IEICE TRANS. INF. & SYST., VOL.E100–D, NO.4 APRIL 2017

INVITED PAPER *Special Section on Award-winning Papers*

Development and Evaluation of Online Infrastructure to Aid Teaching and Learning of Japanese Prosody

Nobuaki MINEMATSU[†a)], ***Senior Member***, **Ibuki NAKAMURA**[††], ***Nonmember***, **Masayuki SUZUKI**[†††], ***Member***, **Hiroko HIRANO**[††††], **Chieko NAKAGAWA**[†††††], **Noriko NAKAMURA**[††††], **Yukinori TAGAWA**[††††††], ***Nonmembers***, **Keikichi HIROSE**[†], ***Fellow, Honorary Member***, ***and*** **Hiroya HASHIMOTO**[†], ***Nonmember***

SUMMARY This paper develops an online and freely available framework to aid teaching and learning the prosodic control of Tokyo Japanese: how to generate its adequate word accent and phrase intonation. This framework is called OJAD (Online Japanese Accent Dictionary) [1] and it provides three features. 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 fundamental conjugations 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 any given sentence and the expected positions of accent nuclei in the sentence. The third feature is technically implemented by using an accent change prediction module that we developed for Japanese Text-To-Speech (TTS) synthesis [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 done by teachers and learners show extremely high pedagogical effectiveness of the developed framework.

key words: *speech training in Japanese, word accent, intonation, speech synthesis, accent prediction, assessment*

1. Introduction

The number of learners who are learning Japanese outside of Japan is increasing and [4] reported that it reached nearly four millions in 2012. In most of the cases, learners' motivation is to acquire good skills to communicate orally with others in Japanese. When the target language is English, which is an international language, learners will have to communicate with many non-native speakers of English, and therefore listeners' acceptability of accented pronunciations is generally high [5]. In the case of Japanese, however, learners will often communicate only with native speakers of Japanese. This will indicate that, if learners want to improve their pronunciation, they have to correct their pronunciation to make it closer to native pronunciation. It is true that many learners are learning Japanese for business and, in this case, they are required to obtain a good ability of public speaking. As is well-known, native speakers of Japanese often use Tokyo Japanese in public even when they are from local regions. In Japanese, prosodic control often depends on dialects and if learners want to learn Tokyo Japanese, they have to learn its unique prosodic control. However, mainly due to time limitation, prosodic teaching is very rare in class and methodologies for prosodic teaching are not well provided even in teachers' training courseware. To solve this situation, in this paper, web-based online infrastructure of learning and teaching of accent and intonation of Tokyo Japanese is developed and assessment of the infrastructure is done objectively and subjectively.

2. The Current Issues of Speech Training of Japanese

One of the main problems in learning how to speak natural Japanese is prosody. Every content word in Japanese has its own lexical accent, 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 Fig. 1. Many learners, however, don't know this fact because it is not always taught in class [6]–[8]. The current situation of teaching Japanese word accents is reported in [9].

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) [10]. Several examples are shown in Fig. 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 very rarely to learners [11] and it is

Fig. 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.

Manuscript received November 7, 2016.
Manuscript publicized December 22, 2016.
[†]The authors are with The University of Tokyo, Tokyo, 113–8656 Japan.
[††]The author is with Fujitsu Limited, Kawasaki-shi, 211–8588 Japan.
[†††]The author is with IBM Research, Tokyo, 103–0015 Japan.
[††††]The authors are with Tokyo University of Foreign Studies, Fuchu-shi, 183–8534 Japan.
[†††††]The author is with Waseda University, Tokyo, 169–8050 Japan.
[††††††]The author is with Nihon University, Tokyo, 102–8275 Japan.
a) E-mail: mine@gavo.t.u-tokyo.ac.jp
DOI: 10.1587/transinf.2016AWI0007

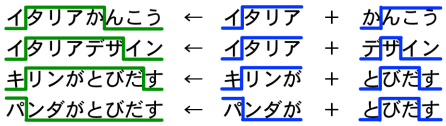


Fig. 2 Not word-based but phrase-based control of accent. The third and forth examples can be read with two accents.

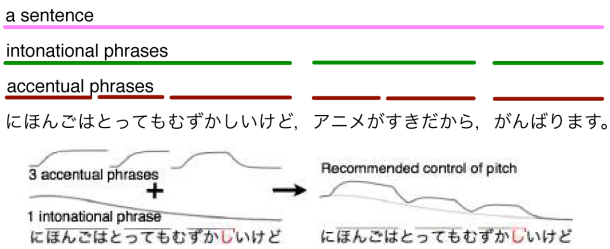


Fig. 3 Hierarchical and prosodic structure of a sentence.

not explained explicitly even in teachers' training courseware. 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 include inadequate control of pitch. This is because the unit of accent control is often not a word but a phrase for reading sentences.

It is very natural that the accent type of a word is different between when it is spoken in isolation and when it is 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 [12]. Further, accent control varies among dialects [13]. 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 [14], 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.

The function of phrase-based accent control refers to grouping words of a phrase into one accent unit or chunk. 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 [15]. That is to say that,

Fig. 4 Visualized prosody in [15].

with those pauses inserted, it takes a longer time for native listeners to comprehend what is said. A large comprehensibility difference between before and after intonation training can be checked in [16]. Generally speaking, however, the relation between intonation control and comprehensibility is not explained explicitly in class. Only a few textbooks such as [15] explain how to control accent and intonation visually (See Fig. 4). One intonational phrase can have multiple accentual phrases. This means that multiple accent nuclei can be found in a single intonational phrase and this complicated prosodic structure will often be problematic especially for beginners. For them, simpler but natural enough prosodic control is desired. By following [15], we derive simplification rules for presenting prosodic control.

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.

In this paper, we attempt to solve the above serious problems by providing the very first online framework for teaching and learning the prosodic control of Tokyo Japanese. 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 any given sentence. Although TTS technologies have been used in CALL development in previous studies [17], [18], a main focus was always put on how to use speech output. In this paper, we don't use speech output but use an internal module of a TTS system as visualizer of the hidden prosodic structure in a given sentence. This prosodic structure is what learners desperately want to know to be linguistically dressed-up.

3. Implementation of the Three Features

3.1 Comprehensive Illustration of Accent Changes

Since word accent changes found in conjugation of verbs and adjectives are relatively regular and systematic, we realized a module 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 fun-

Fig. 5 Illustration of the accent patterns of conjugated forms.

damental conjugations were adopted and their accents are displayed in a table. Figure 5 shows an example. Twenty 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 predictor [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 module does not predict 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 module 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 of fundamental frequencies, so called as the Fujisaki model [19]. By controlling its parameters, it is easy to realize a pitch pattern with complete mora isochrony. The drawn pattern is of course not acoustically realistic, but what has to be presented to learners is not observed pitch patterns but the pitch pattern "images" that teachers want to exhibit 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. 5, 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 the icon of that purpose. These samples can be downloaded onto users' PCs.

Since all the words are directly from the textbooks, we implemented a very flexible and textbook-oriented 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

Fig. 6 Illustration of the pitch pattern of long expressions.

at University of Tsukuba. 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.

3.2 Illustration of the Accent of Long Verbal Expressions

The first feature 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 feature cannot explain the accent control to read verbal expressions in that textbook completely. So, we implemented another module as the second feature to exhibit 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 accent groups based on their word accent (accented and unaccented) and into three conjugation groups based on their conjugation manner. If the accent group and the conjugation group of a given verb are known, the accent pattern of any verbal expression, which is comprised of that verb and one of the 320 postpositional expressions, can be correctly predicted. In the module, by running morphological analysis to an input verbal expression with MeCab [20], the verb and its postpositional expression are detected automatically. Using the attributes of the verb estimated by MeCab, the module can identify the accent group and the conjugation group for that verb. Then, the resulting accent pattern (the accent nucleus position) is visually presented with its pitch curve.

Figure 6 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 module can find the postpositional expression in the database, it shows the accent pattern in a red rectangle. If the module cannot, however,

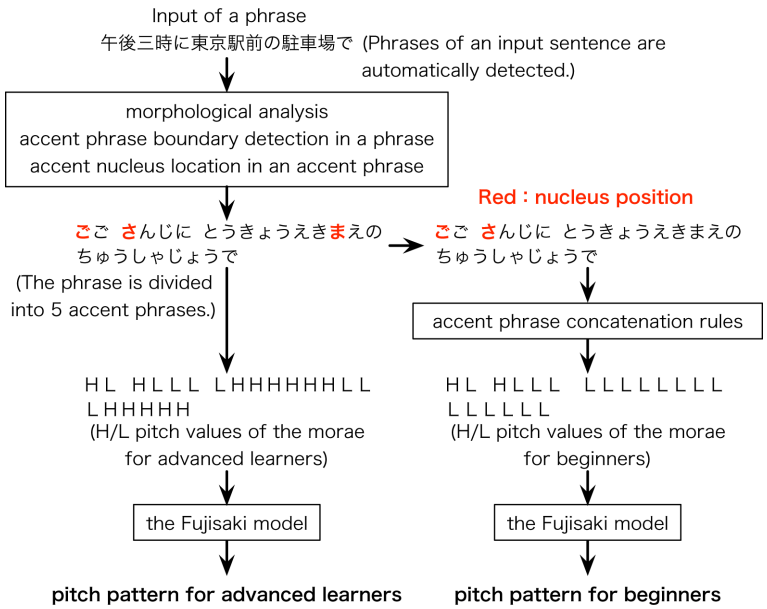


Fig. 7 Generation of original and simplified pitch patterns.

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.

3.3 Illustration of the Pitch Pattern of Any Input Sentence

The first and second features 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 features are not sufficient for learners to learn how to read sentences naturally in Tokyo Japanese. So, as the third feature, we developed a prosodic reading tutor to support learners by presenting the pitch pattern of any given sentence, which is expected to be observed when a native speaker reads that sentence naturally but neutrally.

This feature can be realized by using several internal modules developed for TTS synthesizers. They are morphological analysis (linguistic analysis) [20], accentual phrase boundary detection from text [2], [3], accent nucleus detection for an detected accentual phrase [2], [3], and pitch pattern visualization by the Fujisaki model [19]. As mentioned in Sect. 2, direct visualization of the output of these modules is not good pedagogically because it is sometimes too complicated for learners to follow. As simple pitch patterns as possible with good naturalness are desired to be presented. In tight collaboration with Japanese teachers, we designed simplification rules [15]. Generation of original and simplified pitch patterns is schematized in Fig. 7.

An input sentence is divided into intonational phrases based on punctuation marks and phrase boundary marks (/), which are explicitly given by users. Three analyses of morphological analysis, accentual phrase boundary detection, and accent nucleus detection are run for each intonational phrase. Then, the input intonational phrase is divided into multiple accentual phrases, in each of which the accent

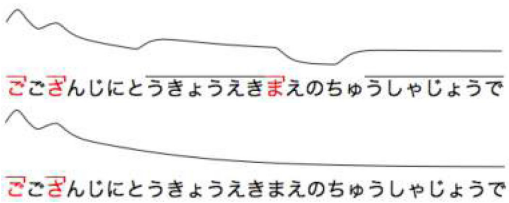


Fig. 8 Original and simplified pitch patterns.

nucleus position is predicted automatically. For advanced learners, these accentual phrases are directly used to visualize the pitch pattern. An example is shown at the top of Fig. 8. For beginners, simplification rules are applied. [15] claims that, for beginning learners, it is a good strategy to focus on the first accent nucleus in an intonational phrase and ignore the remaining nuclei in that phrase. By this strategy, an intonational phrase comes to have only one accent nucleus, which is simple enough with sufficient naturalness. We adopted this strategy basically but modified it slightly by taking account of Japanese listeners' perceptual characteristics. It is shown experimentally in [21] that Japanese listeners are very sensitive to the accent nucleus when it is found at the first mora of an accentual phrase. Following this fact, our simplification rules do not ignore the accent nucleus which is found at the head of an accentual phrase even when that nucleus is not the first nucleus in an intonational phrase. The H/L sequences of multiple accentual phrases are concatenated to generate the final H/L sequence for the entire intonational phrase. An example of the simplified pitch curve is shown at the bottom of Fig. 8.

It should be noted that the first two modules are effectively error-free but the last one sometimes show incorrect pitch patterns due to on-line analysis errors. They can become serious because learners may not be able to detect er-

rors committed by the module. Paying attention to this fact, we carried out only subjective assessment for the first two modules and both subjective and objective assessment for the last one.

4. Assessment of the Implemented Features

4.1 Subjective Assessment of All the Features

We prepared an introductory web page of "Let's use OJAD for accent training!" for subsequent subjective assessments. In this page, the implemented three features 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 explained honestly using some erroneous responses from the tutor.

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 of Japan. Although the subjective assessment was composed of a series of questionnaires, due to space limit, we show the results of only two key 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 Table 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.

4.2 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 Tokyo Japanese, they always ask teachers to detect accent nuclei in their manuscript because it is extremely difficult for learners to detect the nuclei by themselves. In objective assessment here, we impose on learners a task of detecting accent nuclei 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 [22] and 2) a PC-based commercial Japanese speech synthesizer [23].

Table 1 Assessment of the three proposed modules (%).

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

	1st module	2nd module	3rd module
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 module in your class?

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

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 any input sentence as speech output. 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 segmentation into intonational phrases was done for each paragraph by the four teachers. The boundaries for those intonational phrases were explicitly shown to the subjects as punctuation mark or phrase boundary mark in the paragraph. The numbers of intonational phrases are 73, 68, 73, and 70 for p0 to p3, respectively. The actual task imposed on the subjects is detecting the first accent nucleus in each phrase. Figure 9 show an example of the paragraph and an example of the PC desktop.

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

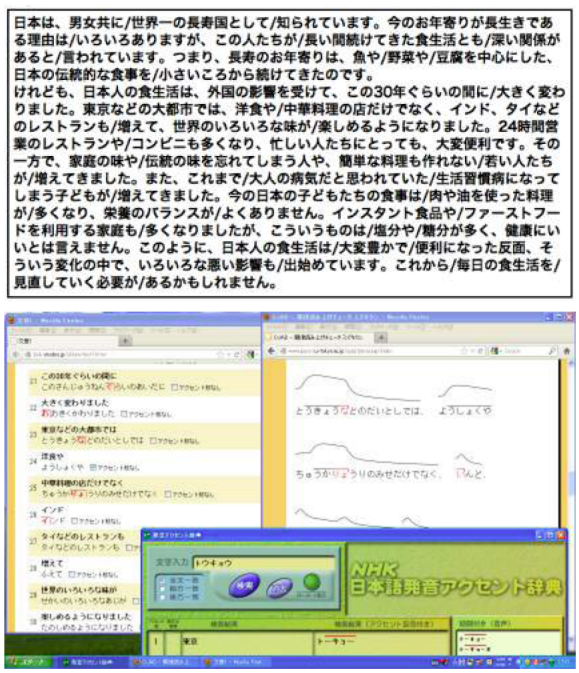


Fig. 9 Examples of the paragraph and the PC desktop.

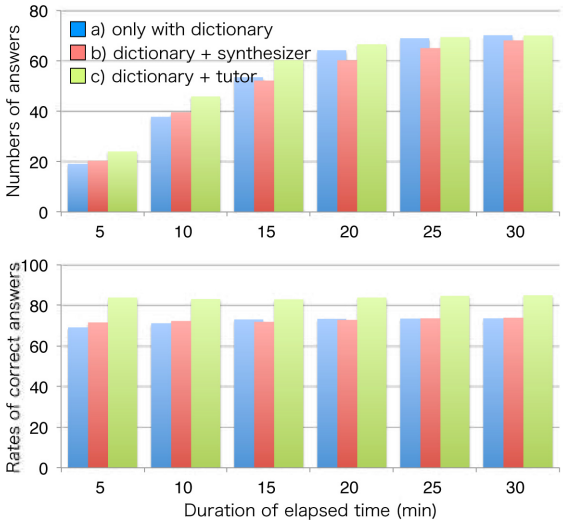


Fig. 10 Results of objective assessments.

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 learners' detection of accent nuclei 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 imposed on ten university students who are native speakers of Tokyo Japanese. Their performance was 61.6%, lower than that of the learners. As described in Sect. 1, this result is reasonable. Native speakers can speak Tokyo Japanese very fluently but they are not good at being aware of accent nuclei. The L1 of 27 subjects out of 36 was tonal languages. The performance of the reading tutor and the synthesizer is 93.2% and 95.9%, respectively, which is by far higher than that of ordinary native speakers.

Results of accent nucleus detection using p1 to p3 are shown in Fig. 10. In the experiment, all the mouse clicks were monitored and recorded in history files. Using the files, the speed of detection and the precision of detection can be compared among the three kinds of facilities of a), b), and c). In Fig. 10, 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 detection and the bottom figure shows the precision of detection 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 detection. 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 detection. 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%) and that of the synthesizer (91.3%). We can say that the learners' judgments often revise the tutors' suggestions for the worse. As implemented in Sect. 3.2, availability or confidence of the system responses should be given to learners probably using colors so that learners can use the system without being afraid of making mistakes. 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. Visual presentation of the prosodic structure is extremely preferred.

5. Conclusions

In this paper, we built an online and freely available framework for teaching and learning the prosodic control of Tokyo Japanese. Three features were implemented using speech synthesis technologies. Both subjective and objective assessment experiments showed extremely high pedagogical effectiveness of the proposed framework. The framework is called OJAD and it was released to the public in August 2012 in an international conference of Japanese education [24]. Since then, we gave more than one hundred tutorial workshops of OJAD in 28 countries. OJAD was translated into 14 non-Japanese languages so that beginning learners can use it easily. Many teachers recognize OJAD as the first and currently only educational tool for teaching and learning the prosodic control of Tokyo Japanese. We received responses from learners who participated in speech contests of Japanese. They claim that it is very difficult to be selected as finalist in a speech contest without practicing with OJAD. We hope that every learner can acquire the dress code of Japanese to speak naturally in public.

References

[1] OJAD, http://www.gavo.t.u-tokyo.ac.jp/ojad/
[2] N. Minematsu, S. Kobayashi, S. Shimizu, and 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, and 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] Japan Foundation, "Survey report on Japanese language education abroad 2012," https://www.jpf.go.jp/j/project/japanese/survey/result/dl/survey_2012/2012_s_excerpt_e.pdf
[5] J. Jenkins, The phonology of English as an international language, Oxford University Press, 2000.
[6] A-Rong-Na and R. Hayashi, "The effect of shadowing training for Mongolian and Chinese learners of Japanese," IEICE Technical Report, SP2009-151, 2010 (in Japanese).
[7] Y. Siriphonphaiboon, "The effectiveness of self-monitoring on Japanese accent learning: an analysis of questionnaire on Thai L1 learners of Japanese," J. Phonetic Science of Japan, vol.12, no.2, pp.17–29, 2008 (in Japanese).
[8] K. Isomura, "The current state of the Japanese accent education in foreign countries," Proc. Autumn Meeting of the Society for Teaching Japanese as a Foreign Language, pp.211–212, 2001 (in Japanese).

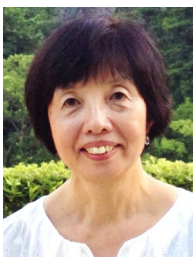
[9] K. Isomura, S. Abe, R. Hayashi, T. Shibata, and N. Minematsu, "The current situation and problems of pronunciation training of Japanese," Proc. Spring Meeting of the Society for Teaching Japanese as a Foreign Language, 2016 (in Japanese).
[10] Y. Sagisaka and H. Sato, "Accentuation rules for Japanese word concatenation," IEICE Trans. Inf. & Syst. (Japanese Edition), vol.J66-D, no.7, pp.849–856, July 1983.
[11] R. Ooyama, "An examination for learning how to speak naturally in Japanese," International Conference on Japanese Language Education, paper session 7-C, 2014 (in Japanese).
[12] S. Kato, G. Short, N. Minematsu, C. Tsurutani, and K. Hirose, "Comparison of native and non-native evaluations of the naturalness of Japanese words with prosody modified through voice morphing," Proc. Int. Workshop on Speech and Language Technology in Education, CD-ROM, 2011.
[13] Z. Uwano, "Word accents of Japanese," in series of Japanese and Japanese Education, published by Meiji-Shoin, 1989 (in Japanese).
[14] NHK Broadcasting Culture Research Institute, Introduction to the New NHK Accent Dictionary, 2016.
[15] C. Nakagawa, N. Nakamura, and S. Ho, Japanese pronunciation drills for advanced oral presentation, Hitsuji-Shobo, 2009 (in Japanese).
[16] OJAD promotion video, https://youtu.be/It-NBJKJd1g
[17] A. Black, "Speech synthesis for educational technology," Proc. SLaTE, CD-ROM, 2007.
[18] M. Eskenazi, "An overview of spoken language technology for education," Speech Communication, vol.51, no.10, pp.832–844, 2009.
[19] H. Fujisaki and K. Hirose, "Analysis of voice fundamental frequency contours for declarative sentences of Japanese," J. Acoust. Soc. Japan (E), vol.5, no.4, pp.233–242, 1984.
[20] MeCab, http://mecab.sourceforge.net/
[21] N. Minematsu and K. Hirose, "Role of prosodic features in the human process of perceiving spoken words and sentences in Japanese," J. Acoust. Soc. Japan(E), vol.16, no.5, pp.311–320, 1995.
[22] NHK Japanese word accent dictionary, published by NHK, 1998.
[23] HOYA service, http://voicetext.jp
[24] N. Minematsu, M. Suzuki, H. Hirano, C. Nakagawa, N. Nakamura, Y. Tagawa, and K. Hirose, "Development of an online Japanese accent dictionary with speech output facility," Proc. Int. Conf. on Japanese Language Education, p.94, 2012.

Nobuaki Minematsu received the doctor of Engineering in 1995 from the University of Tokyo. Currently, he is a full professor there. From 2002 to 2003, he was a visiting researcher at KTH, Stockholm, Sweden. He has a wide interest in speech communication covering from science to engineering. He is a member of IEEE, ISCA, IPA, SLaTE, IEICE, APSIPA, ASJ, PSJ, IPSJ, JSAI, etc.

Ibuki Nakamura received the master of Information Science and Technology in 2015 from the University of Tokyo. Since 2015, he has been working as researcher at Fujitsu Limited.

Masayuki Suzuki received the Ph.D. degree in electrical engineering and information systems from the University of Tokyo in 2013. Since 2013, he has been working as staff researcher at IBM Research - Tokyo. His research interests include speech and spoken language processing. He is a member of ASJ, IEICE, IEEE, and ISCA. He received the Awaya Award from the ASJ in 2013.

Hiroko Hirano received the doctor of Science in 2009 from the University of Tokyo. After she worked as a full-time lecturer and an associate professor at Chinese universities from 2009 to 2015, she works now as a full-time lecturer for Tokyo University of Foreign Studies. She is a member of Society for Teaching Japanese as a Foreign Language (NKG), ASJ, PSJ, JLEM, etc.



Chieko Nakagawa received the doctor of Humanities and Sociology in 2001 from the Ochanomizu University. She teaches Japanese and pronunciation of Japanese for many years at universities and graduate schools. She is a member of Society for Teaching Japanese as a Foreign Language (NKG), Phonetic Society of Japan (PSJ), etc.

Noriko Nakamura earned the Master of Arts at Ochanomizu University in 1996, and completed Doctoral program with expulsion at Ochanomizu University in 2002. She is currently a part-time lecturer at Tokyo University of Foreign Studies, Waseda University and Keio University. She teaches Japanese Pronunciation and Academic Listening to foreign students. She is a member of NKG and PSJ.

Yukinori Tagawa received the Master of Arts in 2002 from the Prefectural University of Kumamoto, and in 2009, completed the doctoral program without a doctoral degree in Japanese Linguistics, Graduate School of Letters, Osaka University. He teaches Japanese, especially pronunciation of Japanese at universities. He is a member of NKG, PSJ, JlEM, JASS, TAPS, etc.

Keikichi Hirose received the doctor of Engineering in 1977 from the University of Tokyo. He was a professor there from 1994 to 2015, and received Professor of Emeritus title. In 1987, he was Visiting Scientist at RLE, MIT, U.S.A. In 2015, he was honored as a Named Person of Merit in Science and Technology by the Mayor of Tokyo. He is a member of ISCA (Board), IEEE, ASA, etc.

Hiroya Hashimoto received the master of Information Science and Technology in 2013 from the University of Tokyo and is now a doctoral candidate at Graduate School of Engineering, the University of Tokyo.