



Comparison of accent theories of Japanese using E2E speech synthesis in terms of their effectiveness for learners to acquire natural prosody

Nobuaki Minematsu¹, Fuki Yoshizawa¹, Tadashi Kumano², Kiyoshi Kurihara², Daisuke Saito¹

¹Graduate School of Engineering, The University of Tokyo,

²Science & Technology Research Laboratories, NHK,

mine@gavo.t.u-tokyo.ac.jp

Abstract

When teaching Japanese prosody to learners, two major theories of lexical accent are available, but they decompose an observed pitch pattern differently into lexical and phrasal components. Which theory is more valid pedagogically? In this work, a candidate answer is sought by comparing the two theories using text-to-speech conversion technologies. An end-to-end speech synthesizer is assumed to be a machine learner, which is built in two ways using a speech corpus annotated differently based on the two theories. Naturalness of the synthesized voices is compared between them, and it is shown that the two theories do not make any significant difference. With this finding in mind, from practitioners' point of view, a pedagogical suggestion is made on which theory should be used in which teaching context. The two theories correspond directly to opposite approaches of teaching and/or learning prosody, i.e., analytic and holistic.

Keywords: accent theories of Japanese, prosody training, end-to-end speech synthesis, naturalness, listening experiments, analytic vs. holistic

1. Introduction

Different teachers sometimes use different strategies to explain the same linguistic phenomena. Selection of the strategies may depend on various factors, such as learners' background knowledge, their cognitive capacity, etc. Some strategies are directly based on real experiences of teachers, i.e., practitioners, while others are purely derived from theories of linguistics, acoustics, education, psychology, etc.

In the present paper, we focus on two theoretical frameworks of Japanese lexical accent [1, 2, 3, 4], both of which attempt to model an observed pitch pattern. Unlike other languages, in Japanese, both lexical prosodic control and phrasal prosodic control are realized primarily as pitch movements. The two theories decompose an observed pitch pattern into the two levels of prosodic control in different ways. This difference may cause some confusion when the two theories are introduced into education. One and the same

prosodic phenomenon may be explained in different ways and, as many teachers and learners are not researchers, they may not understand how to interpret different explanations given from the two theories. A solution to this problematic situation may require experimental validation, where two groups of learners learn Japanese based on the two theories. The paper will also discuss which theory is more effective for the learners to acquire natural prosody. However, it may take a long time to validate experimentally.

In this paper, we attempt to provide a quick and technical solution, although we are very aware that it is not a complete solution. Recently, in speech synthesis, end-to-end (E2E) approaches have become very popular [5]. Being different from module-based development of speech synthesizers [6], the E2E approaches realize the process of text-to-speech conversion as one package, often implemented by training artificial neural networks. Explicit and hand-crafted knowledge for conversion is not needed at all to train an E2E system. A machine finds by itself how to map text to speech phonetically and prosodically to become a fluent reader. In this study, we radically assume that an E2E synthesizer is a simulator of a learner who wants to be a fluent reader *independently of any teacher*. E2E training is generally made with a speech corpus with its transcription, which can include some prosodic symbols. Different prosody annotations may result in different performances of conversion [7]. In experiments, we prepare a single speech corpus with different prosodic annotations based on the two theories. With this corpus, two E2E systems, i.e., two self-learners, are trained and compared with special attention to the naturalness of their prosodic control. Based on the experimental results and from practitioners' viewpoint, a practical and pedagogical suggestion is made to teachers on which theory should be adopted in which teaching context.

2. Accent and intonation in Japanese

2.1. Two theories of lexical accent of Japanese

Every content word has its own lexical accent and is characterized as mora-based pitch movement [8]. A binary value (H/L) is assigned to each mora. This means that 2^N H/L sequences are possible logically for an N-mora word but, in Tokyo Japanese, only N+1 types are allowed as lexical accents. The five accent types for four-mora words are illustrated in Figure 1. The following three facts are often explained in textbooks of Japanese. 1) The number of accent nuclei, which are the morae preceding a rapid pitch downfall and visualized as filled circle in Figure 1, is one or none in a word. 2) The accent type is identified only by the position of the accent nucleus. 3) A rapid pitch movement (L to H or H to L) is always found from the first mora to the second one in a word.

Recently, another very classical view of the lexical accent of Japanese [3] has been revisited, and the most influential accent dictionary of Japanese [4], published primarily for newscasters, has adopted this interpretation and the dictionary has been revised completely. As the two views or theories define the lexical accent differently, this revision caused some confusion among newscasters and teachers. While the above explanation describes the lexical accent is a mora-based H/L sequence, in the classical view, the lexical accent of a word is characterized only as the position of the accent nucleus and pitch control of any other parts in the word is independent of accentuation.

It seems that the first theory gives us a generative view of accentuation while the second theory gives us a discriminative view, but this differentiation is not adequate. As lexical prosodic control and phrasal prosodic control are both realized as pitch movements in Japanese, the second theory can still be interpreted as a generative one. In the following sections, the reason is explained. We also discuss how differently the two theories characterize the interactions between lexical and phrasal pitch controls.

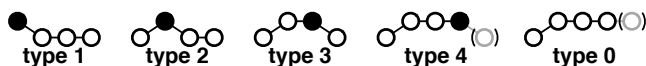


Fig. 1. Five accent types for four-mora words of Japanese

だいがく → だいがくが だいがくは だいがくを
 だいがくも だいがくに だいがくと
 :

Fig. 2. A word + a post-positional word = a bunsetsu

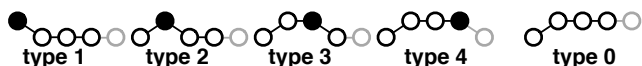


Fig. 3. Five accent types for five-mora bunsetsus of Japanese

2.2. Words, bunsetsus, and phrases

In Japanese, a content word is often concatenated with a short post-positional word to form the shortest phrase, called bunsetsu. The post-positional word

represents the case of the bunsetsu. With different post-positional words, a content word can generate different bunsetsus, shown in Figure 2.

Mora-based pitch control required to read aloud bunsetsus in Tokyo Japanese is shown in Figure 3, which is very similar to Figure 1. In classrooms, reading bunsetsus is taught to learners using Figure 3. Here, initial pitch rises are explicitly visualized because naturalness decreases without pitch rises. In the second and classical theory, this pitch movement is claimed strongly to be irrelevant to lexical accentuation. Then, how should that be interpreted?

Several and consecutive bunsetsus form a longer phrase, over which a global and gradual pitch declination is generally found. An example is shown in Figure 4. The first theory decomposes this pitch control as addition of two accent controls and a single two-bunsetsu-long intonational control [9], visualized in Figure 5. Here, pitch rises are regarded as accentual rises. The second theory explains the same pitch pattern as two consecutive intonational controls, corresponding to two bunsetsus, and pitch rises are regarded as intonational rises, not accentual rises, shown in Figure 6. Here, accent control is realized only as pitch fall, shown as a blue arrow. In the first theory, intonational control is a global declination of pitch, while in the second theory, it is a rapid pitch rise, followed by a subsequent gradual pitch declination. It is very clear that the two theories decompose an observed pitch pattern in different ways.



Fig. 4. An example of pitch movement in Japanese

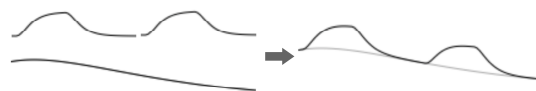
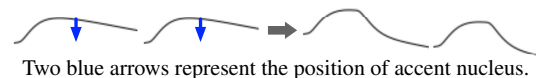


Fig. 5. Interpretation based on the first theory



Two blue arrows represent the position of accent nucleus.

Fig. 6. Interpretation based on the second theory

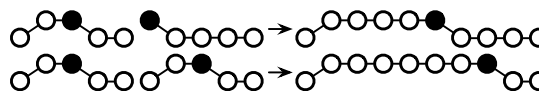


Fig. 7. Accent sandhi (change) found in compound nouns

This theoretical difference can be emphasized in a specific situation, known as accent sandhi. When a noun and another noun are concatenated to form a compound noun to represent a new semantic instance, accent sandhi inevitably takes place [8]. Figure 7 shows two examples, showing that a compound noun is uttered with a long accent type. The principle of accent sandhi for connecting nouns is that the nucleus

of the final noun remains as the nucleus of the resulting compound noun, and that some pitch rises and falls in the original nouns are removed. The two theories do not differ in explaining accent sandhi.

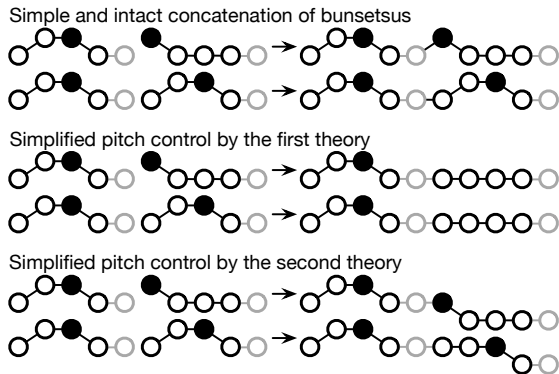


Fig. 8. Simple and intact concatenation of two bunsetsus and simplified pitch control explained by the two theories

Next, a focus is put on pitch control realized when connecting bunsetsus. A sequence of bunsetsus can be uttered naturally enough by concatenating them intactly, shown as top of Figure 8. However, pitch control is often simplified. The two theories explain accent control simplification in different ways, shown also in Figure 8. The first theory claims that the nucleus in the first bunsetsu remains as the nucleus of the resulting phrase, and that some pitch rises and falls in the original bunsetsus are removed. Thus, the pitch movement of the resulting phrase satisfies the three facts or conditions explained in Section 2.1. Since a two-bunsetsu phrase is uttered seemingly with its pseudo lexical accent, in the first theory, this pitch grouping is called accentual phrase. In this theory, it is often explained that bunsetsus as well as nouns are concatenated to form an accentual phrase [1, 2, 8, 9].

In the second theory, however, the concept of accentual phrase is strongly rejected [3, 4]. Simplified pitch movements in two bunsetsus explained by the second theory are shown again in Figure 8, where the utterance-initial pitch rise always stays and the other pitch rises are removed. Pitch falls are never removed, and more than one nucleus are found in the resulting phrase, not satisfying the first condition in Section 2.1. Only pitch downfalls (and their positions) are related to lexical accentuation, and they are never removed.

When native speakers utter the two bunsetsus in Figure 8 with simplified pitch control, which theory can explain acoustic observations more accurately? Answering this question is difficult, because it is widely known that a pitch fall observed in the n -th ($n > 1$) accent nucleus in a phrase becomes significantly smaller than that in the first nucleus, called catathesis [1, 2]. Thus, in the first theory, the n -th nuclei ($n > 1$) are ignored and the lexical accent theory, which explains accent sandhi observed in a compound noun,

is generalized and applied to pitch control when bunsetsus are concatenated to form a longer phrase. Here, some pitch falls and rises in the original bunsetsus are removed to satisfy the conditions in Section 2.1.

In the second theory, however, this generalization is prohibited, and it is claimed strongly that accentuation is a lexical control, not a phrasal control. In this theory, accent is characterized only as pitch downfall, and pitch rises are always intonational (phrasal) control and can be removed depending on context.

In the following section, the two theories are compared in terms of the naturalness of prosodic control realized in two E2E synthesizers, which are trained based on the two theories independently.

3. Experiments

3.1. Speech corpus with prosodic annotation

These days, several broadcasting companies have developed their own E2E synthesizers as virtual news casters [7]. To realize a high performance, not pure text but text with prosodic symbols is often used as input to text-to-speech conversion. In [7], Tacotron2 [5] and WaveNet [10] were used to generate spectrogram and speech waveforms, respectively. In Japanese, there are four kinds of letters and the KaNa system, which is one of them, was used as phonemic symbols for input text, and in [7], a KaNa sequence with prosodic symbols was converted to speech.

Our experimental comparison of the two theories was made on a development framework, which is similar to [7]. To simulate the learning process of Japanese learners, we used a part of the training corpus that was used for virtual casters. After some preliminary experiments, we used 3,000 read-aloud sentences, which were recorded from a professional female narrator. They were digitized at 16 bits and 22.05 kHz. For training virtual casters in [7], prosodic symbols were adequately attached to the full corpus, but special attention was not paid to compactness of prosodic annotations. Even with redundant annotations, machine learning algorithms are highly expected to select or merge those annotations effectively. In our experiments, out of the original prosodic symbols, two subsets were extracted so that they can correspond exactly to the two theories. Table 1 shows the full and original list of prosodic symbols prepared for this experiment. A KaNa sequence with full annotation is shown in Figure 9. It should be noted that the phonetic and/or prosodic meanings of these symbols are not given explicitly or even implicitly to train E2E systems. This is a reason why we regard the training procedure of E2E systems as a process of complete self-learning without any support from teachers.

Table 1. Prosodic symbols used for the experiments

1) inter-phrase symbols	
\$	intonational phrase boundary with a pause
%	intonational phrase boundary without a pause
	bunsetsu boundary [†]
2) accent-based inter-mora symbols	
/	rapid pitch rise
\	rapid pitch fall
-	flat pitch transition
3) other symbols	
#	beginning of a sentence
&	end of a sentence

[†] In spoken Tokyo Japanese, many bunsetsu boundaries are acoustically realized as accentual phrase boundaries.

Full set:

#ヨ\ク-ミ-ルト\$ヒヨ\O-ガ-ガ%ノ/コ\ッ-テ|イ/マ\ス&

Fig. 9. An example of text with the full set of symbols

3.2. Three different styles of prosodic annotation

Table 2 shows the full set and the two subsets of A and B, corresponding to the first and second theories, respectively. The inter-phrase symbols in Table 1 have inclusion relation in themselves. Every intonational boundary is a bunsetsu boundary and, since % is not allowed in prosodic annotation with subset A, every % in the original annotation is replaced by | for subset A (the first theory).

Subset A is derived by taking Japanese textbooks into account in which a sequence of KaNa letters is written always with a white space (“ ”) between two consecutive bunsetsus and always with the accent value (H/L) for every KaNa, as shown in Figure 3. Here, intonational phrase boundaries without a pause are not symbolized but bunsetsu boundary symbols are used instead. In subset B, a pitch rise is treated as inevitable in intonational control, not as accent-based inter-mora pitch transition (/).

The example shown in Figure 9 is re-annotated with subsets A and B in Figure 10. The differences among them seem to be minor, but they are very major theoretically. Comparison between A and B asks a very essential question on whether pitch rises are accentual rises or intonational rises.

Table 2. Three sets of symbols used for comparison

full	\$, %, , /, \, -, #, &	
subset A	\$, , /, \, -, #, &	for the first theory
subset B	\$, %, , , \, -, #, &	for the second theory

Subset A:

#ヨ\ク-ミ-ルト\$ヒヨ\O-ガ-ガ|ノ/コ\ッ-テ|イ/マ\ス&

Subset B:

#ヨ\ク-ミ-ルト\$ヒヨ\O-ガ-ガ%ノ-コ\ッ-テ|イ-マ\ス&

Fig. 10. The same text re-annotated with subsets A and B

3.3. Training of three E2E speech synthesizers

To train speech synthesizers, the training conditions used in [7] were adopted except for the following two

conditions. The first one is the number of training utterances, explained in Section 3.1. The other is pre-training. To simulate learners’ strategy and realize efficient learning, pre-training was introduced. Here, bunsetsus were segmented by referring to the original annotations, and pre-training was conducted with a full set of bunsetsu-segmented speech samples. After that, a main training procedure was run by using the original sentence-long utterances.

For waveform generation, we used the Griffin-Lim vocoder [11] instead of the WaveNet vocoder [10]. With the latter vocoder, naturalness of acoustically realized phonemes will be improved, but this study focuses only on naturalness of prosody acoustically realized from text with prosodic symbols. Further, as the Griffin-Lim vocoder can reduce the time required for training, we used it in the experiments.

3.4. Listening experiments using crowdsourcing

For training the three synthesizers, we used a part of the original speech corpus and for assessment, we used the remaining part of the corpus. Out of it, 100 sentences, which were as long as 60 to 80 morae, were randomly selected. In addition, only the sentences every mora of which was synthesized with adequate phonetic features were further selected. Finally, 25 sentences were randomly selected for assessment.

Three kinds of synthesized speech of the 25 sentences were used in the listening experiments on a crowdsourcing infrastructure, where thirty native adults were recruited. In a trial, two synthetic voices of the same sentence were presented with a 1-sec interval and subjects had to select the synthetic voice that sounded more natural in terms of accent and intonation control. Selection was done in a forced manner. The same pair was presented twice (two trials) but they were presented in a different order for each trial. One subject judged the prosodic naturalness of a pair of synthetic voices in the two trials and s/he judged five pairs, resulting in 10 trials in total.

3.5. Comparison of the two theories in terms of naturalness

Results of the experiments are summarized in Figure 11. Each colored bar represents the ratio of subjects who judged that method X or Y sounded more natural constantly for the two trials, or they sounded similar. Those who answered differently in the two trials were identified as subjects who judged that the two methods realized the same naturalness in terms of prosodic control. The error bars represent 95% confidence intervals for counts of judgment as higher naturalness.

When comparing the full set of symbols to subsets A and B, the number of constant judges is higher in the full set, but significant differences are not observed. Similarly, when comparing the two subsets of

A and B, the number of constant judges is larger in B, but a significant difference in counts of higher naturalness is not observed either between the two methods. We consider that prosodic annotation with the full set is somewhat redundant and similar enough prosodic control was realized only with subset A or B. Any symbol in input text is generally realized acoustically depending on its context. The same phoneme in different contexts is sometimes realized as different phones, called allophones. A prosodic symbol is also considered to be realized acoustically depending on its context and it is called as *allotonic* variation of that prosodic symbol [12]. This will be a reason why subset A or B can realize the same level of naturalness in prosodic control compared to the full set.

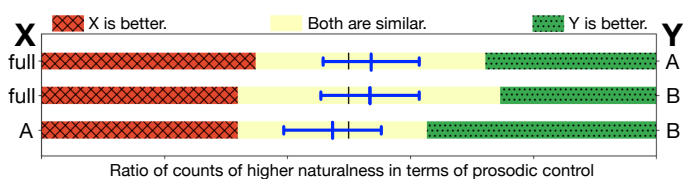


Fig. 11. Summary of results of the listening experiments

3.6. Comparison of the two theories in terms of attention level

Tactron2 uses attention mechanism, which controls the level of attention paid to each symbol in the text. By observing the attention level in the synthesizer trained with the full set, it is possible to discuss which symbols are more important. We calculated means and variances of the attention level of % and / for all the testing sentences. The two symbols are the only difference between subsets A and B. ANOVA showed that / tended to have a significantly ($p < 0.001$) higher level of attention. This may simply indicate that / is more important than %, implying that the first theory is more valid than the second theory, but we should discuss this very carefully. / is placed at a mora boundary exactly at which a pitch rise takes place acoustically, but % is a phrase boundary and is placed before the rise. This alignment gap may have influenced the level of attention. More detailed analysis is needed by using sentences with various prosodic and syntactic structures specified.

4. Practical suggestions to teachers

Results of the listening experiments showed that E2E systems, which were assumed to be machine self-learners, were so flexible that they became good and fluent readers either with subset A or B. Pedagogically speaking, however, we claim that it is not adequate to conclude that the two theories are effectively equivalent as teaching strategy. In this final section, we attempt to make some practical and pedagogical suggestions to teachers on how to differentiate the

two theories functionally depending on the context of teaching Japanese prosody to learners.

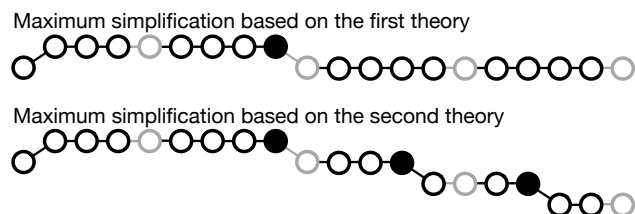


Fig. 12. Maximum simplification of prosodic control based on the two phonological theories of Japanese prosody

The first author developed OJAD [13], Online Japanese Accent Dictionary, which is a web-based educational infrastructure to teach or learn Japanese prosody, i.e., lexical and phrasal control of pitch. Before COVID-19, the first author had carried out more than 150 tutorial workshops of OJAD in over 40 countries. The workshops were held more often as training programs provided to native and non-native teachers of Japanese, not as tutorial lectures to learners. Through a series of the workshops and discussion with teachers, we found that the two theories correspond directly to two opposite ways of teaching Japanese prosody to learners, which are analytic and holistic approach of prosody training. To explain this, two examples of maximum simplification of pitch control are visualized in Figure 12 for concatenating four bunsetsus of accent types 0, 4, 3, 2 in this order. The two examples correspond to the two theories. By the first theory, shown on top in Figure 12, the maximum simplification makes pitch control in the resulting phrase similar to that in a long word. Teachers using the second theory, however, explain the maximally simplified pitch control, shown below in Figure 12, by introducing the principle of *sentence-level* pitch control as *after the initial rise, all falls*. They generalize this principle and apply it to word-level pitch control. Here, pitch rises are treated always as intonational rises and pitch falls are never removed.

On the other hand, as explained in Section 2.2, the first theory of pitch control is based on the lexical accent control, and it applies the lexical control when concatenating bunsetsus to form a phrase or a sentence. It is evident that the first theory supports a bottom-up and analytical approach of teaching Japanese prosody, while the second theory supports a top-down and holistic approach of teaching it to learners.

Figure 13 shows an example of bottom-up and gradual simplification of prosodic control for the four bunsetsus based on the first theory, which starts with intact concatenation of the four bunsetsus, shown with *. In the figure, the ending two bunsetsus are concatenated to form an accentual phrase and then, the first two bunsetsus are connected to form another accentual phrase. Finally, all four bunsetsus are

connected to form a single and long accentual phrase. Figure 14 shows an example of top-down and gradual complication of prosodic control for the four bunsetsus based on the second theory, which starts with the maximally simplified pitch pattern, shown with *. In the figure, an intonational pitch rise is added to the beginning of the third bunsetsu, and it is added again to that of the second bunsetsu. Finally, the complication process reaches the pitch control realized as intact concatenation. Which prosodic control should be used? Native speakers select an adequate one unconsciously based on which bunsetsu should be focused on semantically depending on the context.

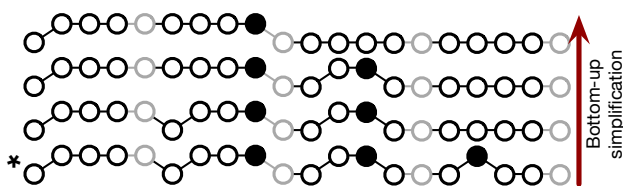


Fig. 13. Bottom-up simplification by the first theory

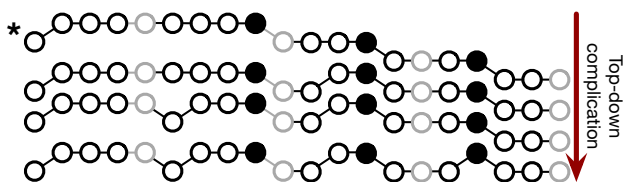


Fig. 14. Top-down complication by the second theory

In classrooms, which approach, analytical or holistic, should be adopted? In our experiments, no significant differences were found in perceived naturalness between the two theories. This indicates that, at least to machine learners, the two theories are equivalent in terms of their effectiveness. Here, we claim that the answer to the above question depends solely on learners' preference and/or teachers' preference. Some learners learn to pronounce words correctly first, and then to compose a sentence using them. In this context, we expect that learners prefer the first theory. They start with intact concatenation of bunsetsus and learn various pitch control by simplifying the initial control of pitch. On the other hand, many other learners learn Japanese by viewing Japanese animations or dramas. Here, they learn by heart orally many sentence-level expressions first, and then decompose them to find the individual words. In this context, we are sure that learners prefer the second theory. They start with a pitch pattern with initial rise and all falls and learn various pitch control by complicating the initial control of pitch.

Of course, teachers' preference may decide which approach to take. For example, if a teacher uses the Verbo-tonal method [14] to teach a language, s/he will prefer the second theory. In this method, prosodic control is taught to learners holistically using body gestures and hand gestures. Those teachers

supporting the Verbo-tonal method often claim that holistic learning should precede analytic learning.

5. Conclusions

In this paper, we attempted to provide an answer to a question on which of the two theories of Japanese lexical accent should be adopted in language classes. We assumed that an E2E speech synthesizer is completely self-learning, and two versions of E2E systems, corresponding to the two theories, were compared. Although no significant differences were found, we made several practical suggestions on which theory should be used in which teaching contexts.

6. References

- [1] Beckman, M. and Pierrehumbert, J., 1986. Intonational structure in Japanese and English, *Phonology Yearbook*, 3, 255–309.
- [2] Pierrehumbert, J. and Beckman, M., 1988. Japanese tone structure, *Linguistic Inquiry Monograph Series*, 15.
- [3] Kawakami, S., 1961. Tone of Japanese words and their concatenation with a euphonic change, *Study of Sound*, 9 (in Japanese).
- [4] Sugihara, M., 2011. Comparison of prosody theory of Tokyo Japanese based on verbal expressions –Toward revision of the NHK accent dictionary–, *Broadcasting Culture Research Institute*, 77–90 (in Japanese).
- [5] Wang, Y., Skerry-Ryan, R., Stanton, D., Wu, Y., Weiss, R. J., Jaitly, N., Yang, Z., Xiao, Y., Chen, Z., Bengio, S., Le, Q., Agiomyrgiannakis, Y., Clark, R., and Saurous, R. A., 2017. Tacotron: Towards end-to-end speech synthesis, in *Proc. INTERSPEECH*, 4006–4010.
- [6] Tokuda, K. and Zen, H., 2009. Fundamentals and recent advances in HMM-based speech synthesis (tutorial), in *Proc. INTERSPEECH*.
- [7] Kurihara, K., Seiyama, N., Kumano, T., and Imai, A., 2018. Evaluation of Japanese end-to-end speech synthesis method inputting kana and prosodic symbols, in *IEICE Technical Report*, SP2018-49 (in Japanese).
- [8] Sagisaka, Y., and Sato, H., 1983. Accentuation rules for Japanese word concatenation, *IEICE Transactions*, J66-D, 849–856 (in Japanese).
- [9] Fujisaki, H., and Hirose, K., 1984. Analysis of voice fundamental frequency contours for declarative sentences of Japanese, *Journal of the Acoustical Society of Japan (E)*, 9, 5, 233–242.
- [10] Van den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A., and Kavukcuoglu, K., 2016. WaveNet: A generative model for raw audio, *ArXiv:1609.03499*.
- [11] Griffin, D. W., and Lim, J. S., 1984. Signal estimation from modified short-time fourier transform, *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 32, 2, 236–243.
- [12] Lei, H., 2016. An acoustic analysis on the allotonic variation of the initial rise in Tokyo Japanese in native speakers and learners, in *Proc. ISAPh*, 48–51.
- [13] Minematsu, N., Nakamura, I., Suzuki, M., Hirano, H., Nakagawa, C., Nakamura, N., Tagawa, Y., Hirose, K., and Hashimoto, H., 2017. Development and evaluation of online infrastructure to aid teaching and learning of Japanese prosody, *Trans. IEICE*, E100-D, 4, 662–669.
- [14] Guberina, P., 2013. *The Verbotonal Method*. Artresor Naklada.