Speech Prosody 2016
31 May - 3 Jun 2106, Boston, USA

Improvement of Naturalness of Learners' Spoken Japanese by Practicing with the Web-based Prosodic Reading Tutor, Suzuki-kun

Nobuaki MINEMATSU[1], *Hiroko HIRANO*[2], *Noriko NAKAMURA*[2], *Kohji OIKAWA*[3]

[1]The University of Tokyo, [2]Tokyo University of Foreign Studies, [3]JASLON

mine@gavo.t.u-tokyo.ac.jp, hirano_hiroko@tufs.ac.jp

Abstract

It is known that learning prosodic control for speaking Japanese is effective to reduce syntactic and lexical ambiguity of learners' spoken Japanese and to improve its naturalness and comprehensibility [1]. In conventional curriculum, however, prosody training has not been provided satisfactorily for learners partly because teaching materials for prosody training are limited. In our previous studies [2, 3, 4, 5], we built a web-based system that, for any given text, can illustrate the prosodic control required to read that text in Tokyo Japanese and also provide a high-quality synthetic voice for that text based on the visualized prosody. Although this system is currently used by many learners, assessment of the system was done only in terms of users' satisfaction [3]. In this study, the effectiveness is examined quantitatively by using eighty Chinese learners of Japanese. We compare the improvements of naturalness realized by practicing 1) with synthetic voices, 2) with visualized prosody, and 3) with both of them. Experimental results show that simultaneous use of auditory and visual instructions is the most effective but that visualized prosody can help learners more than auditory prosody.
Index Terms: Japanese education, prosody, accent and intonation, OJAD, Suzuki-kun, speech synthesis, naturalness

1. Introduction

A language learner attempts to acquire good skills of speaking the target language and every target language has its unique difficulties. It is the case with Japanese and one of the main problems in learning Japanese pronunciation is prosody. Every content word of Japanese has its own lexical accent. It is mora-based pitch accent and binary values (High/Low) are assigned to each mora. This logically means 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, called accent types. Namely, the lexical accent of a particular N-mora word is one of those N types. Many learners, however, don't know this fact because it is rarely taught explicitly in class [6, 7, 8].

When a native speaker speaks, accent control is done by a unit of not word but phrase (i.e. the accentual phrase). 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 pitch control. This fundamental mechanism of accent control is also taught very rarely [9].

Phrase-level accent control directly means that the pitch pattern of a word is often different between when it is spoken in isolation and when spoken in context [10]. 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 of Japanese, is generally not high. Further, accent control varies among dialects [11]. It is not uncommon that native teachers whose native dialect is not Tokyo Japanese confess that they are unconfident in teaching accent control. Tokyo Japanese is the common dialect and used in business. Since many Asian learners are learning Japanese for business, a good infrastructure for learning accent control has been requested from them.

The function of phrase-based accent control is regarded as the grouping of words of a phrase into one accent unit. Another grouping mechanism is also present in Japanese. It is termed the intonational phrase and it is composed of one or more accentual phrases. In other words, accentual phrases of an intonational phrase make up one intonation unit. Between consecutive intonation units, a pause is often inserted and when the utterance is transcribed, a punctuation mark is often placed there. If additional pauses are inserted due to non-nativeness or disfluency, comprehensibility is easily degraded for lack of adequate phrase structures [1]. Readers can check loss of comprehensibility in [12]. Generally speaking, the relation between intonation control and comprehensibility is not explained explicitly in class.

It seems that teachers have not explained Japanese prosodic control sufficiently to human learners. However, engineers have explained it intensively to machine learners for a few decades. This is because the aim of Text-to-Speech (TTS) systems of Japanese is the conversion of input text into spoken Tokyo Japanese with good naturalness and comprehensibility. In our previous studies [2, 3, 4, 5], several modules or techniques embedded in Japanese TTS systems were transferred to educational software. In 2012, we launched OJAD (Online Japanese Accent Dictionary [13]) [2, 3], 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 [4, 5]. 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 [14]. Further, the visualized prosody is realized as audio with a synthetic voice using an HMM-based speech synthesizer. We can find several previous studies [15, 16, 17], where fundamental frequencies of learners' utterances and teachers' utterances were plotted for visual comparison. Suzuki-kun was built with a different objective in mind. It can illustrate the full and hierarchical prosodic control of any *given text* and generate its 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 the above function. OJAD has been already translated into 14 non-Japanese languages and is now used internationally[1]. As for the effectiveness of OJAD, we examined users' satisfaction only with questionnaires [2, 3]. In this paper, we quantitatively investigate the effectiveness of Suzuki-kun on the improvement of naturalness in terms of prosodic control.

[1]The number accesses for the last six months is about 250 thousands.

doi: 10.21437/SpeechProsody.2016-53

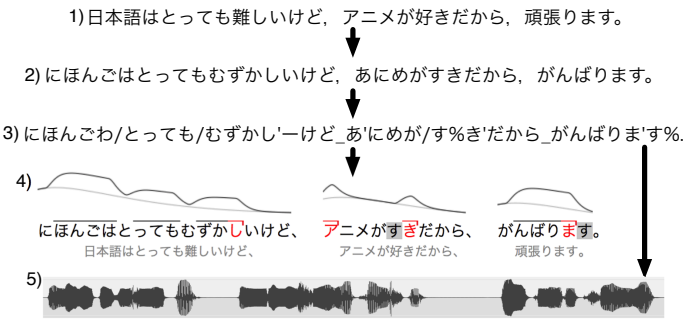

Figure 1: Prediction and easy-to-understand visualization of prosodic features for a given Japanese text and its waveform generated by an HMM-based speech synthesizer

2. The prosodic reading tutor, 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 used as model utterances [18]. Generally speaking, a TTS engine does not read an input text directly but processes its corresponding phoneme sequence with the prosodic symbols attached and predicted by an internal module of the engine.

Fig. 1 shows 1) an original text, 2) its phonemic transcript as a Hiragana sequence, 3) output from a prosody prediction module that we developed in [19], and 4) visual output from Suzuki-kun. In 3), the prosodic features are assigned as symbols. ' is an accent nucleus. / and _ indicate an accentual phrase boundary without a pause and that with a pause, respectively. The latter also functions as intonational phrase boundary. This symbolic representation is defined by JEITA[2] [20] 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. 3) claims that this sentence should be divided into three intonational phrases and that the phrases contain three, two, and one accentual phrase(es) from the head to the end of the sentence. This prosodic hierarchy is illustrated in 4) by using Fujisaki's model. The values of the model parameters are determined by applying a set of rules to 3), which were constructed through collaboration with teachers [2, 3].

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 but it is a pitch pattern that should be used as *mental target* in prosody training to read this sentence[3]. Suzuki-kun can produce synthetic voices with four different speaker identities (2 males and 2 females) of 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 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 the pitch pattern of a sentence will even vary from speaker to speaker.

[2]Japan Electronics and Information Technology Industries Association

[3]It might be better to deform the shape of the mental target dependently on learners' L1 but the current version of Suzuki-kun illustrates the mental target independently of their L1.

3. Experimental assessment of Suzuki-kun's instructions

3.1. The objective of the assessment

We can find many textbooks with an audio CD as attachment, which contains readings of all the sentences in the textbooks. A very small number of those textbooks, [1] 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 a very complicated set of rules for prosodic control to read new texts with natural-sounding prosody. This will cause their reading to become accented. By using Suzuki-kun, however, learners can know the prosodic control *instantly* and by reading many new texts under the guidance of Suzuki-kun, they will be able to acquire its rules in an implicit manner.

With a focus on these distinctive characteristics of Suzuki-kun, in this paper, we conduct experiments to compare the improvement of prosodic control in terms of naturalness that are realized under the following four conditions: 1) practice with auditory prosody only (use of synthetic voices), 2) that with visualized prosody only, and 3) that with both auditory prosody and visualized prosody. For comparison, the 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* feedback, 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, recording was always done after a short five-minute self-practice in each condition.

3.2. Subjects and texts adopted for the experiments

A fair comparison of learners' reading performance under different conditions is sometimes difficult because different groups of learners with as close to the same proficiency level as possible have to be selected for the different conditions. Similarly, different sets of text with as close to the same difficulty level as possible have to be prepared for the different conditions. In this study, 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 [21], we asked Chinese learners who had learned Japanese for one to two years to participate in our experiments. Eighty learners joined the experiments[4].

As for passages selection, two paragraphs were adopted, text-1 and text-2, both from the same unit of a textbook, written for a second semester Japanese course. They had 197 and 196 characters respectively. Another short paragraph was prepared and used only to explain Suzuki-kun's visualization of prosody.

Prior to the experiments, we had asked all the learners to fill in questionnaires to know 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 Tab. 1, so that each group will have a similar skill of reading. Further, we asked all the learners to download the texts and the synthetic voices in advance that would be used for experiments. These materials were controlled by password, which was provided immediately before practice for recording. Only with the password, the learners can access to the meterials.

[4]They were recruited at Oikawa's Camp for speech training [22].

Table 2: The experimental design by using the six groups of learners

	A (11)	B (10)	C (16)	D (18)	E (13)	F (12)
session 1	condition-0 **text only** practice for **text-1**	condition-0 **text only** practice for **text-2**	condition-0 **text only** practice for **text-1**	condition-0 **text only** practice for **text-2**	condition-0 **text only** practice for **text-1**	condition-0 **text only** practice for **text-2**
	recording	recording	recording	recording	recording	recording
session 2	condition-1 **text + auditory** practice for **text-1**	condition-1 **text + auditory** practice for **text-2**	condition-2 **text + visual** practice for **text-1**	condition-2 **text + visual** practice for **text-2**	condition-3 **text + both** practice for **text-1**	condition-3 **text + both** practice for **text-2**
	recording	recording	recording	recording	recording	recording
session 3	condition-3 **text + both** practice for **text-1**	condition-3 **text + both** practice for **text-2**	condition-3 **text + both** practice for **text-1**	condition-3 **text + both** practice for **text-2**	condition-3 **text + both** practice for **text-1**	condition-3 **text + both** practice for **text-2**
	recording	recording	recording	recording	recording	recording
session 4	condition-3 **text + both** practice for **text-2**	condition-3 **text + both** practice for **text-1**	condition-3 **text + both** practice for **text-2**	condition-3 **text + both** practice for **text-1**	condition-3 **text + both** practice for **text-2**	condition-3 **text + both** practice for **text-1**
	recording	recording	recording	recording	recording	recording

Table 1: The number of learners in each group

group	A	B	C	D	E	F
#learners	11	10	16	18	13	12

3.3. The experimental design for the assessment

The experimental design using the six groups of learners is illustrated in Tab. 2. 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, recording was carried out. In session 3, all the learners practiced the same text again but they did with both kinds of prosody training stimuli. 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 (condition-3), and then recording was done. In Fig. 2, 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 (80×4=320) was rated in terms of the naturalness in accent and intonation realization for the utterances. Four experienced native teachers of Japanese, two of whom are the second and the third authors, participated in this rating. A seven-degree scale was used, where 1 and 7 indicate very poor and native-like, respectively. For each session, the averaged score over the four teachers was assigned to each learner.

3.4. Subjective assessment after the experiments

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

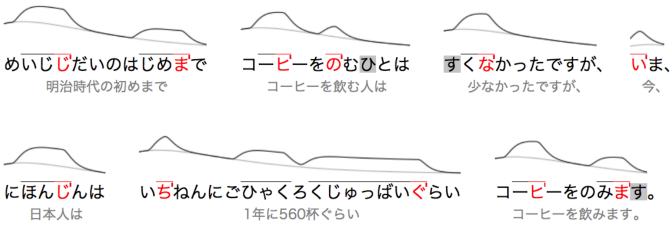

Figure 2: An example of visualized prosody of text-1

Table 3: The reading performance of each group

session 1	A	B	C	D	E	F
mean	4.93	4.05	4.56	3.82	5.08	4.29
s.d.	1.04	0.87	0.91	1.01	0.97	0.77
session 4	**A**	**B**	**C**	**D**	**E**	**F**
mean	5.52	4.83	5.11	4.13	5.23	4.73
s.d.	0.75	1.06	0.95	1.43	0.86	1.12

Table 4: The reading performances after re-selection

	A(11)	B(7)	C(15)	D(10)	E(11)	F(10)
mean	4.93	4.54	4.67	4.50	4.84	4.53
s.d.	1.04	0.37	0.84	0.44	0.81	0.58

4. Results and discussion

As told in Sect. 3.2, the eighty learners had been divided into six groups based on the duration of their learning Japanese and their native dialect. The reading performance of a group, however, may show an unignorable difference from that of another group. To compare the original reading performance among the groups, for each of sessions 1 and 4, we calculated mean and standard deviation of the scores that were assigned by native teachers. In session 1, the learners read a new text with no prosodic instructions and in session 4, they read another new text with two kinds of prosodic instructions. Tab. 3 shows their performances. It was found clearly that relatively poorer learners had been assigned to groups B, D, and F. To realize an even distribution of the performance among the groups, we re-selected the learners whose averaged performance over the four sessions varied from 4.0 to 6.5. The performances of session 1 after re-selection are shown in Tab. 4. T-tests showed no sig-

Table 5: The reading performance 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	+visualized	+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** %

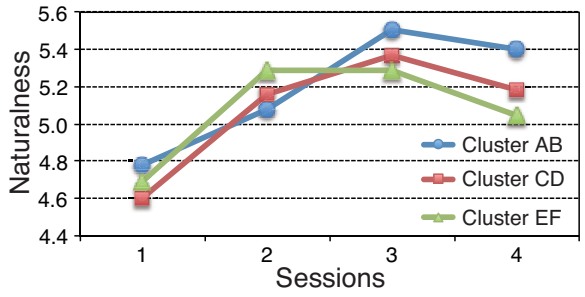

Figure 3: Naturalness scores for each session

nificant difference between any group pair. The parenthesized numbers are the number of the resulting learners for each group.

Since the re-selection process decreased the number of available learners for each group and t-tests between A and B, C and D, and E and F in Tab. 4 did not show any significant difference, the following discussion concerns the three clusters of AB, CD, and EF. Tab. 5 shows the reading performances of the three clusters for each session. Fig. 3 illustrates the performances. In session 1, even at the significance level of 10%, no significant difference was found between any pair of clusters.

Between sessions 1 and 2, naturalness is improved irrespective of the kind of prosodic instruction. At the significance level of 5% (p-value is 5%), however, only clusters CD and EF show significant improvements. This 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 problems in the quality of synthesized voices, but we have evidence to refute this possibility. For Suzuki-kun, four TTS systems were selected out of ten [23] that could generate synthetic voices with quality of a high enough level that the voices could be used in a language class. Further, the authors have already held three-hour tutorial workshops on OJAD seventy nine times internationally but no remarks of insufficient quality of the synthetic voices were 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 of visualized prosody. Why is visualized prosody 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 them. If this discussion is valid, the learner will prefer the order of visualized→auditory to that of auditory→visualized.

A preference for visualized prosody is found in the responses from the learners to the questions that were posed after recording. 80 % of the learners judged a combination of both kinds of prosody as the most effective. If they are presented sequentially, 73 % of the learners preferred the order of visualized→auditory. A main reason for this is that learning prosodic control is easier with visualized prosody than with auditory prosody. They say that, if given visualized prosody, they can select the segments of synthetic reading in advance to which they wish to pay close attention. From this, it appears appropriate to claim that when teachers introduce prosodic training to beginner learners, they should illustrate prosodic control for sentences before they present audio of those sentences.

The authors have received some emails from teachers who introduced prosody training to their class with 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 were examined. In China, all primary schoolers learn Mandarin as a common dialect[5]. They often learn the tones of Mandarin with visual illustrations. A strong preference for visualized prosody found in this paper 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, the findings above will be verified with high confidence.

In session 4, a new text was read with both kinds of prosody 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 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. 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 may imply that with sequential presentation of prosodic instructions, learners have acquired a better strategy to modify their reading compared to the case when both kinds of prosody are available simultaneously.

5. Conclusions and future work

Suzuki-kun is very unique in that it can visualize the full prosodic control for any Japanese text. It can also provide synthetic speech based on the visualized prosody. Quantitative assessment was done focusing on how learners' spoken Japanese changed with self-practice using Suzuki-kun. Significant improvements of prosodic naturalness were observed and visualized prosody was found to be by far more effective than auditory prosody. In future work, we will investigate whether similar preferences are found in learners whose L1 is not Chinese. Further, we are interested in developing a new module to rate learners' utterances through detecting prosodic errors and developing a prosody visualizer for other languages than Japanese.

6. Acknowledgement

The authors would like to thank Ms. Yi YANG (The University of Tokyo) and Ms. Shuju SHI (Beijing Language and Culture University) for their contributions of translating questionnaires and collecting responses from the learners. This work was supported financially by Grant-in-Aid for Scientific Research 26118002 and 26240022.

[5] If Chinese people speak in their native dialects, oral communication is difficult among people with different dialects.

7. References

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

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

[3] N. Minematsu, I. Nakamura, M. Suzuki, H. Hirano, C. Nakagawa, N. Nakamura, Y. Tagawa, K. Hirose, and H. Hashimoto, "Development and evaluation of online infrastructure to support teaching and learning of Japanese accent and intonation," *IEICE Trans*. vol.J96-D, no.10, 2496–2508 (2013, in Japanese).

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

[5] N. Minematsu, "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*, vol.19, no.1, 18–31 (2015, in Japanese)

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

[7] K. Isomura, "The current state of the Japanese accent education in for-eign countries, " Proc. Autumn meeting of the Society for Teaching Japanese" 211-212 (2001, in Japanese)

[8] Y. Siriphonphaiboon, "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 (2008, in Japanese)

[9] R. Ooyama, "A experimental study of speech training for natural Japanese," paper session 7-C, International Conference on Japanese Language Education (2014)

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

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

[12] A 1-min promotion video clip of OJAD: `https://goo.gl/k8rqKF`

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

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

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

[16] WinPitch: `http://www.winpitch.com`

[17] J. Komissarchi and E. Komissarchik, "Better Accent Tutor – Analysis and visualization of speech prosody," *Proc. InSTILL*, 86–89, (2000)

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

[19] 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)

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

[21] *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`

[22] JASLON: `http://www.jaslon.net`

[23] Speech synthesis software N2 and its TSS library SDK: `http://www.kddilabs.jp/products/audio/n2tts/product.html`