Development of a web framework for teaching and learning Japanese prosody: OJAD (Online Japanese Accent Dictionary)

Ibuki Nakamura†, Nobuaki Minematsu†, Masayuki Suzuki†, Hiroko Hirano‡, Chieko Nakagawa∗
Noriko Nakamura⋆, Yukinori Tagawa∗, Keikichi Hirose†, Hiroya Hashimoto†

† The University of Tokyo, Tokyo, Japan
‡ Northeast Normal University, Jilin, China
∗ Waseda University, Tokyo, Japan
⋆ Tokyo University of Foreign Studies, Tokyo, Japan

{nakamura,suzuki,mine,hiroya,hirose}@gavo.t.u-tokyo.ac.jp

10.21437/Interspeech.2013-575

Abstract

This paper introduces the first online and free framework for teaching and learning Japanese prosody including word accent and phrase intonation. This framework is called OJAD (Online Japanese Accent Dictionary) [1] and it provides three functions. 1) Visual, auditory, systematic, and comprehensive illustration of patterns of accent change (accent sandhi) of verbs and adjectives. Here only the changes caused by twelve kinds of fundamental conjugation are focused upon. 2) Visual illustration of the accent pattern of a given verbal expression, which is a combination of a verb and its postpositional auxiliary words. 3) Visual illustration of the pitch pattern of an any given sentence and the expected positions of accent nuclei in the sentence. The third function is implemented by using an accent change prediction module that we developed for Japanese text-to-speech (TTS) synthesizers [2, 3]. Experiments show that accent nucleus assignment to given texts by the proposed framework is much more accurate than that by native speakers. Subjective assessment and objective assessment by teachers and learners show very high pedagogical effectiveness of the framework.
Index Terms: language education, Japanese prosody, accent sandhi, OJAD, speech synthesis, assessment experiments

1. Introduction

It is often said that a language learner has to acquire good skills of reading, writing, speaking, and listening of the target language if he/she wants to be a good user of that language. Each target language has its own difficulties of learning, which may also depend on the first language of learners. For example, Japanese has three kinds of writing systems, KANJI (Chinese characters), KATAKANA, and HIRAGANA and this fact often increases the difficulty for learners to acquire good writing skills in Japanese in an adequate manner. Teaching and learning of Japanese pronunciation has its unique problems for teachers and learners. One of the problems is Japanese prosody. Japanese word accent is mora-based pitch accent but many learners, who are learning Japanese in colleges or universities in Japan, don't know this fact because it is rarely taught in a classroom [4]. This is partly because of time constraints but we can say there are other reasons. One of these reasons is the fact that Japanese word accent is easily and variously changed due to its context [5]. It should take time to teach the complicated mechanism of accent change. Another reason is teachers' awareness of the mora-based pitch attribute. Japanese phonology explains that each mora in a given sentence has its pitch attribute of high or low. The pitch attribute of a mora is determined by the word which includes the mora and the context which surrounds the word [5]. However, native Japanese are not always good at labeling each mora as H/L in a given utterance although they can speak Japanese fluently. For them, word accent control is so automatic that they are not required to be aware of the H/L value of each mora in conversation. The Japanese three writing systems do not require writers to mark the pitch attribute visually. This fact can also be considered another reason of native Japanese's low awareness of the pitch attribute[1]. This is the case with native teachers of Japanese, not a small number of whom are not good at pitch labeling. In this situation, it is difficult to teach word accent control in reading sentences.

In Japanese, regional accents are often exposed as differences of word accent control [6]. In the Japanese society, public speaking is almost always done in Tokyo dialect, which is referred to as the standard Japanese. Many learners of Japanese are from other Asian countries and their mother tongue is often a tonal language[2]. Of course, they are much more sensitive to and aware of syllable-based pitch control than normal Japanese. In the current situation of Japanese education, even when learners are eager to know how to control word accent in the standard Japanese, they have to fail in finding good learning materials because no good teaching/learning material of Japanese word accent control exists to the best of the authors' knowledge.

In this paper, we attempt to solve this serious problem by providing the very first online and free framework for teaching and learning Japanese prosody including word accent and phrase intonation. In development, we use an accent sandhi prediction module [2, 3] that we developed for Japanese TTS systems to visualize the prosodic structure hidden in a given sentence. Although TTS technologies have been used in CALL development in previous studies [7, 8], a main focus was always put on how to use synthesized speech output. In this paper, we don't use speech output but use the internal module of a TTS system as visualizer of the hidden prosodic structure in a given sentence because this prosodic structure is what learners want to know. It should be noted that direct visualization of the output of the accent sandhi module is not good pedagogically. The output of the module is often too complicated for learners to learn. For them, as simple prosodic structure as possible with naturalness enough is desired. In tight collaboration with Japanese teachers, the forth to the seventh authors, we design simplification rules to derive the simplified prosodic structure [9]. This paper also shows some experimental results of subjective and objective assessment of the proposed framework.

[1] This performance can be compared to low phonological awareness of dyslexic individuals.

[2] Japanese is not a tonal language.

2. Realization of the three functions

2.1. Comprehensive illustration of accent changes

Since the accent changes found in conjugation of verbs and adjectives is relatively regular and systematic, we decided to realize a system that can show the accent changes due to conjugation of these words. Users type verbs and/or adjectives of interest to know their accent changes. Here, twelve kinds of fundamental conjugation were adopted and their accents are displayed in a table. Fig. 1 shows an example. Seven widely-used textbooks were selected and all the verbs and adjectives found in them were manually extracted. The total number of words is about 3,500 and that of their conjugated forms is about 42,000. The accent pattern of each form was obtained as follows. 1) Automatic estimation of the accent pattern of the form by using an accent nucleus position estimator [2, 3] and 2) manual inspection of the results and manual correction if needed. The resulting accent patterns were stored in a database. It should be noted that the system does not estimate the accent patterns of the conjugated forms of an input word on the fly but searches the database for the accent patterns. We can say that the system is effectively error-free unless users commit typing errors.

The pitch curve of each form is drawn on its HIRAGANA representation by using the generation process model, so called as the Fujisaki model [10]. By controlling the model parameters, it is easy to realize a pitch pattern of complete mora isochrony. The drawn pattern is of course not realistic but what has to be presented to users is not observed pitch patterns but the pitch pattern "images" that teachers want to show to learners. This is the reason why we adopted the Fujisaki model.

Each of these forms was read aloud by a voice actor and a voice actress. About 84,000 speech samples were recorded and they were segmented semi-automatically using voice activity detection techniques. In Fig. 1, by clicking a blue/pink icon, users can listen to a male/female speech sample of each form, respectively. A series of the samples on a row or on a column can be heard by clicking an icon of that purpose. These samples can be downloaded onto users' PCs or portable devices.

Since all the words are directly from the textbooks, we implemented a very flexible user interface to search the database. Instead of typing specific verbs or adjectives, users can indicate a specific lesson of a specific textbook to know the accent patterns of the conjugated forms of all the verbs and adjectives that are introduced to that textbook for the first time in that lesson. Each word is assigned "difficulty level to learn" as attribute, which is from another Japanese word database developed at University of Tsukuba [11]. Using this attribute, for example, users can know the accent patterns of all the verbs and adjectives of a specific textbook that beginners should learn. Other useful options are available for practical use in a classroom. Interested readers should visit the web site of OJAD [1].

2.2. Illustration of the accent of long verbal expressions

The first function only illustrates the accent patterns of the twelve fundamental conjugated forms of verbs and adjectives. Since Japanese is an agglutinative language, a verb can be combined with multiple postpositional and auxiliary words. For example, verb "断る" (refuse) can be concatenated to "そう", "に", "なる", "た", "こと", "が", and "ある" in this order. By conjugation, "断る" is finally changed into "断りそうになったことがある" and this kind of long verbal expressions can be found even in a textbook for beginners. This means that only the first function cannot explain the accent control to read verbal expressions in that textbook completely. So, we developed another system as the second function to show the accent pattern of a given long verbal expression. By inspecting a Japanese textbook for beginners manually, we found 320 kinds of postpositional expressions combined to verbs. Then, we checked the possible accent patterns for each of the postpositional expressions. Japanese verbs can be clustered into two groups (起伏式 and 平板式) based on their word accent and into three groups (第 1 グループ to 第 3 グループ) based on their conjugation manner. If the accent-based group and the conjugation-based group of a given verb is known, the accent pattern of any verbal expression comprised of that verb and one of the 320 postpositional expressions can be correctly predicted. In the system, by running morphological analysis to an input verbal expression with MeCab [12], the verb and its postpositional expression are detected automatically. Using the attributes of the verb estimated by MeCab, the system can identify the accent-based group and the conjugation-based group for that verb. Then, the resulting accent pattern (the accent nucleus position) is visually presented with its pitch pattern drawn by the Fujisaki model.

Figure 1: Illustration of the accent patterns of conjugated forms

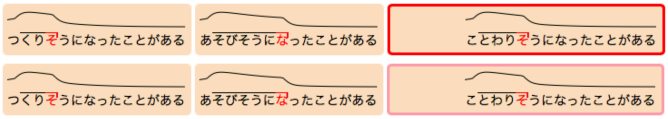

Figure 2: Illustration of the pitch pattern of long expressions

Fig. 2 shows several examples of illustrating the pitch pattern of a long verbal expression. It is easily expected that a verbal expression including an unknown postpositional expression can be typed as input. Our database contains the information of the 320 expressions only. The top right figure shows the result of typing "断りそうになったことがある" and the bottom right one shows that of typing "断りそうになったことがあるのだが", where "のだが" is added to the first query. If the system can find the postpositional expression in the database, it shows the accent pattern in a red rectangle. If the system cannot, however, it shows the accent pattern of the most similar expression, which is found in the database, in a pink rectangle. Availability of the system response is indicated by color.

2.3. Illustration of the pitch pattern of any input sentence

The first and second functions only focus upon verbs and adjectives. Word accent changes are not only found in these words but also in other words such as nouns. This means that the first two functions are not enough for learners to read sentences adequately in textbooks. So, as the third function, we developed

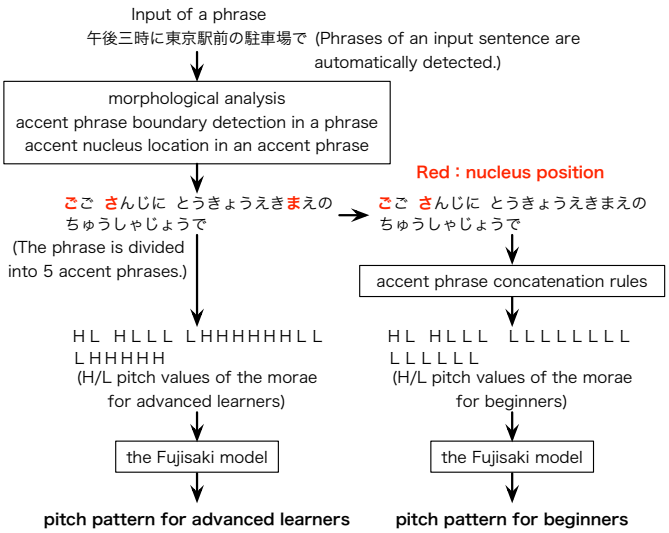

Figure 3: Generation of original and simplified pitch patterns

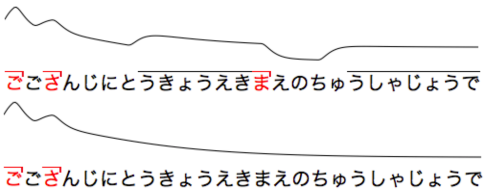

Figure 4: Original and simplified pitch patterns

a prosodic reading tutor to support learners by presenting the pitch pattern of an any given sentence, which is expected to be observed when a native speaker reads that sentence neutrally.

This function can be realized easily by using several internal modules developed for TTS synthesizers. They are morphological analysis (linguistic analysis) [12], accent phrase boundary detection from text [2, 3], accent nucleus location for an detected accent phrase [2, 3], and pitch pattern visualization by the Fujisaki model [10]. As mentioned in Sect. 1, direct visualization of the output of these modules is not good pedagogically because it is sometimes too complicated for learners to learn. As simple pitch patterns as possible with naturalness enough are desired to be presented. In tight collaboration with Japanese teachers, we designed simplification rules [9]. Generation of the original and simplified pitch patterns is schematized in Fig. 3.

An input sentence is automatically divided into phrases by punctuation marks and phrase boundary marks (/), which are explicitly given by users. Three analyses of morphological analysis, accent phrase boundary detection, and accent nucleus location are run for each phrase. Then, the input phrase is divided into multiple accent phrases, in each of which the accent nucleus position is estimated automatically. For advanced learners, these accent phrases are directly used to visualize the pitch pattern. An example is shown at the top of Fig. 4.

For beginners, simplification is needed. In some accent phrases, there is no accent nucleus. [9] claims that learners should focus on the first nucleus found in a phrase (not an accent phrase). The remaining accent nuclei found in that phrase after the first nucleus can be ignored to generate a very simple pitch pattern with sufficient naturalness. In this paper, we adopted this strategy but modified it considering Japanese listeners' perceptual characteristics. It is shown in [13] that Japanese listeners are very sensitive to the accent nucleus when it is found at the first mora of an accent phrase. So, our simplification rules remove the accent nuclei other than the first accent nucleus in a phrase and the accent nuclei found at the first mora of an accent phrase which is longer than three morae. After removal, the H/L sequences of the multiple accent phrases are concatenated to generate the H/L sequence for the entire phrase. Using the resulting H/L sequence, our simplified pitch pattern is generated. An example is shown at the bottom of Fig. 4.

It should be noted that the first two systems are effectively error-free but the last one sometimes show incorrect pitch patterns due to on-line analysis errors. They can sometimes become serious because learners may not be able to judge whether the presented information is correct or not. Considering this, we carried out only subjective assessment for the first two systems and both subjective and objective assessment for the last one.

Table 1: Assessment of the three proposed systems (%)

a) How useful do you think the system is to learners?

	1st system	2nd system	3rd system
Very useful	71.0	54.8	62.7
Rather useful	29.0	45.2	28.8
Not so useful	0.0	0.0	8.5
Not useful at all	0.0	0.0	0.0

b) Do you want to use the system in your class?

	1st system	2nd system	3rd system
Yes, definitely	38.7	29.0	42.6
Yes, if needed	59.7	64.5	50.0
No	1.6	6.5	7.4

3. Assessment of the proposed systems

3.1. Subjective assessment of all the proposed systems

We prepared an introductory web page of "Let's use OJAD for accent training!" for the subsequent subjective assessment. In this page, the functions of the proposed three systems were explained and some example exercises for accent training were provided. The fact that the prosodic reading tutor sometimes show incorrect pitch patterns was also explicitly explained using some erroneous responses from the system.

We asked teachers of Japanese to join the subjective assessment test after learning how to use OJAD in the above page. Eighty teachers joined the test, two thirds of whom were teaching Japanese outside Japan. Although the subjective assessment was composed of a series of questionnaires, due to space limit, we show the results of only two fundamental questionnaires: a) How useful do you think the system is to learners? and b) Do you want to use the system in your class?

Results of the two questionnaires are shown in Tab. 1 in the form of percentage. Considering that teaching Japanese prosody is just only one aspect of Japanese language education, we consider that the eighty teachers of Japanese recognize very high pedagogical effectiveness of the proposed framework.

3.2. Objective assessment of the prosodic reading tutor

In many Japanese classes, public speaking is introduced in their syllabus. Here, when learners want to speak in the standard Japanese, it is often true that they always ask teachers to locate accent nuclei in their manuscript. In objective assessment here, we give learners a task of accent nucleus location in several Japanese paragraphs using the prosodic reading tutor and two other facilities available currently. The two facilities are 1) a PC-based word accent dictionary [14] and 2) a PC-based

commercial Japanese speech synthesizer [15]. Since the accent dictionary is widely used in Japanese education, we compared the following three conditions: a) only with the word accent dictionary, b) with the dictionary and the synthesizer, and c) with the dictionary and our prosodic reading tutor.

The word accent dictionary only shows the accent pattern of an isolatedly pronounced word. So, its usefulness in this task is expected to be low. The speech synthesizer can present the pitch pattern of an any input sentence as auditory stimulus. The difference between b) and c) lies basically in the mode of presentation, auditory or visual. The synthesizer sometimes present incorrect pitch patterns, similarly to the tutor. The objective assessment test was done after explaining the limitation of each facility. The dictionary contains only the word accent of isolatedly pronounced words and both the synthesizer and the tutor sometimes present incorrect pitch patterns.

Four paragraphs, p0 to p3, were prepared, which were judged by four Japanese teachers to belong to the same reading difficulty level. Manual phrase segmentation was done for each paragraph by the four teachers and phrase boundaries were explicitly shown to the subjects as punctuation mark or phrase boundary mark in the paragraph. The numbers of phrases are 73, 68, 73, and 70 for p0 to p3, respectively. The actual task given to the subjects is locating the first accent nucleus in each phrase. Fig. 5 show an example of the paragraph used and an example of the PC desktop image in the experiment.

The subjects were 36 learners of Japanese who had fundamental knowledge of Japanese word accent. p0 was presented to all the subjects without any facility to know their original performance. The number of facilities is three and we have other three paragraphs. Considering the ordering effect, we can prepare 36 different combinations of the paragraphs and the facilities. The 36 subjects were used to cover all the combinations.

Results of location were compared to the correct positions prepared by the four teachers. The original performance of the subjects was 68.2%. The same task of p0 was given to ten university students who are native speakers of Tokyo dialect. Their performance was 61.6%, lower than that of the subjects. As described in Sect. 1, this result is reasonable. Native Japanese speak the standard Japanese very fluently but they are not good at locating accent nuclei consciously. The first languages of 27 subjects out of 36 were tonal languages. The performance of the reading tutor and the synthesizer is 93.2% and 95.9%, respectively, which is much higher than that of native speakers.

Results of accent nucleus location using p1 to p3 are shown in Fig. 6. In the experiment, all the mouse clicks were monitored and recorded in history files. Using the files, the speed of location and the precision of location can be compared among the three kinds of facilities of a) to c). In Fig. 6, the x-axis indicates the duration of elapsed time and the y-axis means the numbers of effective mouse clicks (answers) in the top and the rates of correct answers in the bottom. The top figure shows the speed of location and the bottom figure shows the precision of location by a function of elapsed time. No significant difference is found among the three facilities in the top figure, indicating that the tutor was unexpectedly ineffective to reduce the time required for accent nucleus location. This is because of the prior knowledge on incomplete performance of the tutor and it seemed that the subjects used the tutor very carefully. On the other hand in the bottom figure, the tutor is found to be significantly effective to increase the precision of accent location. It is also found that the synthesizer is also ineffective in this figure.

It is interesting that the performance in c) (84.8%) is lower than the original performance of the tutor for p1 to p3 (91.0%)

Figure 5: Examples of the paragraph and the PC desktop

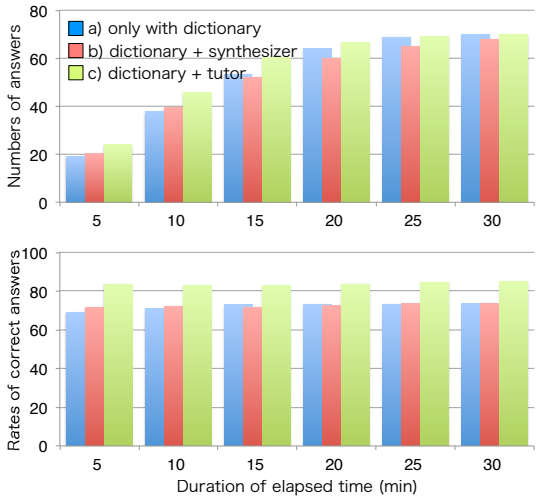

Figure 6: Results of objective assessment

and that of the synthesizer (91.3%). We can say that the subjects' judgments always revise the tutors' suggestions for the worse. As implemented in Sect. 2.2, availability of the system responses should be given to learners as useful information.

After the experiments, we asked the subjects how useful each facility was. 37.5%, 30.0%, and 82.5% of them said that a), b), and c) were "very useful", respectively.

4. Conclusions

In this paper, we developed the very first online and free framework for teaching and learning Japanese prosody including word accent and phrase intonation. Both subjective and objective assessment experiments showed very high pedagogical effectiveness of the proposed framework.

This framework, called OJAD, was released to the public in August 2012 in an international conference of Japanese education [16]. After that, by using Google Analytics, all the accesses to OJAD have been recorded in a history file. The file shows that the number of accesses is around 16,000 as of March 2013 and about half of them are from outside Japan. Considering a fact that 72% of teachers of Japanese are non-native outside Japan [17], we can say that not a small number of learners are using OJAD as the only information source to know how to control word accent when trying to speak the standard Japanese.

5. References

[1] OJAD: http://www.gavo.t.u-tokyo.ac.jp/ojad/

[2] N. Minematsu, S. Kobayashi, S. Shimizu, K. Hirose, "Improved prediction of Japanese word accent sandhi using CRF," *Proc. INTERSPEECH*, CD-ROM (2012)

[3] M. Suzuki, R. Kuroiwa, K. Innami, S. Kobayashi, S. Shimizu, N. Minematsu, K. Hirose, "Accent sandhi estimation of Tokyo dialect of Japanese using conditional random fields," Trans IEICE, J96-D, 3, 655–654 (2013 in Japanese)

[4] R. A, R. Hayashi "The effect of shadowing training for Mongolian and Chinese learners of Japanese", *IEICE Technical Report*, SP2009-151, 19–24 (2010 in Japanese)

[5] Y. Sagisaka, H.Sato, "Accentuation rules for Japanese word concatenation," *Trans. IEICE Jpn.*, 66D, 7, 849–856 (1983 in Japanese)

[6] Z. Uwano, "Word accents of Japanese," in series of *Japanese and Japanese Education*, published by Meiji-Shoin (1989 in Japanese)

[7] A. Black, "Speech synthesis for educational technology," *Proc. SLaTE*, CD-ROM (2007)

[8] M. Eskenazi, "An overview of spoken language technology for education," *Speech Communication*, 11, 51, 832–844 (2009)

[9] C. Nakagawa, N. Nakamura, S. Ho, *Japanese pronunciation drills for advanced oral presentation,* published by Hitsuji-Shobo (2009 in Japanese)

[10] H. Fujisaki, K. Hirose, "Analysis of voice fundamental frequency contours for declarative sentences of Japanese," J. Acoust. Soc. Japan (E), 5, 4, 233–242 (1984)

[11] http://jisho.jpn.org

[12] MeCab: http://mecab.sourceforge.net/

[13] N. Minematsu, K. Hirose, "Role of prosodic features in the human process of perceiving spoken words and sentences in Japanese," J. Acoust. Soc. Japan(E), 16, 5, 311–320 (1995)

[14] NHK Japanese word accent dictionary, published by NHK (1998)

[15] HOYA service: http://voicetext.jp

[16] N. Minematsu, M. Suzuki, H. Hirano, C. Nakagawa, N. Nakamura, Y. Tagawa, K. Hirose, "Development of an online Japanese accent dictionary with speech output facility," Proc. Int. Conf. on Japanese Language Education (ICJLE), pp.94 (2012)

[17] *The statistical situation of Japanese education in foreign countries,* reported by JAPAN FOUNDATION (2009) http://www.jpf.go.jp/j/japanese/survey/result/dl/survey_2009/2009-05.pdf