IIAI Open Conference Publication Series
IIAI Letters on Informatics and Interdisciplinary Research
Vol. 006, LIIR334
DOI: https://doi.org/10.52731/liir.v006.334

Quantitative Evaluation of Perceived Acceptability of Accentual Patterns in Four-Mora Japanese Words Based on Lexical Attributes –for Implementation into CALL System for Japanese Language Learners–

Ikuyo Masuda-Katsuse [*], Ayako Shirose [†]

Abstract

We generated four-mora Japanese word utterances with various pitch accent features and mapped the perceived pitch accent acceptability of native Japanese speakers onto a distribution of pitch accent features. The results confirmed that even for words classified as having the same accent type, the acceptability distribution differed depending on the word types. These results demonstrate the diversity of pitch accent perception and provide quantitative support from the perspective of auditory perception for knowledge about accent rules reported in linguistics. We also discussed the potential application of these results to CALL systems for Japanese language learners.

Keywords: acceptability distribution, CALL system, lexical properties, pitch accent

1 Introduction

Spoken Japanese exhibits a pitch accent, which is determined by the pitch contour of each mora. In the context of four-mora words in standard Japanese, typical accent types are classified as type 0 (low-high-high-high), type 1 (high-low-low-low), type 2 (low-high-low-low), or type 3 (low-high-high-low). The rules governing these accents are intricate and multifaceted, posing significant challenges for learners who seek to acquire accurate pitch accent production.

A quantitative assessment of pitch accents for assisting learners is crucial. Deep learning offers a powerful framework for extracting meaningful pitch accent feature representations for such quantitative assessments. When the primary objective is simply classifying words into their respective accent types, deep learning models have demonstrated high levels of performance [1]. However, since spoken language is inherently diverse, a critical question arises: can feature representations learned from training data annotated with limited labels, such as accent type, adequately capture such inherent diversity? To investigate this question, we previously investigated the relationship between the feature representation of pitch accent and its audible adequacy by native speakers and concluded that even words categorized into the same accent type may exhibit different distributions of acceptability depending on their part of speech [1].

Our overarching goal is to develop a robust Computer-Assisted Language Learning (CALL) system that can effectively evaluate learners' pitch accent production and provide insightful feedback. This system must consider both the lexical properties of words and the subjective acceptability judgments of native speakers. Our previous study [2] proposed a CALL system that leverages the distribution of subjective acceptability ratings, extracts the pitch accent features from speech produced by learners during practice sessions, and subsequently plots their production

[*] Kindai University, Fukuoka, Japan
[†] Tokyo Gakugei University, Tokyo, Japan

within the distribution of acceptability ratings provided by native speakers (Figure 1, bottom graph). This visualization allows learners to gain a clear understanding of how "appropriately" their pitch accent is perceived by native speakers. Moreover, they can interactively explore the acceptability space by clicking on any arbitrary point within the graph. Such interaction enables them to listen to transformed speech samples that exhibit pitch accent characteristics corresponding to the selected point for facilitating direct comparison between their own production and highly acceptable examples.

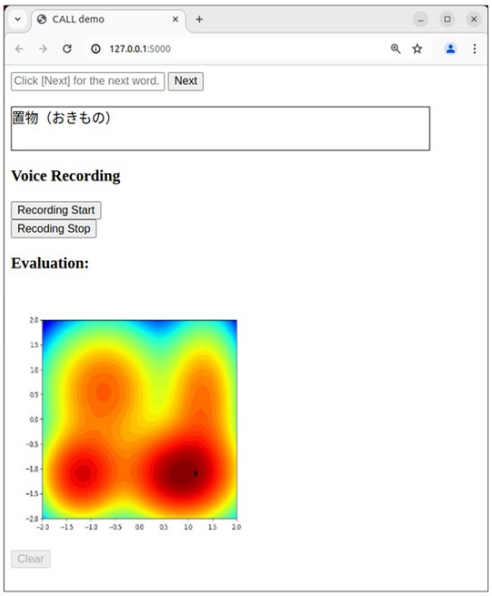

Figure 1: Sample screen of our CALL system: Black dot indicates where learner's voice is positioned.

Even though the acceptability distributions may differ by not only accent types but also the other lexical properties, detailed classification of lexical properties is inappropriate for language learners. We must build a CALL system with accent evaluation and feedback based on the minimum required distributions of acceptability. Therefore, in this study, we extended our previous research to examine the relationship between acceptability distributions and word types and identified the types of acceptability distribution that can be provided as feedback in a CALL system based on the lexical properties of practice words.

2 Modeling of Pitch Accent Features and Correlation with Perceived Acceptability

2.1 Modeling Pitch Accent Features

We constructed a new dataset of 4-mora words using our previously employed multiple speech corpora [1] and partitioned it into three sets: training (4,500 words per accent type), testing (400 words per accent type), and subjective evaluation (50 words per accent type).

Following our previous research [1], our model is based on the S3VAE model, which is an autoencoder model that takes a fundamental frequency (F0) as input, encodes it into a latent variable representing F0 characteristics, and reconstructs the F0. Latent variables are decomposed

into both time-varying and time-invariant latent variables. The former represent voiced/unvoiced information at each time step; the latter represent the characteristics of pitch accent. We retrained the model using a new, larger training dataset than in our previous study.

The distribution of the latent variables representing the characteristics of the pitch accents in the test data is shown in the left panel of Figure 2. Color-coded numerical markers indicate the accent types of the speech input to the model. Type 0 is centered around coordinates (1, -1), type 1 around (1, 1), type 2 around (-1, 1), and type 3 around (-1, -1). By manipulating the latent variables representing the pitch accent characteristics, F0s can be reconstructed with various characteristics. The right panel of Figure 2 shows the reconstructed F0s plotted based on their coordinates in the latent space.

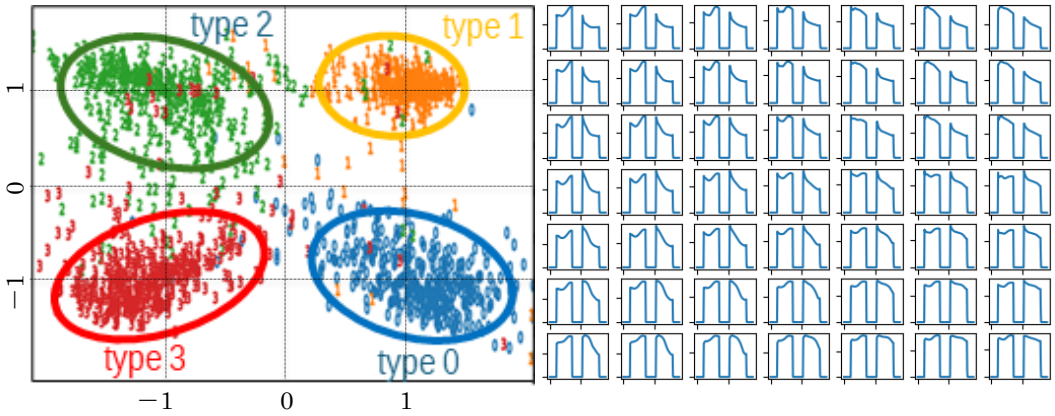

Figure 2: (Left) Distribution of latent variables representing pitch accent features. (Right) Illustrative examples of fundamental frequency (F0) reconstructed from individual latent factors.

2.2 Subjective Rating Experiment

For our rating experiment, we first selected words with a word familiarity score [3] of 4.0 or higher that matched the accent types found in accent dictionaries [4, 5]. To examine the influence of the linguistic attributes on pitch accent acceptability, we selected words for each accent type and ensure a balanced distribution across various parts of speech and word types, such as origin (native Japanese, Chinese, or foreign) or structure (simple or compound). Finally, we selected words for each accent type so that the total number of words for each type was 50. Table 1 shows each lexical property and the number of words.

Forty-nine university students (24 females and 25 males, aged around 20, native speakers of Japanese) participated in the rating experiment. The questionnaire results indicated that many participants experienced dialects from multiple regions. However, when classified according to the Zones map of Japanese pitch accent [5] based on their primary dialect exposure, 32 were from the Tokyo Japanese, 12 from the Keihan Japanese, 1 intermediate Japanese, and 4 from non-accented region.

We obtained informed consent from our participants after explaining the experiment. Each participant was presented with 50 words, and each was assigned randomly to one of the 49 types of F0, ensuring no overlap among the participants. They listened to the speech stimuli through earphones or headphones. After listening to each stimulus as many times as they desired, they rated the acceptability of the pitch contour for each word on a 5-point scale from 1 (inappropriate) to 5 (appropriate).

Table 1: Lexical properties of words in rating experiment

Accent type	Part of speech	Detailed lexical properties	Number of words
0	Noun	Native Japanese compound noun	10
		Sino-Japanese compound noun	10
	Verb	Simple verb	10
		Compound verb	10
	Adverb		5
	Onomatopoeia		5
1	Noun	Sino-Japanese compound noun	10
		Noun of loan words	10
	Adverb		15
	Onomatopoeia		15
2	Noun	Native Japanese compound noun	10
		Sino-Japanese compound noun	10
		Noun of loan words	20
	Adverb		10
3	Noun	Native Japanese compound noun	10
	Verb	Simple verb	10
		Compound verb	10
	Adverb		10
	Adjective		10

3 Analyses of Differences in Average Rating Distributions

Words were grouped based on their word types, and we calculated their mean ratings in the latent space to obtain average rating distributions. We employed a two-dimensional Kolmogorov-Smirnov test (2D-KS test) [6] to statistically determine the variations in the distributions of average ratings for words with different word types. This test determines whether two probability distributions differ based on finite samples. The following two approaches were taken to acquire a finite number of samples.

3.1 Differences in Average Rating Distributions for Lexical Properties

The KS test requires a finite number of samples drawn from an arbitrary probability distribution. We improved the power of the 2D-KS test using the bootstrap method. Although too few samples cannot adequately reflect distribution, the 2D-KS test becomes computationally expensive with too many samples. Therefore, we treated the integers obtained by multiplying the rating average on a 5-point scale by 50 as the number of observations at a corresponding coordinate.

First, we conducted a 2D-KS test to compare the distributions between different parts of speech. We also conducted this test for the detailed word types that can be compared within parts of speech, such as compound and simple verbs. The test results are presented in the "2D-KS test for unmodeled samples" column (Table 2).

3.2 Differences in GMM Applied for Rating Averages for Lexical Properties

In the previous section 3.1, we directly generated samples from the distribution of average ratings.

However, in our CALL system, to provide feedback on the subjective acceptability by native speakers in a more comprehensible manner, we applied a Gaussian Mixture Model (GMM) to the distribution of acceptability [1]. Thus, we needed to verify whether the distributions modeled by this GMM varied based on the lexical properties of words.

Similar to the previous study [1], we applied GMM to the probability distribution, which is considered as the distribution of the average ratings. We set the maximum number of mixture components to 4 and selected the number of components and variances using the Bayesian Information Criterion (BIC) [7]. To conduct 2D-KS tests for GMMs estimated for different word groups with different lexical properties, we needed a finite number of samples. Since the estimated GMM is a continuous probability density function, we first discretized the distribution by dividing the range of -2.1 to 2.1 on each axis into 21 equal-width bins and integrating the probability density function within each one. Then we multiplied this integrated probability by 1,000 to obtain a discrete frequency distribution, which was used as the sample size. Finally, we performed 2D-KS tests whose results are shown in the "2D-KS test for GMM-generated samples" column (Table 2).

3.3 Results and Discussions

The yellow cells in Table 2 highlight significant differences in the part-of-speech contrast, regardless whether the samples were unmodeled or generated using a Gaussian mixture model. In this section, we explain the reasons for these significant differences and examine the groups of word types that constitute acceptability distribution used for feedback in CALL systems.

3.3.1 Significant differences in accent type 0

Significant differences were found in accent type 0 between the onomatopoeias and the other parts of speech. Figure 3 shows the probability distributions obtained by applying a Gaussian mixture model to the average ratings for nouns, onomatopoeias, and verbs in accent type 0. Nouns peak around accent type 0, but onomatopoeias have high ratings not only around accent type 0 but also around type 1. Fukumachi [8] argues that the accent type of an onomatopoeia is determined by the subsequent word. For example, *tokotoko* can be either type 0 or 1, depending on the word that follows it. When a noun follows, as in "*tokotoko* aruki" (a *pitter-patter* walk), it is type 0. However, when a verb follows, as in "*tokotoko* aruku" (to walk *pitter-patter*), it is type 1. In our experiment, since onomatopoeias were presented in isolation, the participants' ratings may have reflected which pattern they imagined.

Furthermore, we found significant differences between verbs and the other parts of speech in accent type 0. Figure 3 shows that verbs received higher average ratings in areas where accent types 0 and 3 were concentrated. According to the Japanese Accent Dictionary [5], compound verbs of two verbs, where the first verb has an undulating accent, are generally flat (type 0). However, middle-aged and younger people tend to pronounce them with a mid-high accent (type 3). Perhaps this tendency is reflected in the distribution of the average ratings. Therefore, we divided the verbs of accent type 0 into simple and compound and applied GMM to each group (Figure 4). These results show that simple verbs peak around accent type 0 in the accent feature distribution, while compound verbs peak around type 3. Furthermore, statistical tests confirmed a significant difference (Table 2). These results support the idea that, for the university students in this experiment, compound verbs labeled as accent type 0 are more likely to be perceived as having a type 3 accent.

Table 2: Comparisons of acceptability distributions based on word types within each accent type using the Kolmogorov-Smirnov test

Accent type	Contrast	2D-KS test for unmodeled samples	2D-KS test for GMM-generated samples
0	Nouns — Verbs	d = 0.0702 p = 0.0136*	d = 0.0896 p = 0.00871**
	Nouns — Adverbs	d = 0.0402 p = 0.402	d = 0.0449 p = 0.502
	Nouns — Onomatopoeias	d = 0.122 p = 7.98e-07**	d = 0.143 p = 1.75e-06**
	Verbs — Adverbs	d = 0.0865 p = 0.00105**	d = 0.101 p = 0.00206**
	Verbs — Onomatopoeias	d = 0.144 p = 1.59e-09**	d = 0.146 p = 1.08e-06**
	Adverbs — Onomatopoeias	d = 0.150 p = 5.30e-10**	d = 0.138 p = 4.86e-06**
	Native Japanese compound nouns — Sino-Japanese compound nouns	d = 0.0409 p = 0.371	d = 0.0884 p = 0.0103*
	Compound verbs — Simple verbs	d = 0.0919 p = 0.000279**	d = 0.0985 p = 0.00277**
1	Nouns — Onomatopoeias	d = 0.0558 p = 0.0672	d = 0.0685 p = 0.0825
	Nouns — Adverbs	d = 0.0473 p = 0.161	d = 0.0616 p = 0.151
	Adverbs — Onomatopoeias	d = 0.0341 p = 0.504	d = 0.0395 p = 0.663
	Sino-Japanese compound nouns — Nouns of loan words	d = 0.0558 p = 0.0735	d = 0.0506 p = 0.352
2	Nouns — Adverbs	d = 0.0177 p = 0.994	d = 0.0382 p = 0.702
	Native Japanese compound nouns — Sino-Japanese compound nouns	d = 0.0455 p = 0.186	d = 0.0441 p = 0.523
	Native Japanese compound nouns — Nouns of loan words	d = 0.0316 p = 0.615	d = 0.0296 p = 0.928
	Sino-Japanese compound nouns — Nouns of loan words	d = 0.0295 p = 0.704	d = 0.0265 p = 0.971
	Compound nouns — Simple nouns	d = 0.0157 p = 0.999	d = 0.0172 p = 1.00
3	Nouns — Verbs	d = 0.0622 p = 0.0318*	d = 0.0839 p = 0.0175*
	Nouns — Adjectives	d = 0.0769 p = 0.00396**	d = 0.0803 p = 0.0264*
	Nouns — Adverbs	d = 0.0665 p = 0.0183*	d = 0.0913 p = 0.00741**
	Verbs — Adverbs	d = 0.0203 p = 0.981	d = 0.0446 p = 0.516
	Verbs — Adjectives	d = 0.0450 p = 0.237	d = 0.0732 p = 0.0543
	Adverbs — Adjectives	d = 0.0558 p = 0.0777	d = 0.0746 p = 0.0459*
	Compound verbs — Simple verbs	d = 0.0603 p = 0.0395*	d = 0.0521 p = 0.313

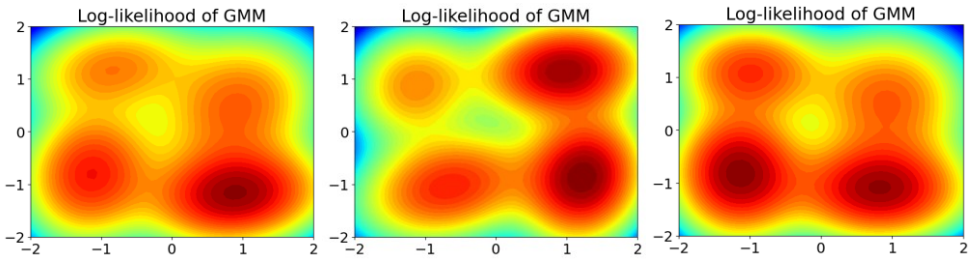

Figure 3: Distributions of average ratings of words with accent type of 0, categorized by part of speech

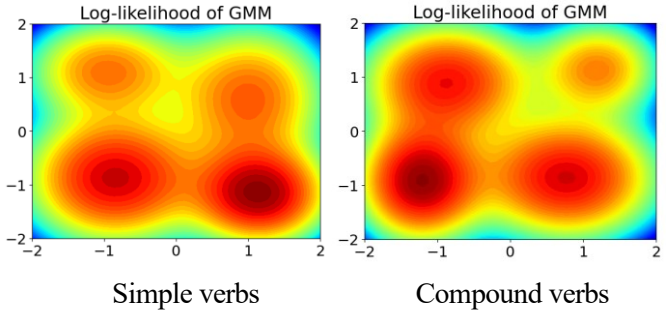

Figure 4: Distributions of average ratings of simple and compound verbs with accent type 0

3.3.2 Significant differences in accent type 3

Table 2 shows significant differences in accent type 3 between the nouns and the other parts of speech. Figure 5 presents the distribution of a GMM applied to the average ratings of the nouns and adjectives in accent type 3. For nouns, high ratings were obtained not only around accent type 3 but also around type 0. An examination of the components of nouns assigned to accent type 3 revealed that they were compounds consisting of noun + native Japanese nouns, noun + verb, or adjective + noun. Linguists have reported that these compound nouns can also be type 0 [5]. In various accent dictionaries, these words have been labeled as both types 0 and 3. Our results corroborated that four-mora nouns with accent type 3 are highly acceptable with both accent types 0 and 3.

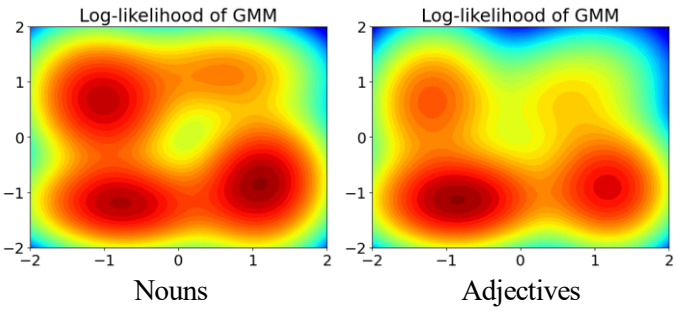

Figure 5: Distributions of average ratings of nouns and adjectives with accent type 3

3.3.3 Classification of acceptability distributions based on word types

Based on the above discussion, we propose the following categorization of the acceptability distributions for evaluating learners' pitch accents:

- **Accent type 0:** Onomatopoeia, compound verbs, and other word types (noun, simple verb, adverb).
- **Accent type 1:** All parts of speech (nouns, adverbs, onomatopoeias).
- **Accent type 2:** All parts of speech (nouns, adverbs).
- **Accent type 3:** Nouns and other parts of speech (verbs, adverbs, adjectives).

However, considering that the acceptability distribution of accent type 0 onomatopoeias is likely to vary depending on the subsequent word, an experiment is needed in which participants rate the appropriateness of pitch accents when presented with both an onomatopoeia and a subsequent word. This step will allow us to obtain a new acceptability distribution.

3.3.4 General discussions and future works

Although the participants in this experiment exhibited diverse dialectal experiences, our current findings can be primarily explained within the framework of existing linguistic knowledge regarding standard Japanese. In our experiment, the influence of the subjects' dialectal backgrounds was likely more pronounced as listeners than as speakers. Subjects are not only exposed to the dialect of their region of residence but also to standard Japanese and other regional dialects on a daily basis through various media. We believe that the listener-based evaluation employed in this study effectively reveals accents that deviate from the naturalness of Japanese, extending beyond the inherent diversity of dialects and other forms of speech. It is crucial to acknowledge that standard Japanese does not encompass the entirety of the Japanese language. Our ultimate goal is to develop a learning support system that comprehensively integrates its diverse facets to ensure that no dialect is excluded.

As indicated by the seminal Fujisaki model [9], the pitch pattern of speech is determined not only by the word itself but also by the preceding and succeeding linguistic context. However, the accent type of individual words remains relatively stable across different contexts, with such exceptions as onomatopoeia, as previously discussed in 3.3.1. Therefore, this study's findings are valuable for learning at the word, clause, phrase, and sentence levels. Furthermore, although this study focused on four-mora words encompassing various word types, our research methodology is expected to be applicable to words with other mora counts.

This study primarily analyzed variations in the acceptability distribution based on word type. However, numerous other factors can also influence these distributions, including the subject's place of origin, gender, and the speaker's gender. The statistical analysis method employed in this study enables the detection of differences not only in the location of the peaks of the appropriateness distribution (i.e., the accent type) but also in the distribution's variance. Such detailed differences hold significant linguistic interest. As a future research direction, we plan to implement a new experimental design that specifically targets these attributes for linguistic analysis, while carefully considering the balance among the number of words, subjects, and speakers for each attribute. The results of this future research will be reported in a separate publication.

4 Conclusions

We evaluated not only the accent types of learners' spoken words but also their acceptability to native Japanese speakers by correlating the latent space where the characteristics of pitch accents

in four-mora Japanese words are distributed with the "acceptability" of these pitch accents as perceived by native speakers. We found that the distribution of average ratings by native speakers is influenced by the lexical properties of the words. The onomatopoeia grouped into accent type 0 can be more appropriate with accent type 1 depending on the following word, the compound verbs of accent type 0 can also be appropriate with accent type 3, and accent type 3 nouns can also be appropriate with type 0.

Linguistic research on accent has traditionally prioritized a speaker-centric perspective. This study offers a novel approach by investigating the perception of accented speech from the listener's standpoint. Our results provide quantitative support from the perspective of auditory perception for knowledge about accent rules reported in the field of linguistics. We also discussed the potential application of these results to CALL systems for Japanese language learners.

Acknowledgement

This work was supported by JSPS KAKENHI Grant Number 24K04017.

References

[1] I. Masuda-Katsuse, “Feature learning of Japanese pitch accents and applications to Japanese speech education,” Proc. 2023 14th IIAI International Congress on Advanced Applied Informatics (IIAI-AAI), 2023, pp.188-193.

[2] I. Masuda-Katsuse and A. Shirose, “CALL system using pitch-accent feature representations reflecting listeners’ subjective adequacy,” Proc. 14th INTERSPEECH2024, 2024, pp.5206-5207.

[3] S. Amano and K. Kondo, Lexical Properties of Japanese, Sanseido, 1999. (in Japanese)

[4] NHK Japanese Pronunciation Dictionary, NHK Publishing, Inc., 2002. (in Japanese)

[5] H. Kindaichi (Supervisor) and K. Akinaga (Editor), Shinmeikai Japanese accent dictionary, second edition, Sanseido, 2014. (in Japanese)

[6] W. H. Press, S. A. Teukolsky, W. T. Vetterling, B. T. Flannery, “Do Two-Dimensional Distributions Differ?,” Numerical recipes in C (2nd edition), Cambridge University Press, 1992, pp. 645-649.

[7] G. E. Schwarz, Gideon E. , “Estimating the dimension of a model,” Annals of Statistics, 6 (2), 1978, pp. 461-464.

[8] K. Fukamachi, “Usage Classification of Onomatopoeia using Acoustic Features with Accent Position,” Proc. the 36th Annual Conference of the Japanese Society for Artificial Intelligence, 2022, 1P5-GS-6-04, pp. 1-3. (in Japanese)

[9] H. Fujisaki, In Vocal Physiology: Voice Production, Mechanisms and Functions, Raven Press, 1988.