels are generally more susceptible to the pressure of speech rate. The LSR values of the /a/ and /o/ vowels remained above 1.75 in the dialogue, showing that the duration contrast of open vowels was well preserved.

TABLE 2. Duration characteristics and ratios of long and short vowels for the five vowel categories in different registers

Vowel types	Average conversation duration (ms)	Average duration of written language (ms)	Shortening rate (%)	Dialogue LSR	LSR (Literal Language)
/a/	82	105	21.9	1.75	2.38
/i/	71	98	27.6	1.62	2.25
/u/	65	95	31.7	1.54	2.29
/e/	79	103	23.3	1.71	2.35
/o/	84	107	21.5	1.78	2.36

C. THE INFLUENCE MECHANISM OF PROSODIC BOUNDARIES AND SPEECH RATE

The interaction between prosodic boundaries and speech rate was quantified using regression analysis. Speech rate was measured in syllables per second (SPS), averaging 5.8 SPS in conversation and 4.2 SPS in written language. Prosodic boundaries were divided into two levels: intonation phrase boundaries and discourse boundaries. Vowels before the boundary showed a duration extension effect, while vowels after the boundary showed compression. In the model, the interaction term between speech rate and prosodic position explained 41% of the vowel duration variation, while speech rate alone contributed 28%, and prosodic position contributed 19%. The synergistic effect of the two was stronger than their independent effects. Table 3 shows that the prolongation effect of the pre-boundary position in written language reaches 18 milliseconds, while in dialogue it is only a relative prolongation, but the absolute duration is still shorter than the written language baseline. A negative speech rate influence coefficient indicates that increased speech rate leads to duration reduction; vowels within phrases are most sensitive to changes in speech rate, with an absolute coefficient of 0.71. A compression effect of 12 milliseconds appears at the post-boundary position, indicating an acceleration phenomenon after prosodic reset. The prolongation effect at the utterance end position exists in both registers, but in dialogue, this effect is partially offset by overall speech rate pressure.

TABLE 3. Distribution of vowel duration under different prosodic positions and speech rates

Rhythmic position	Conversation duration (ms)	Written language duration (ms)	Speech rate influence coefficient	Boundary elongation effect (ms)
Front position of the boundary	95	128	-0.62	+18
post-boundary position	64	89	-0.48	-12
inside the phrase	76	99	-0.71	0
The beginning of the speech	88	115	-0.55	+14
End of speech	92	125	-0.59	+16

V. CONCLUSION

Through the application of systematic analysis of a bimodal corpus, the paper has found a vast difference in the vowel duration in natural conversation and written reading of Japanese language. The primary effect of the stylistic type on vowel length is confirmed by a mixed-effects model, showing that word-final positions where the weakening of prosodic features is especially sensitive also show more pressure with the pressure to reduce the duration of the word. The five types of vowels are sensitive to changes in style, the high-back vowel /u/ taking the most weakness, whereas the duration contrast of open vowels /a/ and /o/ is only slightly lessened. The association between prosodic limits and speaking rate forms the primary force of variation of duration; the effect of prolongation preceding the limit is stronger in written language, whereas the effect of compression acceleration follows the limit. This study establishes the dynamic regulation law of vowel duration in natural speech flow, providing operable parameters for Japanese prosody teaching, particularly emphasizing the weakening of word-final positions and high vowels in cultivating learners' naturalness of speech.

AUTHOR CONTRIBUTIONS

Peng Liu designed, collected and analyzed the data, and drafted the manuscript. Peng Liu conducted the study, critically revised the manuscript for important intellectual content, and gave final approval of the version to be published. Peng Liu participated fully in the work, take public responsibility for appropriate portions of the content, and agreed to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

CONFLICTS OF INTEREST

The author declares no conflict of interest.

FUNDING

This work was supported by Research on Japanese Loanwords from the Viewpoint of Cultural Accommodation, 2022XJLXBSJJ07.

ACKNOWLEDGEMENTS

Not applicable.

REFERENCES

[1] Y. Huang, "On the cooperativeness of Japanese people from the perspective of Japanese dialogue - focusing on the modification type of "obstacle source" in dialogue, " Modern Linguistics, vol. 11, no. 12, pp. 5992-5999, 2023, doi: 10.12677/ML.2023.1112803.

[2] J. Wang and M. He, "Japanese Language and Culture Promotion Strategies and Their Implications for the Development of International Chinese Language Education, " Comparative Studies in Cultural Innovation, vol. 7, no. 11, pp. 176-180, 2023, doi: 10.3969/j.issn.2096-4110.2023.11.035.

[3] Y. Song, "Practice of "work-based" teaching reform in Business Japanese listening and speaking courses, " Journal of Hubei Open Vocational College, vol. 33, no. 14, p. 3, 2020, doi: CNKI: SUN: HBHS.0.2020-14-071.

[4] L. Haiqi, "The relationship between Japanese intonation and age group - based on the Japanese daily conversation corpus, " Modern Linguistics, vol. 12, no. 12, pp. 285-291, 2024, doi: 10.12677/ml.2024.12121148.

[5] X. Huang and C. Liu, "A comparative study of kinship terms in China and Japan based on discourse politeness theory— taking the terms of address between sisters as an example, " Modern Linguistics, vol. 12, no. 7, pp. 319-325, 2024, doi: 10.12677/ml.2024.127561.

[6] U. Dessi, "Culture as Power: Buddhist Heritage and the Indo-Japanese Dialogue, " Edited by MadhuBhalla. New York: Routledge; 2021. xxiv + 213 p, doi: 10.1111/rsr.17014.

[7] Y. C. Gomez, "How Do Japanese EFL Learners Fail to Engage in a Collaborative Dialogue in a Group Project? International Journal of Curriculum Development and Practice, " vol. 24, no. 1, pp. 31-45, 2021, doi: 10.18993/jcrdaen.24.1_31.

[8] J. Shan, Y. Nishihara, R. Yamanishi, et al., "Japanese Language Listening Practice Support Using Dialogue Scenes of Anime with Estimated Levels, " The Journal of the Society for Art and Science, vol. 20, no. 2, pp. 108-119, 2021, doi: 10.3756/artsci.20.108.

[9] Y. Hayashi, Y. Kondo, and Y. Ishii, "Automated speech scoring of dialogue response by Japanese learners of English as a foreign language, " Innovation in Language Learning and Teaching, vol. 18, no. 1, pp. 32-46, 2024, doi: 10.1080/17501229.2023.2217181.

[10] H. Sugiyama, "Achievements and challenges of a large-scale deep learning based Japanese chatting dialogue system, " THE JOURNAL OF THE ACOUSTICAL SOCIETY OF JAPAN, vol. 78, no. 5, pp. 271-276, 2022, doi: 10.20697/jasj.78.5_271.