# OJAD: Web-based Prosodic Reading Tutor of Japanese

Nobuaki Minematsu

Graduate School of Engineering, the University of Tokyo, Japan
*mine@gavo.t.u-tokyo.ac.jp*

## Abstract

Learning prosodic control for speaking Japanese is effective to reduce syntactic and lexical ambiguity and to improve the comprehensibility of learners' spoken Japanese. Good prosody can also improve its naturalness. In the conventional curriculum however, prosody training has not been provided satisfactorily for learners partly because of scarcity of teaching materials. To facilitate prosody training with speech technologies, we developed a web-based system called OJAD (Online Japanese Accent Dictionary) [1] and in OJAD, a prosodic reading tutor of Japanese, Suzuki-kun, is provided. In the current paper, technical development of Suzuki-kun and its effectiveness is described. Experiments show that Suzuki-kun's visualized prosody can help learners more effectively than model utterances.

**Keywords:** OJAD, Suzuki-kun, Japanese prosody, improvement of naturalness

## 1. Introduction

Learners attempt to acquire good speaking skills in the target language, and every target language has its own unique difficulties. This is also the case with Japanese and one of the main problems in learning how to speak natural Japanese is prosody. Every content word in Japanese has its own lexical accent[1], much like English. It has mora-based pitch accent and binary values (High/Low) are assigned to each mora. This means logically that $2^N$ H/L sequences are possible for an N-mora word but, in Tokyo Japanese, only N sequences are allowed as lexical accents, which are called accent types. Namely, the lexical accent of a particular N-mora word is one of those N types. The four accent types of four-mora words are illustrated in Figure 1. Many learners, however, don't know this fact because it is not always taught in class [2-4]. The current situation of teaching Japanese word accents is reported in [5].

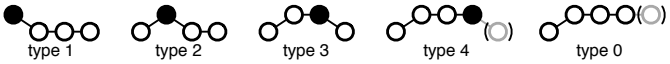


**Figure 1:** The four accent types of four-mora words in Japanese. A filled circle is an accent nucleus, which is the mora position immediately before a rapid and local pitch downfall.

When a native speaker speaks, accent control is often achieved not by a unit of word but by a unit of phrase (i.e. the accentual phrase) [6]. Several examples are shown in Figure 2. If an accentual phrase has M morae, it has one of the M accent types. In other words, that phrase is pronounced in a similar way to an M-mora word in terms of accent control. This fundamental mechanism of speech production in Japanese is also taught rarely to learners [7] and it is not explained explicitly even in teachers' training programs. If a sentence is read aloud by using the lexical accent of its constituent words as it is, in most cases, the read sentence will result in including inadequate control of pitch. This is because the unit of accent control is often not a word but a phrase for reading sentences.

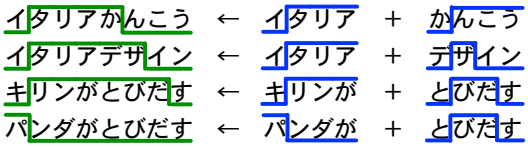


**Figure 2:** Not word-based but phrase-based control of accent The third and forth examples can be read with two accents.

It is very natural that the accent type of a word is different between when it is spoken in isolation and when spoken in context. Native speakers acquire context-dependent accent control implicitly and when they speak, they change word accent almost unconsciously. This is why accent awareness of native speakers, even native teachers, is generally not high although they are still sensitive to inadequate accent control exhibited by learners [8]. Further, accent control varies among dialects [9]. It is not uncommon that native teachers whose native dialect is not Tokyo Japanese are unconfident in teaching accent control. It is true, however, that Tokyo Japanese is the common dialect and is said to be the "dress code" of Japanese [10], which is often used in business or in public. Since many learners are learning Japanese for business, a good infrastructure for learning the accent control of Tokyo Japanese has been requested from those learners.



10.21437/ISAPh.2016-3

The function of phrase-based accent control refers to grouping words of a phrase into one accent unit or chunk[2]. In Japanese, another grouping mechanism is also present. It is termed the intonational phrase, which is composed of one or more accentual phrases. In other words, several accentual phrases of an intonational phrase make up one intonation unit or chunk. Between consecutive intonation units, a pause is often inserted and when the utterance is transcribed, a punctuation mark is often placed there. Figure 3 shows the conceptual and hierarchical structure of Japanese prosody embodied when reading a sentence. An example of integrating accent control and intonation control is also illustrated. If additional pauses are inserted at inadequate word boundaries due to non-nativeness, the comprehensibility of that read sentence is easily degraded [12]. That is to say that, with those pauses inserted, it takes a longer time for native listeners to comprehend what is said. Generally speaking, however, the relation between intonation control and comprehensibility is not explained explicitly in class. Comprehensibility differences between before and after intonation training can be checked in [13].

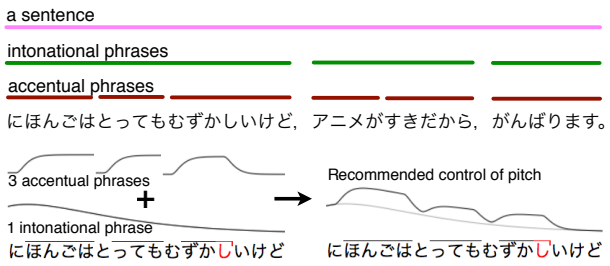


**Figure 3:** Hierarchical and prosodic structure of a read-aloud sentence

It seems that teachers have not explained Japanese prosodic control sufficiently and satisfactorily to human learners. However, we can claim that engineers have explained it very intensively to machine learners for some decades. Speech technologies used to enable a machine to read a given text aloud are called speech synthesis or Text-To-Speech (TTS) technologies. The aim of Japanese TTS is generally the conversion of input text into spoken *Tokyo* Japanese with good naturalness and comprehensibility. To realize this, engineers have implemented programs in machines with adequate prosodic knowledge. If those programs are not implemented, Japanese customers will reject synthesized voices because the voices are somewhat different from the "dress code" of Japanese. As is well known, Japanese are strict regarding technologies, but they seem to be lenient or indulgent to human learners.

In some of our previous studies [14-17], several modules or techniques embedded in Japanese TTS systems were transferred to educational software. In 2012, we launched OJAD [14,15], a web-based system to learn word accent and its control. Following this, in 2014, we added a prosodic reading tutor module to OJAD, Suzuki-kun [16,17]. For any given text, it can predict the full and hierarchical control of accent and intonation required to read that text in Tokyo Japanese and illustrate the pitch control using Fujisaki's model [18]. Further, the visualized prosody is realized as audio with a synthetic voice using an HMM-based speech synthesizer. We can find several previous studies [19-21], where pitch contours of learners' utterances and teachers' utterances were plotted for visual comparison. Suzuki-kun was built with a different objective in mind, not for comparison between teachers and learners but for guidance to read given sentences in Tokyo Japanese.

Suzuki-kun can illustrate the full and hierarchical prosodic control of any given text and generate synthetic speech based on the visualized prosody. In the long history of Japanese education, Suzuki-kun is the first and currently only prosodic reading tutor with this function. OJAD has been already translated into 14 non-Japanese languages and is now used internationally. As for the effectiveness of OJAD, however, we examined users' satisfaction only with questionnaires [14,15]. In this article, we quantitatively investigate the effectiveness of Suzuki-kun on the improvement of prosodic naturalness by using sixty-four Chinese learners.

## 2. Development of Suzuki-kun

Over the past decade, the quality of synthetic voices has been drastically improved and we can find some cases where these voices are being used as model utterances in a language class [22]. Generally speaking, a TTS engine does not read input text directly but processes its corresponding phoneme sequence with various prosodic symbols attached by a prosody prediction module.

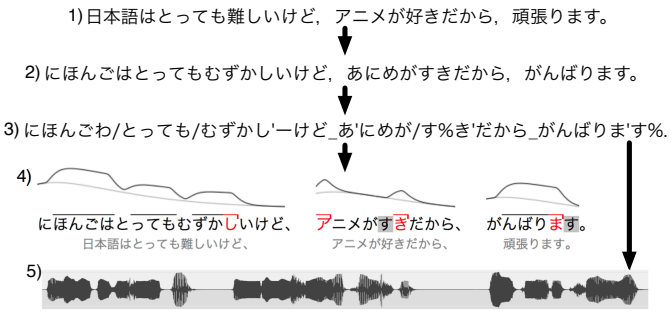


**Figure 4:** Prediction and visualization of prosodic features for given text and generation of its waveform with an HMM-based speech synthesizer

Figure 4 shows 1) an original Japanese text, 2) its phonemic transcript as a Hiragana sequence, 3) out-

put from a prosody prediction module that we have developed [23], and 4) visual output from Suzuki-kun. In 3), the prosodic features are predicted only from the input text and assigned to the text as symbols. ‘’’ represents an accent nucleus. ‘/’ and ‘_’ indicate an accentual phrase boundary without a pause and that with a pause, respectively. The latter also functions as an intonational phrase boundary. This symbolic representation is defined by JEITA (Japan Electronics and Information Technology Industries Association) [24] and is widely used in the Japanese community of TTS synthesis. 3) includes a full description of the hierarchical structure of prosody required to read this text naturally. It claims that this sentence should be divided into three intonational phrases and that these intonational phrases contain three, two, and one accentual phrase(s) from the head to the end of the sentence, respectively. This prosodic hierarchy is illustrated in 4) by using Fujisaki’s model. The values of its model parameters are determined by applying a set of rules to 3), which were derived through collaboration among teachers and engineers [14,15].

Further, ‘%’ in 3) is an unvoicing operator and a mora with an unvoiced vowel in 4) is drawn in a gray patch. Acoustically speaking, unvoiced sounds do not have fundamental frequencies but, educationally speaking, the pitch pattern should be drawn even over such segments and it can be drawn easily with Fujisaki’s model. In other words, the pitch pattern in 4) is *not* a pitch pattern that is expected to be observed acoustically when this sentence is read by a native speaker. It is a pitch pattern that should be used as *mental target* in prosody training to read this sentence. Suzuki-kun can produce synthetic voices with four different speaker identities (2 males and 2 females) for any given text and the voices are obtained by running four different HMM-based synthesizers, one for each voice. Here, 3) is input into the synthesizers and the pitch pattern drawn by Fujisaki’s model is not used for synthesis. If the pitch patterns are acoustically extracted from these voices, they will show some differences from Fujisaki’s pattern. We can consider these deviations to still fall in the distribution of natural speech. This is because acoustically observed pitch patterns will even vary from speaker to speaker.

## 3. Effects of Suzuki-kun

### 3.1. Objective of the assessment experiments

We can find some textbooks that have an audio CD as attachment of all the sentences contained in the textbooks. A few of those textbooks, [12] for example, illustrate the pitch patterns of all the sentences. What is unique to Suzuki-kun is that it can illustrate the full and hierarchical prosodic control, accent and intonation, for any given text and can provide a reading of that text based on the visualized prosody. If only conventional textbooks are available, learners have to guess very complicated rules for Japanese prosodic control to read new texts with natural-sounding prosody. This often causes their reading to become accented. By using Suzuki-kun, learners can understand the prosodic control instantly and they may be able to acquire the appropriate rules in an implicit manner by reading many new texts under the guidance of Suzuki-kun.

With a focus on these distinctive characteristics of Suzuki-kun, in this article we conduct experiments to compare the improvement of prosodic control in terms of naturalness that are realized under the following conditions: 1) practice with auditory prosody only (use of synthetic voices), 2) that with visualized prosody only, and 3) that with both auditory and visualized prosody [25]. For comparison, learners’ reading only with text, which is referred to as condition 0 hereafter, is also collected. It should be noted that no attention is paid to how long the improvement in naturalness due to Suzuki-kun’s help lasts. This is because Suzuki-kun’s primary advantage is its instant instructions on prosody and we believe that long-term acquisition of prosodic control will depend on continued use of Suzuki-kun for reading new texts. In the experiments below, recording was always done after a short five-minute self-practice in each condition.

### 3.2. Subjects and texts used for the experiments

A fair comparison of learners’ reading performance under different conditions is sometimes difficult. This is because different groups of learners with as close to the same proficiency level as possible have to be selected for different conditions. Similarly, different sets of text with as close to the same difficulty level as possible have to be prepared for these different conditions. In this study, the L1 of the learners was controlled to make fair comparison even easier. Since the largest number of learners of Japanese outside of Japan is found in China [26], we asked Chinese learners who had learned Japanese for more than one year but less than two years to participate in our experiments. Sixty-four learners joined the experiments and they were recruited at Oikawa’s Summer Camp for speech training [27].

As for passage selection, two passages were adopted, text-1 and text-2, both from the same unit of a textbook, which is written for a second semester Japanese course. The passages had 197 and 196 characters respectively and their content covered drinking coffee. Another short passage was also prepared only to explain how Suzuki-kun illustrates

**Table 1:** The experimental design using the six groups of learners

| | A (11) | B (7) | C (15) | D (10) | E (11) | F (10) |
|---|---|---|---|---|---|---|
| session 1 | condition-0<br>**text only**<br>practice for **text-1** | condition-0<br>**text only**<br>practice for **text-2** | condition-0<br>**text only**<br>practice for **text-1** | condition-0<br>**text only**<br>practice for **text-2** | condition-0<br>**text only**<br>practice for **text-1** | condition-0<br>**text only**<br>practice for **text-2** |
| | recording | recording | recording | recording | recording | recording |
| session 2 | condition-1<br>**text + auditory**<br>practice for **text-1** | condition-1<br>**text + auditory**<br>practice for **text-2** | condition-2<br>**text + visual**<br>practice for **text-1** | condition-2<br>**text + visual**<br>practice for **text-2** | condition-3<br>**text + both**<br>practice for **text-1** | condition-3<br>**text + both**<br>practice for **text-2** |
| | recording | recording | recording | recording | recording | recording |
| session 3 | condition-3<br>**text + both**<br>practice for **text-1** | condition-3<br>**text + both**<br>practice for **text-2** | condition-3<br>**text + both**<br>practice for **text-1** | condition-3<br>**text + both**<br>practice for **text-2** | condition-3<br>**text + both**<br>practice for **text-1** | condition-3<br>**text + both**<br>practice for **text-2** |
| | recording | recording | recording | recording | recording | recording |
| session 4 | condition-3<br>**text + both**<br>practice for **text-2** | condition-3<br>**text + both**<br>practice for **text-1** | condition-3<br>**text + both**<br>practice for **text-2** | condition-3<br>**text + both**<br>practice for **text-1** | condition-3<br>**text + both**<br>practice for **text-2** | condition-3<br>**text + both**<br>practice for **text-1** |
| | recording | recording | recording | recording | recording | recording |

Japanese prosody. No content word is shared between the short passage and the two coffee passages.

Prior to the experiments, we had asked all the learners to fill in questionnaires to learn their native dialect, the textbooks that they used, and the entire duration of learning Japanese. With their responses, we carefully divided the learners into six groups, shown in Table 1. The parenthesized numbers at the top are the numbers of learners comprising each group. Further, we asked all the learners to download the text and the synthetic voices that would be used for experiments in advance. These materials were controlled by password and it was provided immediately before practice for recording.

### 3.3. Experimental design for assessment

The experimental design using the six groups of learners is also illustrated in Table 1. In session 1, all the learners started practicing text-1 or text-2 for five minutes (condition-0), and then recording was done. Groups A, C, and E used text-1 and B, D, and F used text-2. In session 2, all the learners practiced the same text but under different conditions. Here, the six groups were merged into three clusters, AB, CD, and EF. Learners of cluster AB and those of cluster CD used auditory prosody (condition-1) and visualized prosody (condition-2), respectively. Cluster EF used both kinds of prosody (condition-3). After a five-minute practice session, recording was carried out. In session 3, all the learners practiced the same text again but they did with both kinds of prosody instructions. In sessions 2 and 3, cluster EF read the same text under the same condition repeatedly. In the final session, all the learners practiced a new text with both kinds of prosody instructions (condition-3), and then recording was done. In Figure 5, the visualized prosody of the second sentence in text-1 is shown.

For efficiency, all the utterances were recorded with the learners' mobile phones and sent wirelessly to the experimenters' PC after all the sessions were completed. Each of these utterances (64x4=256) was rated in terms of the naturalness in accent and intonation realization for the utterances. Four experienced native teachers of Japanese participated in this rating. A seven-degree scale was used, where 1 indicated very poor, and 7 indicated native-like. For each session, the averaged score over the four teachers was assigned to each learner.

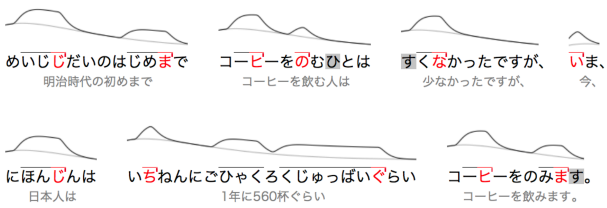


**Figure 5**: The visualized prosody of a sentence in text-1

### 3.4. Subjective assessment after the experiment

After the four recording sessions, two questions were posed to the learners. One was on their preference regarding the four different conditions, 0) text only, 1) +auditory prosody, 2) +visualized prosody, and 3) +both. The other was regarding which order of presentation is preferred, visualized → auditory, or auditory → visualized, if two kinds of prosody instructions are presented sequentially. For both questions, reasons were also requested, but answering was voluntary.

### 3.5. Results and discussion

At first, to verify that the distribution of the learners' original reading performance among the six groups were approximately the same, the mean and standard deviation of each group were calculated for session 1. T-tests were conducted between A and B, C and D, and E and F. No significant difference was observed at the significance level (p-value) of 10%. This means that text-1 and text-2 have very similar difficulty of reading. Then, we formed three clusters AB, CD, and EF by ignoring content difference be-

tween text-1 and text-2. Further, t-tests were carried out again between each pair of the three clusters for session 1. No significant difference was observed at p=10%. This indicates that the reading performance is very similar among the three clusters. The following discussion concerns the three clusters of AB, CD, and EF. Table 2 shows the reading performances of the three clusters for each of the four sessions and Figure 6 illustrates them visually.

**Table 2:** The reading performances of each cluster

| | AB(18) | CD(25) | EF(21) |
|---|---|---|---|
| session 1 | text only | text only | text only |
| mean / s.d. | 4.78 / 0.85 | 4.60 / 0.70 | 4.69 / 0.71 |
| session 2 | +auditory | +visual | +both |
| mean / s.d. | 5.08 / 0.56 | 5.16 / 0.77 | 5.29 / 0.69 |
| $p$ (1→2) | 8.55 % | **0.09** % | **0.005** % |
| session 3 | +both | +both | +both |
| mean / s.d. | 5.51 / 0.75 | 5.37 / 0.70 | 5.29 / 0.63 |
| $p$ (2→3) | **0.54** % | 16.6 % | 100 % |
| session 4 | +both | +both | +both |
| mean / s.d. | 5.40 / 0.83 | 5.18 / 0.88 | 5.05 / 0.73 |
| $p$ (3→4) | 51.5 % | 12.7 % | 10.1 % |
| $p$ (1→4) | **0.33** % | **0.23** % | **3.10** % |

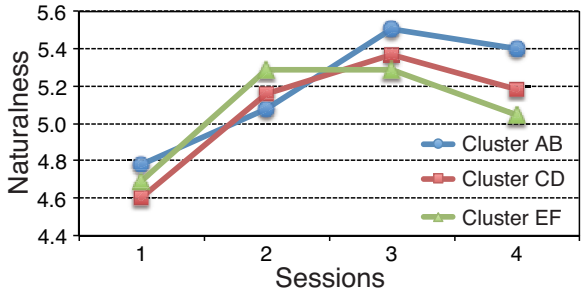


**Figure 6:** Naturalness scores for each session and each cluster

Between sessions 1 and 2, naturalness is improved irrespective of the kind of prosody instruction. At the significance level of 5%, however, only clusters CD and EF show significant improvements. This clearly indicates that visualized prosody shows a much larger effect (p=0.09%) than auditory prosody (p=8.55%). It is possible that this might be due to the problem in the quality of synthesized voices, but we have evidence to refute this possibility. For Suzuki-kun, four TTS systems were selected from ten candidates [28] that could generate synthetic voices with quality of a high enough level that the voices could be used as model utterances in class. Further, our collaborators and we have already held three-hour tutorial workshops on OJAD more than a hundred times in 27 countries but no remarks on insufficient quality of the synthetic voices have been received from the participants.

From session 2 to 3, on the other hand, a significant improvement is found at p=5% only in cluster AB, where visualized prosody is introduced for the first time. This result shows again a large effect from visualized prosody. Why is it so powerful? Generally speaking, adult learners learn a new language with the help of visual stimuli such as symbols. When they learn prosodic control, they may also tend to rely on visual stimuli. If this discussion is valid, the learners will prefer the order of visualized → auditory to that of auditory → visualized when two kinds of prosody instructions are given sequentially.

A preference for visualized prosody is certainly found in the responses of the learners to the questions that were posed after recording. Fully 80% of the learners judged a combination of both kinds of prosody instructions as the most effective. If they are presented sequentially, 73% of the learners preferred the order of visualized → auditory. One main reason for this is that for a given sentence, learning prosodic control to read that sentence in Tokyo Japanese is much easier with visualized prosody. Some learners say that given visualized prosody at first, they can identify in advance the segments to which they should pay closer attention during listening. From this, it appears appropriate to argue that when teachers introduce prosodic training to beginner learners, they should illustrate prosodic control for sentences before they play audio of those sentences. From the experiments, we can also say that the effects of "listen and repeat" using only audio materials is limited.

We received some emails from teachers who have introduced prosody training to their class using OJAD. They reported larger-than-expected effects of Suzuki-kun's prosody visualization on increasing the learners' prosodic naturalness. In the current study, only Chinese learners' performance was examined. In China, all primary schoolers learn Mandarin Chinese as a common dialect[3]. In class, they often learn its tones by using visual symbols and illustrations such as curved arrows. A strong preference for visualized prosody found in the experiments might be attributed to this learning strategy. However, similar preference is to be expected in the case of learners with different L1s. If it is observed experimentally from those learners, teachers' findings above will be verified with high confidence.

In session 4, a new text was read with both kinds of prosody instructions but the scores are lower than those of session 3 in every cluster although the differences are not significant even at the level of 10%. The decrease in naturalness can be explained reasonably as follows. Before recording in session 3, the learners had already practiced the text for fifteen minutes in total but for session 4, recording started after only five minutes of practice. It is clear, however, that the scores of session 4 are significantly higher than those of session 1, where recording started after a five-minute practice session. The improvements from session 1 to 4 are significant at p=1% in clusters AB and CD but significant only at

p=5% in cluster EF. These results might imply that, with prior training based on sequential presentation of prosody instructions, learners have acquired a better strategy to correct their reading even when both instructions are presented simultaneously.

## 4. Conclusions

In this article, we firstly explained the issues of learning and teaching Japanese prosody and the reasons why we developed Suzuki-kun. Then, we described its technical development. Suzuki-kun is very unique in that it can visualize the full and hierarchical prosodic control for any Japanese text. It can also provide high-quality synthetic speech based on the visualized prosody. Finally, experimental assessment was done focusing on how learners' spoken Japanese changed with self-practice using Suzuki-kun. Significant improvements of prosodic naturalness were observed and it is interesting that visualized prosody was found to be more effective than auditory prosody. In future work, we will investigate whether similar improvements and preferences are found in non-Chinese learners. Further, we are interested in developing a new module to detect prosodic errors and provide some corrective feedback.

## 5. Acknowledgements

This work was supported financially by MEXT/JSPS KAKENHI Grant Numbers JP26118002, JP26240022, and JP16K13237.

## 6. References

[1] Project OJAD, Online Japanese Accent Dictionary (OJAD), http://www.gavo.t.u-tokyo.ac.jp/ojad/

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

[3] Siriphonphaiboon, Y. 2008. The effectiveness of self-monitoring on Japanese accent learning: an analysis of questionnaire on Thai L1 learners of Japanese, *Journal of the phonetic science of Japan*, 12, 2, 17-29 (in Japanese).

[4] Isomura, K. 2001. The current state of the Japanese accent education in foreign countries, *Proc. Autumn Meeting of the Society for Teaching Japanese as a Foreign Language*, 211-212 (in Japanese).

[5] Isomura, K., Abe, S., Hayashi, R., Shibata, T., and Minematsu, N. 2016. The current situation and problems of pronunciation training of Japanese, *Proc. Spring Meeting of the Society for Teaching Japanese as a Foreign Language* (in Japanese).

[6] Sagisaka, Y. and Sato, H. 1983. Accentuation rules for Japanese word concatenation, *Trans. IEICE*, J66-D, 7, 849-856 (in Japanese).

[7] Ooyama, R. 2014. 日本語の自然な発話習得に関する一考察, paper session 7-C, International Conference on Japanese Language Education (ICJLE).

[8] Kato, S., Short, G., Minematsu, N., Tsurutani, C., and Hirose. K. 2011. Comparison of native and non-native evaluations of the naturalness of Japanese words with prosody modified through voice morphing, *Proc. SLaTE*, CD-ROM.

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

[10] NHK Broadcasting Culture Research Institute 新・NHK アクセント辞典・ポイント解説, http://www.nhk.or.jp/bunken/forum/2016/accent.html

[11] Ohta, S. 2010. On the intriguing relationship between accent patterns of place names in Japan and the Latin accent rule, *Journal of cross-cultural studies*, 4, 1-14 (in Japanese).

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

[13] Project OJAD, Promotion video of OJAD with speech output function, https://youtu.be/It-NBJKJd1g (English version) https://youtu.be/kPJifu2aBXg (Japanese version)

[14] Nakamura, I., Hirano, H., Minematsu, N., Suzuki, M., Nakagawa, C., Nakamura, N., Tagawa, Y., Hirose, K., and Hashimoto, H. 2013. Development of a web framework for teaching and learning Japanese prosody: OJAD (Online Japanese Accent Dictionary), *Proc. INTERSPEECH*, 2554-2558.

[15] Minematsu, N., Nakamura, I., Suzuki, M., Hirano, H., Nakagawa, C., Nakamura, N., Tagawa, Y., Hirose, K., and Hashimoto, H. 2013. Development and evaluation of online infrastructure to support teaching and learning of Japanese accent and intonation, *IEICE Trans.* J96-D, 10, 2496-2508 (in Japanese).

[16] Minematsu, N., Hashimoto, H., Hirano, H., and Saito, D. 2015. Development of a prosodic reading tutor of Japanese --effective use of TTS and F0 modeling techniques for CALL--, *Proc. SLaTE*, 189.

[17] Minematsu, N. 2015. Development of an online infrastructure for teaching Japanese prosody based on information processing of speech and text corpora, *Journal of the Phonetic Society of Japan*, 19, 1, 18-31 (in Japanese).

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

[19] Martin, P. 2010. Learning the prosodic structure of a foreign language with a pitch visualizer, *Proc. Speech Prosody*.

[20] WinPitch, http://www.winpitch.com

[21] Komissarchi, J. and Komissarchik, E. 2000. Better Accent Tutor --Analysis and visualization of speech prosody, *Proc. InSTILL*, 86-89.

[22] Pellegrini, T., Costa, A., and Trancoso, T. 2012. Less errors with TTS? A dictation experiment with foreign language learners, *Proc. INTERSPEECH*.

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

[24] Japan Electronics and Information Technology Industries Association (JEITA), http://www.jeita.or.jp

[25] Minematsu, N., Hirano, H., Nakamura, N., and Oikawa, K. 2016. "Naturalness improvement of learners' spoken Japanese by using the web-based prosodic reading tutor, Suzuki-kun," *Proc. Speech Prosody* 252-256.

[26] JASLON, http://www.jaslon.net

[27] Japan Foundation, 2013. The statistical figures of Japanese education in foreign countries, https://www.jpf.go.jp/j/about/press/dl/0927.pdf

[28] KDDI Laboratory, 2015. Speech synthesis software N2 and its TSS library SDK,

http://www.kddi-research.jp/products/n2.html

[1] In this article, the term 'accent' is used to indicate word-based local pitch movement and that of intonation refers to more global pitch movement. The former includes steep pitch rise or fall but the latter generally corresponds to gradual pitch declination.

[2] Phrase-based grouping or chunking of words is sometimes explained to have a function of signaling generation of a new semantic instance in an utterance [11].

カラー + しゃしん → カラーしゃしん (type-4)　カラーしゃしん(type-1+type-0)

For example, カラー(color) has type-1 accent and しゃしん(photo) has type-0 accent as lexical accent, but カラーしゃしん has type-4 accent. If カラーしゃしん is pronounced as simple concatenation of a type-1 word and a type-0 word, then, native listeners will recognize that word pair not as colored photo but probably as color and photo. By grouping or chunking words in a single accent form, those words will have a new semantic instance. As explained before, different dialects often have different lexical accent systems. Even if a non-Tokyo accent is used by learners, the comprehensibility of their spoken Japanese will not be degraded largely. In the case of a compound expression such as カラーしゃしん, however, when its constituent words are not grouped in a single accent form, their comprehensibility will be degraded. Given カラーしゃしん , each word will be recognized correctly but the word pair will not be recognized correctly. By using a single accent form, the word pair can generate a new meaning.

[3] In China, people with different dialects cannot communicate orally when they speak in their dialects. For smooth oral communication, every child in China learns 普通語 in primary school, which literally means the common language and is very close to Mandarin Chinese.