Kashihara, A. et al. (Eds.) (2024). Proceedings of the 32nd International Conference on Computers in Education. Asia-Pacific Society for Computers in Education

Construction of a Japanese Language Learning Support System that Enables Word Accent Learning

Satoru KOGURE[a*], Kazuki TOMITA[a], Yasuhiro NOGUCHI[a], Koichi YAMASHITA[b], Tatsuhiro KONISHI[a], & Makoto KONDO[a]

[a]*Faculty of Informatics, Shizuoka University, Japan*
[b]*Faculty of Business Administration, Tokoha University, Japan*
*kogure@inf.shizuoka.ac.jp

Abstract: In second language learning, learners learn four skills: reading, writing, listening, and speaking. However, compared with the other three skills, speaking skills have rarely been studied in environments wherein they can be learned. Therefore, we proposed a method for recognizing Japanese word accents using the pitch average of the whole word and that of each mora as part of speaking skill learning. We demonstrated that this method outperforms the pitch drop detection method. Using the proposed accent recognition method, we embedded a function that enabled learners to learn the accents of target words into the feedback phase of an existing Japanese language learning support system. This allowed learners to speak the target words and receive feedback on whether their accents were correct.

Keywords: Japanese language leaning, word accent recognition, dictogloss

1. Introduction

In second language learning, learners learn the four skills of reading, writing, listening, and speaking. There are many second language learning support systems and applications for reading, writing, and listening skills. However, there are not as many systems and applications for speaking as for the other three skills. Several second language learning support systems and applications allow learners to learn speaking skills. There are also several services that provide a second language learning environment based on conversations with native speakers.

In addition to learning word pronunciation, it is also necessary to learn word accents and sentence intonation as part of speaking skills. In our interviews with Japanese language teachers, noted that sentence intonation was the next item that Japanese learners who had reached a certain level of proficiency should study. As a prerequisite for learning to speak, it is necessary for learners to have reading skills, such as vocabulary and grammatical understanding, and listening skills, such as listening comprehension.

The dictogloss environment proposed by Wajnryb (1990) is a second language learning environment that enables learning of the four skills. We also developed a dictogloss learning support system that enabled learners to learn reading, writing, and listening skills, and the learning of “pronunciation” among speaking skill.

This study aimed to implement the “word accent” learning environment of speaking skills into the existing dictogloss system.

2. Related Works

2.1 Several Systems, Applications, Environments, and Services of Second Language Learning for Speaking Skills

Duolingo is a game-like service used to teach a second language. Learners can learn the four skills of reading, writing, listening, and speaking either on the website or through the mobile application. *Rosetta Stone* is a learning service that focuses on the learning styles used to learn first languages: seeing, hearing, and speaking. Because no native languages are involved in learning a second language, one can learn how a child learns a language. *Babbel* provides a language learning environment that focuses on the acquisition of conversational skills. *Pearson Speaking Test* is a service provided by Pearson that automatically assesses English speaking abilities. Pearson also offers an essential four-skills test for English.

The first three learning systems and applications provide a wealth of learning content. They also have features that dynamically capture the learner's learning history and comprehension status, and review areas where the learner is having difficulty. However, they do not have a function for teachers to register new learning content based on the content used in their own classes.

Italki is a service that allows learners to have one-on-one lessons with personal native speakers or qualified language teachers. *Hello Talk* provides an environment in which, for example, an English learner whose native language is Japanese is paired with a Japanese learner whose native language is English, and the pair teaches each other their native languages. *Language Exchange* and *Tandem* also provide a learning environment for exchanging native and target languages, as well as chat, voice and video communication.

These environments are appropriate for language learning in the sense that they provide a very important activity for language learning: communication with native speakers. However, they require matching with language pairs and have limitations as the pairs learn simultaneously.

2.2 Dictogloss Environment

Dictogloss is a language learning method proposed by Wajnryb (1990) in which multiple language skills are acquired through cooperative interaction. Many classroom practices have been reported in actual language learning, and a several second language teachers believe that this learning environment is effective for second language learning. As a specific advantage, learners can learn listening, writing, reading, and speaking skills by listening to a text read by the teacher, while taking notes and reproducing the same text in consultation with other learners. However, this environment is not suitable for self-study, because it requires a cooperative learner.

Therefore, we developed a Japanese dictogloss learning support system that enables self-study for listening, writing, and reading skills by implementing a cooperative learning agent and a teacher agent (Kogure et al., 2017). In addition, we have implemented a dictogloss system that enables pronunciation learning for speaking (Kogure et al., 2020).

The dictogloss system used in this study was based on a previous system (Kogure et al., 2017) which we reimplemented to enable learning of both Japanese and English, and additionally implemented a function that enabled word accent evaluation.

2.3 Accent Recognition

Although there are many studies on word accents, most do not focus on recognizing word accents but on using word accents to improve the system for other tasks. Mikhailava et al. (2022) studied ways to improve speech recognition accuracy by considering word accent habits due to native language differences. Parikh et al. (2020) proposed a neural network-

based method that estimates the native language and converts the native accent into natural speech of the target language (Parikh et al., 2020).

2.4 Japanese Word Accent Corpus

The Online Japanese Accent Dictionary (OJAD) is a dictionary for Japanese learners (Minematsu et. al, 2012). OJAD contains approximately 9,000 Tokyo dialect accents for nouns, 42,300 accents for 12 basic conjugations for approximately 3,500 usages, and accent data for approximately 300 other postverbal expressions. In addition, male and female voice samples are provided for the verbs. In the present study, we used correctly accented speech of verbs in the OJAD as part of our experimental evaluation.

Although pronunciation evaluation has already been implemented for speaking learning, there are few self-study learning environments that focus on leaning word accents and sentence intonation. With “sentence intonation” as a future goal, we decided to construct a mechanism that could evaluate “word accent” and improve it so that it could be used with the existing dictogloss system.

3. Japanese Word Accent Recognition

3.1 Definition of Japanese Word Accents

Japanese word accents are expressed through changes in the pitch of sounds within words, called pitch accents. The pitch accent differs from the stress accent of English words. In Japanese word accents, the first and second morae have different pitch heights. The mora of the high pitch immediately before it decreases when changing from high to low pitch in a word is called the accent nucleus. Accent types were classified according to the presence and position of accent nuclei. There are three types of accents: Type 0 accents with no accent nucleus, Type 1 accents with an accent nucleus at the beginning of the word, and Type n accents in the middle and end of the word. Figure 1 shows the presence and absence of accent nuclei and accent types in four-mora words. “Ga” in the fifth mora of the “i mo o to” example in Figure 1 is a particle. This indicates that the particle “ga” associated with the word “i mo o to”, lowers the pitch. For Type 4 accented words with four morae (in Figure 1), the detector can only detect the downstroke when the next word (in this case, the particle "ga") appears. In this study, we judged that a Type-m accent could not be detected for a m-mora word, and in this case, we judged the word to be Type 0 (without an accent nucleus).

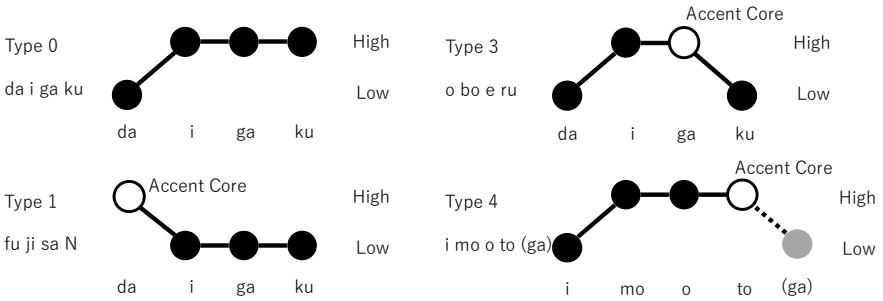

Figure 1. The Pitch Graph Examples for Each Accent Type.

3.2 Japanese Word Accent Recognition Focusing on "Pitch Drop”

The presence or absence of accent nuclei and their location are important for identifying Japanese word accents. Kitamura et al. (2019) proposed a method to identify accents by estimating the presence and location of accent nuclei at pitch-drop locations. For details of the pitch-drop method, please refer to Kitamura et al. (2019).

3.3 Accent Recognition Focusing on per-Mora Pitch Averages

In linguistics, many previous studies have defined Japanese accents, including what Japanese accents are and what happens to them in dialects. However, there has been little research on accent recognition systems for second language learners.

Therefore, we estimated the presence of an accent nucleus and the location of the accent nucleus for words with N morae using the following algorithm:

(1) The system determined the average μ_{all} of the pitches of the voiced intervals of all morae.
(2) The system finds the average μ_n of the pitches of the voiced interval of the nth mora.
(3) The system assigns a "high" or "low" label to each mora based on the following criteria.
(3-1) If $\mu_n \geq \mu_{all}$, the system assigns the "high" label to the mora.
(3-2) If $\mu_n < \mu_{all}$, the system assigns the "low" label to the mora.
(4) The system identifies the high part of the pair with the largest pitch-mean difference ($diff_{\mu_n} = \mu_n - \mu_{n+1}$) among several pairs that are changing from *high* (nth mora) to *low* ($(n+1)$th mora) as the accent nucleus.
(5) The system identifies the accent type from the identified accent nuclei.

As described in Section 3.2, we implemented a program to identify the accent type using *Praat* script.

3.4 Performance Evaluation of Accent Recognition for Two Methods

We investigated the performance of the two methods in discriminating Japanese word accents. We collected 304 utterances of the types mentioned below. The maximum number of morae in these utterances was four. The distributions of accent types were 120 for Type-0, 76 for Type-1, 32 for Type-2, 26 for Type-3, and 46 for Type-4.

(1) A total of 140 utterances of 20 words with correct accents were prepared by seven undergraduate and graduate students in the university of information science and 24 utterances of 12 words from OJAD in one male and one female voice for a total of 164 utterances.
(2) Twenty words uttered by seven undergraduate and graduate students with incorrect accents prepared by us for a total of 140 utterances.

For (1), the system judged a word as "correct" if it was recognized as the correct accent, and as "miss" if it was identified as any other accent. For (2), the system judged the word as "false alarm" if it was incorrectly identified as the correct accent and as "correct reject" if it was incorrectly identified as the correct accent. Accent recognition was performed for each Type-p accent. We calculated $n_{c,p}$ (the number of "correct" words in Type-p word), $n_{m,p}$ (the number of "miss" words in Type-p word), $n_{fa,p}$ (the number of "false alarm" words in Type-p word), $n_{cr,p}$ (the number of "correct reject" words in Type-p word) for the recognition results.

To evaluate the correct recognition performance, the precision and recall for each p-type accent were calculated using the following equations:

$$precision_p = \frac{n_{c,p}}{n_{c,p}+n_{fa,p}} \qquad recall_p = \frac{n_{c,p}}{n_{c,p}+n_{m,p}}$$

We also calculated the F-measure, which is a comprehensive measure of the precision and recall, and the accuracy. These were calculated using the following equation:

$$F_p = \frac{2precition_p \cdot recall_p}{precision_p + reclall_p} \qquad Accuracy_p = \frac{n_{c,p}+n_{cr,p}}{n_{c,p}+n_{m,p}+n_{fa,p}+n_{cr,p}}$$

Table 1 lists the results of accent recognition using Kitamura et al.'s (2019) method described in Section 2.2, and proposed average pitch method described in Section 2.3.

Table 1. *The Results of Accent Recognition of All Type with Pitch Drop Method (Section 3.2) and Proposed Average Method (Section 3.3)*

	Pitch Drop Method	Proposed Average Method
Precision	0.591	**0.970**
Recall	0.354	**0.585**
F-measure	0.443	**0.730**
Accuracy	0.520	**0.766**

Tables 1 shows that the proposed method using average pitch is more accurate than the pitch drop method for recognizing accents. In accent evaluation, we considered it important that "the system is error-free in its detection." In other words, we considered precision, which reflects "the number of errors in the recognition results," to be important. From this perspective, Proposed average method shows that the precision was >0.95, which indicates sufficient accuracy.

4. Dictogloss Learning Support System with Accent Recognition

We reimplemented a dictogloss system (Kogure et al., 2017) capable of learning both Japanese and English. In the feedback phase of this system, we implemented a feature that allowed the output of the feedback images generated from the accent recognition results described in Section 3. The accent recognition server asked the learner to utter the keywords set in the lesson selected by the Japanese dictogloss learning support system, recognizes speech for the uttered words, detected the falling and rising edges of the fundamental frequency, and estimates the accent graph. Specifically, when the feedback phase was executed in the dictogloss system, the following procedure was used to perform word accent recognition and output the results as images:

The system sends "information about words to be learned" from the dictogloss system to the accent identification server. Then, the accent identification server sends a trigger to the speech recording server to record the learner's speech. The accent identification server recognizes the words in the recorded speech and determines whether the learner is speaking the target word. If the server determines that the learner is not speaking, it points to the system individually. If the server determines that the target word has been spoken correctly, it proceeds to the next step. The accent identification server sends the correct syllable sequence and the speech file to the "syllable segmentation server" to obtain the correspondence between the speech frame and the mora. The accent identification server generates *TextGrid* information from the correspondence. *TextGrid* is a framework for mapping syllables and phonemes to frames in *Praat* based on multiple criteria. The accent identification server measures the fundamental frequency of each frame using *Praat*. Using the results and mora correspondences, the accent recognition server performs word accent recognition using *Praat*. The accent identification server uses *Praat* to generate an image containing the speech waveform, correct syllable sequence, recognized accent pitch graph, and correct accent pitch graph. The accent identification server sends the file path of the generated image to the dictogloss system. The dictogloss system displays an image of the received image file path in a feedback window.

Figure 2 shows the feedback image displayed on the system when the user utters "ki ma tsu shi ke N". The system displays the waveform of the recorded speech and the pitch graph recognized by *Praat*. The system displays the recognized accent graph in blue and the correct accent graph in red.

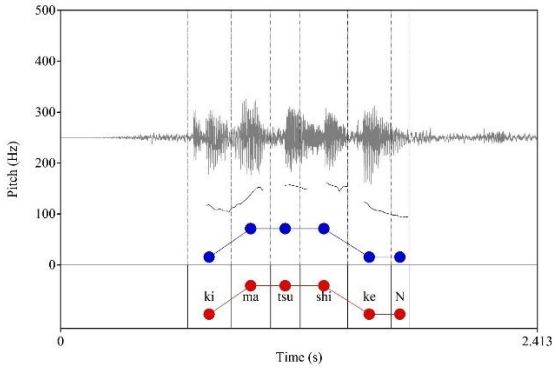

Figure 2. Feedback Images of Word Accent Recognition Result on the Feedback Phase on Dictogloss System.

5. Conclusion

We propose a word accent recognition method based on the relationship between the average pitch of an entire word and the average pitch of each mora unit. The experimental evaluation showed that the proposed method outperformed the pitch-drop detection method. Furthermore, we developed a Japanese dictogloss learning support system that provides feedback on word accent recognition. The learner spoke the target word in the microphone. The system presents the learner with an image on a word accent recognition feedback screen that simultaneously includes the speech waveform, pitch graph, recognized word accent graph, and correct word accent graph.

In the future, we plan to develop a framework for recognizing the accents of collocations and intonation of sentences and implement the system.

Acknowledgements

This work was supported by JSPS KAKENHI Grant Number JP21K12172, JP24K15205.

References

Boersma, P., & Weenink, D. (1992-2024) Praat: Doing phonetics by computer [Computer program]. Version 6.2.06, retrieved 19 March 2024 from https://www.fon.hum.uva.nl/praat/.

Kitamura, T., Amakawa, Y., & Hatano, H. (2019). Occurrence condition s of delayed fundamental frequency fall in Japanese word of Tokyo dialect speakers, *Speech Study, Journal of the Phonetic Society of Japan, 23*, 165-173 (in Japanese).

Kogure, S., Okugawa, K., Noguchi, Y., Konishi, T., Kondo, M., & Itoh, Y. (2017). Improvement of the situational dialog function and development of learning materials for a Japanese dictogloss environment, *Proceedings of ICCE2017*, 104-106.

Kogure, S., Hakamata, H., Noguchi, Y., Konishi, T., Kondo, M., & Itoh, Y. (2020). Development of Japanese dictogloss learning support environment for pronunciation learning of Japanese speech, *Proceedings of ICCE2020*, 261-263.

Mikhailava V, Lesnichaia M, Bogach N, Lezhenin I, Blake J, & Pyshkin E. (2022). Language accent detection with CNN using sparse data from a crowd-sourced speech archive. *Mathematics*. *10*(16):2913, https://doi.org/10.3390/math10162913.

Minematsu, N., Kobayashi, S., Shimizu, S., & Hirose, K. (2012). Improved prediction of Japanese word accent sandhi using CRF. *Proceedings of INTERSPEECH 2012*.

Parikh, P., Velhal, K., Potdar, S., Sikligar, A., & Karani, R. (2020). English language accent classification and conversion using machine learning, *Proceedings of the International Conference on Innovative Computing & Communications (ICICC) 2020,* http://dx.doi.org/10.2139/ssrn.3600748.