# USING LOCAL PHRASE DEPENDENCY STRUCTURE INFORMATION IN NEURAL SEQUENCE-TO-SEQUENCE SPEECH SYNTHESIS

*Nobuyoshi Kaiki*[†], *Sakriani Sakti*[†‡] *and Satoshi Nakamura*[†‡]

†Nara Institute of Science and Technology, Japan, ‡RIKEN AIP, Japan

## ABSTRACT

We introduce end-to-end text-to-speech synthesis (TTS) with prosodic symbols that represent phrase components based on local syntactic dependency structures for synthesizing Japanese speech with natural prosody. We propose two TTS models: 1) one with prosodic symbols representing the syntactic dependency distance at the phrase boundaries and 2) another with prosodic symbols that reflect a superimposed model of the phrase and accent components based on an F0 generation control mechanism. Using these two models, we observed 1) pause insertion that indicates the phrase boundary and 2) F0 resetting at the right-branching boundaries. To verify the effectiveness of these two proposed models against the conventional model using only accent components, we conducted an AB test as a subjective evaluation. Our result confirmed that synthetic speech with natural prosody, which reflects the corresponding intention to the utterance, was generated using the local phrase dependency information of sentences and the F0 generation model in a Japanese end-to-end TTS.

***Index Terms***—Neural end-to-end text-to-speech speech synthesis, local phrase dependency structure, prosodic symbol

## 1. INTRODUCTION

Recently, the sound quality of text-to-speech synthesis using neural networks has significantly improved. Much progress has also been made in the research and the development of deep learning. End-to-end speech synthesis, which directly integrates speech using text strings and phonemes as input, has also been widely studied.

However, the naturalness of prosody remains inadequate, especially in such synthetic speech as reading storybooks and dialogues. Moreover, the difficulties of producing the naturalness of prosody vary depending on the language. For example, in Japanese, the addition of accent information dramatically improves the naturalness of prosody [1] [2].

Several regularizations, proposed to improve synthetic speech quality, were mainly based on the phrase structure of sentences [3]-[6]. The F0 pattern model is represented by the superimposition of a local accented phrase component (accent component) and a component spanning the same prosodic cohesion (phrase component) that maintains the global descent property. This property is expressed by the dependency relation between adjacent phrases. When the phrase immediately before the phrase boundary directly modifies the phrase immediately after it (i.e., the left-branching boundary), the global descent property is preserved. On the other hand, when a phrase modifies a more backward nonadjacent phrase (the right-branching boundary), so-called F0 resetting occurs. The analysis and regularization of pause insertions have also shown that long pauses are more likely to be inserted at right-branching boundaries that contain punctuation marks [7].

Fujimoto [1] showed that the naturalness of synthesized speech can be improved by simultaneously inputting pitch height and phoneme one-hot vectors into Tacotron 2 [9]. However, further adding input with full context information, including such linguistic details as a word's part of speech, contributed very little to the naturalness of the synthesized speech.

Kurihara [8] also demonstrated that the naturalness of synthesized speech was improved more when such prosodic symbols as phonemes, accents, and phrase boundaries were added to the input of Tacotron 2 [9] compared to the input of only phoneme symbols. However, that work failed to specify how the boundary information was inserted. In addition, since the effect of F0 resetting at phrase boundaries has not been clarified, further verification is required.

We applied information on the syntactic dependency distance of sentences to Japanese end-to-end speech synthesis to produce more natural synthetic speech. To model the control mechanism of prosody generation, we also investigated a model that incorporates prosodic control symbols for two components with different properties: accent and phrase components. We also present our evaluation results of the obtained synthetic speech for these two models using prosodic symbols representing the local phrase dependency structure of the syntax.

## 2. PROPOSED MODEL UTILIZING PROSODY SYMBOLS TO REPRESENT PHRASE STRUCTURE

Figure 1 shows the process flow of the TTS used by Kurihara [2]. Given the text, the front-end module creates phonemic and prosodic symbols. The latter are shown in

Table 1. This paper enhanced and expanded the prosodic symbols proposed by Kurihara [2]. Our modified prosodic symbols (shown in Table 2) include those representing the local phrase dependency structure of sentences. These prosodic symbols are input to Tacotron 2 with phonemic symbols.

The main difference with Kurihara's approach [2] is that instead of using # to represent the accent phrase boundary, we use #1 to #6, indicating the distance of the syntactic dependency of the accent phrases, to represent the local phrase dependency structure of a parse tree. In addition, these syntactic dependency distances are inserted at all the accented phrase boundaries. #1 marks the phrase boundaries where the preceding phrase directly modifies the following adjacent phrase (i.e., left-branching boundary). For example, Fig. 2 shows two possible syntax trees for the syntactically ambiguous sentence, *The policeman chased the thief who ran away*, and the phonemic and prosodic symbols corresponding to each model. In the first syntax tree, #4 is placed at the phrase boundary between *the policeman* and *ran* because *the policeman* is syntactically attached to *chased*. Since *ran* directly modifies *away*, #1 is placed at the phrase boundary between *ran* and *away*. However, for processing purposes, when the dependency distance between the phrases is six or more, the phrase structure is expressed as #6.

We also propose a model based on the F0 generation control mechanism. Table 3 shows the prosody symbols used in it. For the accent component, we use control symbol "/" to indicate the rising edge of the accent command and control symbol "\" to indicate the falling edge. For the phrase component, prosody symbols #2 to #6 indicate the length of the syntactic dependency, assuming that the length indicates the strength of the phrase command.

In this experiment, we abbreviated the model that uses only accent information and no prosody marks to represent the phrase structure as the baseline model and compare it with our two proposed models. The model, which additionally inputs prosody marks to represent the phrase structure information (Table 2), is denoted as proposed model 1, and the model with the F0 generation control mechanism (Table 3) is denoted as proposed model 2.

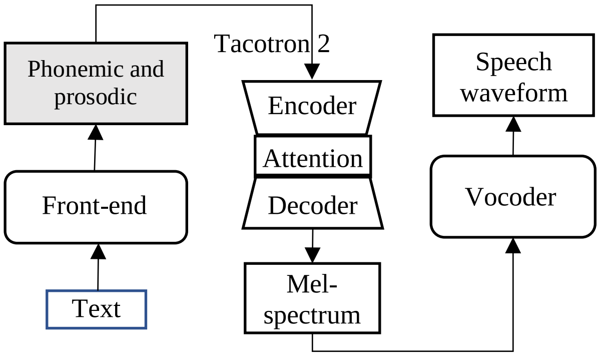

Fig. 1 Process flow of TTS using phonemic and prosodic symbols

In natural speech prosody, pauses are often inserted when the text has a punctuation mark. However, pauses might not be inserted even when there is a punctuation mark; they may be inserted even when there is a punctuation mark at the phrase boundaries where the syntactic dependency distance is long [7]. To represent this phenomenon in the end-to-end model, we used a comma "," instead of the prosodic pause, "_".

## 3. SPEECH DATABASE

In this study, we used the speech data of a single speaker reading a story rather than individual sentences to improve the naturalness of reading voices. We created a new dataset for our experiment using the online text and an oral transcription of the *Arabian Nights* and its reading voice: 190 sentences, average 8 minutes and 21 seconds, total 26 hours and 26 minutes [11].

Table 1 Prosodic symbols (baseline)

| Feature | Prosodic symbols |
|---|---|
| Initial rising | ^ |
| Accent nucleus | ! |
| Accent phrase boundary | # |
| EOS (declarative) | （ |
| EOS (interrogative) | ? |
| Pause | _ |

Table 2 Prosodic symbols representing local phrase dependency structure (proposed 1)

| Feature | Prosodic symbols |
|---|---|
| Initial rising | ^ |
| Accent nucleus | ! |
| Syntactic dependency distance<br>(accent phrase boundary) | #1, #2, #3,<br>#4, #5, #6 |
| Punctuation mark | , |

Table 3 Prosodic symbols based on F0 generation mechanism (proposed 2)

| Feature | Prosodic symbols |
|---|---|
| Accent command (rising) | / |
| Accent command (falling) | \ |
| Phrase command<br>(syntactic dependency distance） | #2, #3,<br>#4, #5, #6 |
| Punctuation mark | , |

The text was manually stripped of unnecessary symbols and line breaks and divided into sentence units based on punctuation, line breaks, and so on. Such information as phonemes, accent types, and accent phrase boundaries was automatically assigned using Open JTalk [12]. Morphological analysis results from Open JTalk were input into ChaboCha [13], and the phrase boundary information, which indicates a sentence's phrase structure, was automatically assigned based on the obtained syntax tree representation.

The speech data were automatically segmented into sentence units using the CTC Segmentation [14] of ESPnet

Input: 警官は走って逃げる泥棒を追いかけた。(The policeman chased after the thief who ran away)

Phoneme: ke ekaNwa ha shi clte ni geru do robooo o ikake ta
Baseline (accent): ke^ekaNwa# ha^shi!clte# ni^ge!ru# do^robooo# o^ikake!ta(
Proposed 1 (accent + prosodic symbols representing local phrase dependency structure):
Syntax tree 1: ke^ekaNwa#4 ha^shi!clte#1 ni^ge!ru#1 do^robooo#1 o^ikake!ta
Syntax tree 2: ke^ekaNwa#1 ha^shi!clte#3 ni^ge!ru#1 do^robooo#1 o^ikake!ta
Proposed 2 (prosodic symbols based on F0 generation mechanism):
Syntax tree 1: ke/ekaNwa\#4 ha/shi\clte ni/ge\ru do/robooo\ o/ikake\ta
Syntax tree 2: ke/ekaNwa\ ha/shi\clte#3 ni/ge\ru do/robooo\ o/ikake\ta

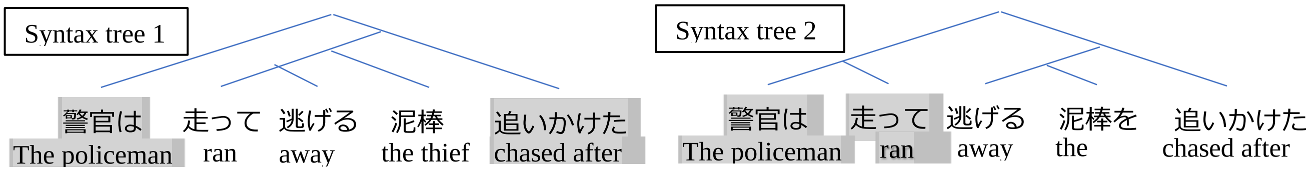

Fig. 2 Comparison of phonemes and prosodic symbols utilized in baseline and proposed method based on two syntax tree candidates

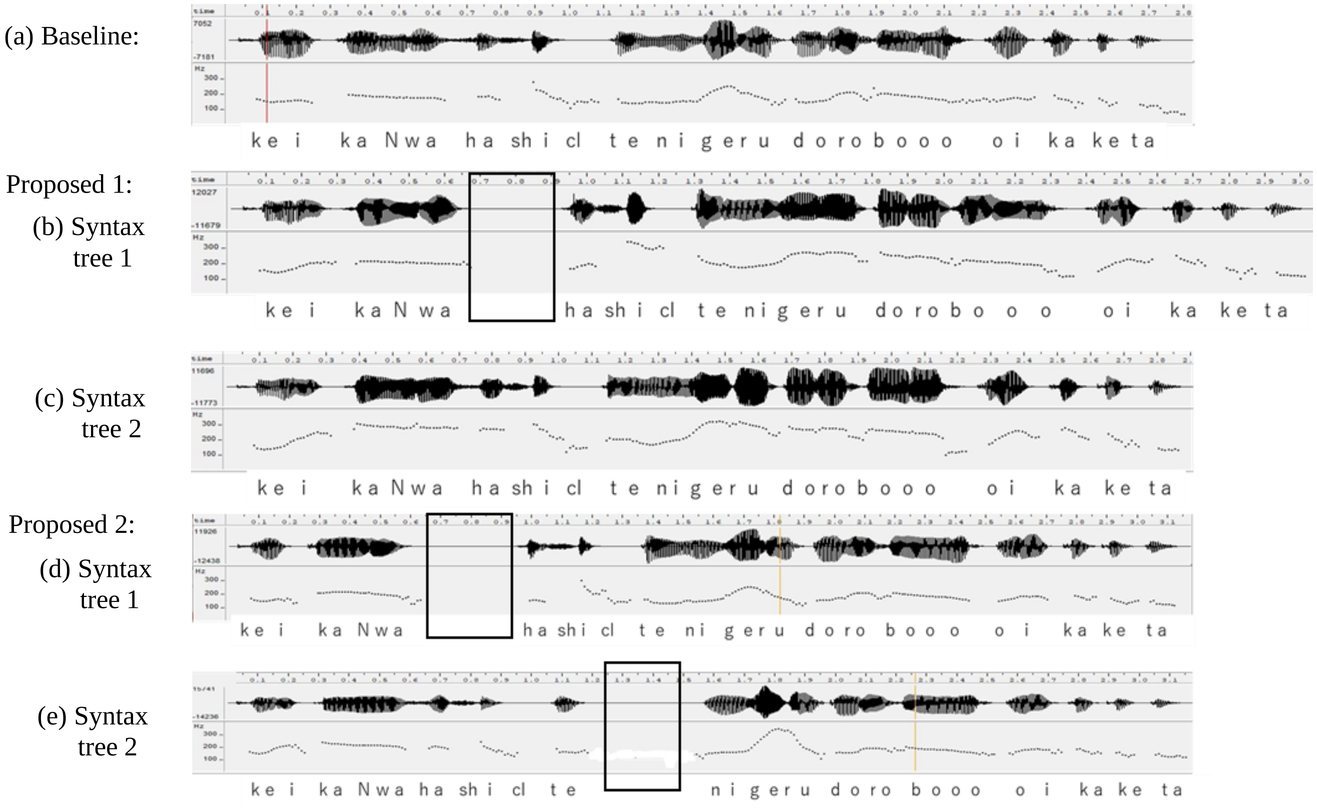

Fig. 3 Pause generation that reflects phrase dependency structure of syntax trees †

( † Synthesis speech samples available at https://sites.google.com/view/synthesis-speech/%E3%83%9B%E3%83%BC%E3%83%A0 )

[15]. In addition, phoneme-based alignments [16] corresponding to the readings obtained from Open JTalk were taken for each sentence, and complete context labels were created with other information from Open JTalk. During these processes, 11,615 sentences were used as the dataset, excluding sentence segmentation errors, phoneme alignment errors, and sentences longer than 20 seconds.

## 4. EXPERIMENT

In this experiment, Japanese Tacotron 2 [9] [10], provided by ESPNet2 [14], generated a mel-spectrum from the input sequence. Parallel-Wavegan [17] was used as the vocoder to generate speech waveforms from the mel-spectrum. Of the 11,615 sentences, 11,115 were used for training and 250 each were used for validation and testing.

### 4.1 Objective evaluation

In this section, we show that the synthesized speech of the trained proposed model reflects the phrase structure of the syntax tree: 1) the pauses indicating phrase breaks are automatically inserted at the right-branching boundary without punctuation marks, and 2) the phenomenon of F0 resetting occurs. Using a baseline (conventional) model and our two proposed models, we input the syntactically ambiguous sentence, *The policeman chased the thief who ran away*, and generated synthetic speech to investigate its speech waveform and F0. This syntactically ambiguous sentence is assumed to have two syntax trees (Fig. 2). Fig. 3(a) shows the synthesized speech waveform and the F0 generated by the baseline model using only accent control symbols. Figs. 3(b)-(e) show the synthetic speech waveforms and the F0 corresponding to the two syntax trees generated by proposed models 1 and 2. For syntax tree 1, no pauses were generated at the right-branching boundary between *policeman is* and *ran* in the baseline synthetic speech model. In contrast, proposed models 1 and 2 both generated pauses that indicate phrase boundaries that clarify the syntax. For syntax tree 2, at the right-branching phrase boundary between *ran* and *away*,, the synthetic speech of model 2 similarly contains a pause indicating a phrase boundary that clarifies the syntactic disambiguation. No pauses were generated in the baseline or in proposed model 1.

Next we show an example where F0 resetting occurs at the right-branching boundary. Fig. 4 shows the three possible syntax trees for such syntactically ambiguous sentences as *I like the white house with the big roof*. Fig. 5(a) shows the F0 contour of the synthesized speech produced by the baseline model, which is based only on accent control symbols. Figs. 5(b)-(d) show the F0 contour of the synthetic speech produced by proposed model 2 for the three syntax trees in Fig. 3. For syntax tree 2, where *white* is syntactically attached to *house*, F0 resetting occurs at the phrasal boundary between

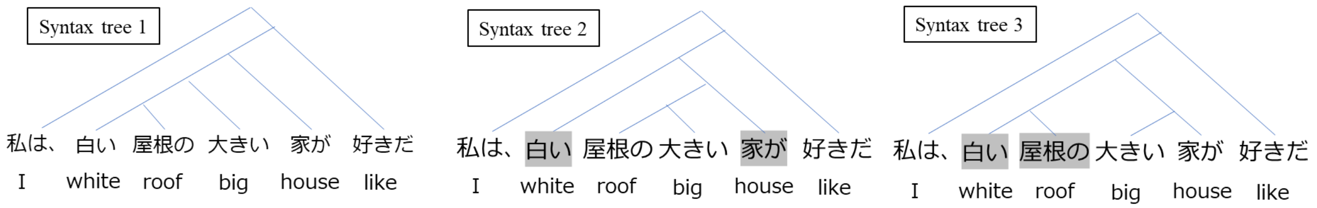
Syntax tree 1
私は、 白い 屋根の 大きい 家が 好きだ
I white roof big house like
Syntax tree 2
私は、 白い 屋根の 大きい 家が 好きだ
I white roof big house like
Syntax tree 3
私は、 白い 屋根の 大きい 家が 好きだ
I white roof big house like

Fig. 4 Three different phrase dependency trees

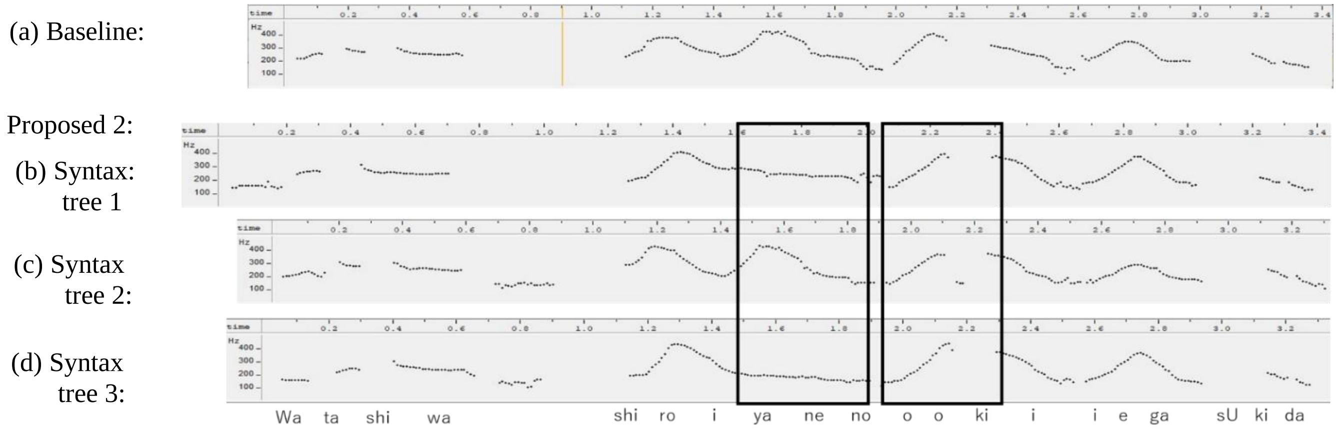

Fig. 5 F0 contour of synthesis speech of three different phrase dependency trees

*white* and *roof*, which is right-branching, and the F0 contour reflects the tree structure. However, for syntax tree 3, where *white roof* is syntactically attached to *house*,, the phrase boundary between *roof* and *large*, which is also right-branching, has the same F0 contour as syntax trees 1 and 2, which are left-branching boundaries. No F0 resetting was observed. However, the maximum F0 value of the accent phrase *big* in syntax tree 3 was slightly larger than the maximum F0 value of the accent phrase in the other syntax trees.

These results show that the baseline models did not generate pauses or F0 resetting at the right-branching boundaries, although the two proposed models did generate them. Note that these phenomena do not always occur at the right-branching boundary; they can be learned and generated based on the speech data used for training.

## 4.2 Subjective evaluation

To comprehensively evaluate the speech synthesized by the proposed model, which introduced and learned the prosodic symbols that use the information of the phrase structure, we conducted listening experiments on the naturalness of the synthesized speech.

### *4.2.1 Evaluation data*

For the evaluation data, we used data randomly extracted from 250 sentences for evaluation other than the data used for training and validation at the time of the model creation. However, to verify the effect of the local phrase dependency structure information in the listening experiment, we used 20 sentences, excluding those that satisfied at least one of the following three conditions:

1) sentences in which the phrase boundary consists only of a left-branching boundary;

2) sentences with three or fewer accent phrases;

3) sentences with synthesized speech longer than eight seconds.

For the baseline model and each of the two proposed models, 20 sentences were synthesized, for a total of 60 sentences. They were used as evaluation data for the listening experiment.

### *4.2.2 Evaluation method*

We assessed the quality of the baseline and the proposed models focusing only on the naturalness of prosody. As different systems may have the same overall naturalness quality but different only in prosody, we evaluated them with a paired comparison instead of the mean opinion score (MOS) test. The 13 evaluators were all native Japanese speakers. The evaluation was done by listening to a pair of synthetic speeches and making a forced judgment as to which synthetic speech had more natural prosody. 60 pairs of 20 sentences each were randomly presented to the listeners: baseline model and proposed model 1, baseline model and proposed model 2, and proposed model 1 and proposed model 2. We also randomized the presentation order of each pair of synthetic speech models.

### *4.2.3 Evaluation results and discussion*

We obtained the following two results from our subjective evaluation experiment (Fig. 6). .

The synthesized speech of proposed models 1 and 2, which use prosodic symbols with information on a sentence’s phrase structure, is significantly (1% level) more natural than the synthesized speech of the baseline model, which only uses prosodic symbols with accent components (proposed model 1: 68%, proposed model 2: 62%).

In the comparison between the baseline model and proposed models 1 and 2, proposed model 1 was judged more natural than the synthesized speech of proposed model 2. The difference between proposed models 1 and 2 was not significant. In a direct comparison of the synthesized speech of models 1 and 2, the synthesized speech of model 2 was judged 53% more natural than that of model 1. Similarly, there was no significant difference between them. In this subjective evaluation, it was impossible to determine the difference in naturalness between the synthesized speech of proposed models 1 and 2. The improvement in prosodic naturalness of the two proposed methods compared to the baseline can be attributed to the fact that they were able to generate the insertion of pauses and the F0 resetting according to the intention of the sentence. However, among the proposed models, it seems that there was no significant difference. The reason might be because the accent component can correctly represent the Japanese accent type in both methods. Furthermore, the assumption that the F0 resetting does not occur when the length of the syntactic dependency is 1 is almost correct.

# 5. CONCLUSION

We introduced new prosodic control symbols that represent phrase components based on the local phrase dependency structure to end-to-end speech synthesis for

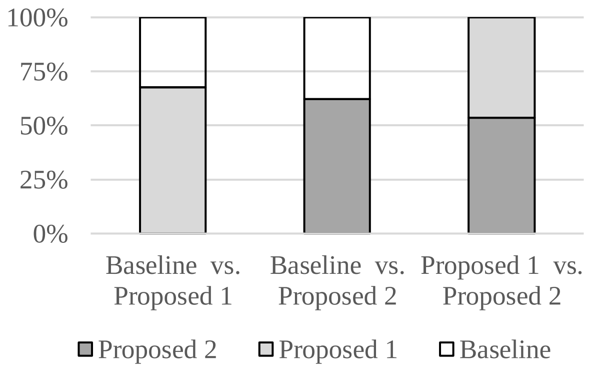

Fig. 6 Subjective evaluation (AB test) of naturalness for prosodic symbols

synthesizing Japanese speech with natural prosody. To represent a sentence’s local phrase dependency structure, we proposed two new models: 1) one that adds prosodic symbols to the phrase boundary to indicate the distance of the syntactic dependency, and 2) another that adopts prosodic control symbols that reflect a superimposed model of phrase and accent components based on a F0 generation control mechanism.

After examining the synthetic speech using these two models, we confirmed the following prosodic cues at the right-branching boundary, reflecting the structure of the sentence syntax tree: 1) Pauses are produced even though there are no punctuation marks, and 2) F0 resetting of the phrase component was observed.

We conducted subjective evaluation experiments on the synthesized speech of the end-to-end speech synthesis using these two new proposed models and a baseline (conventional) model. Proposed model 1, which introduced a new prosodic control symbol to indicate the depth of the phrase boundaries representing the local phrase dependency structure, improved the naturalness of prosody in the synthesized speech by 68% more than the baseline (conventional) model, which used only accent information as the prosodic control symbol. For proposed model 2, which uses prosodic control symbols based on a F0 generation control mechanism, the naturalness of the prosody of the synthesized speech was improved by 62% more than the baseline model.

These experimental results confirmed that by incorporating information about the local phrase dependency structure of a sentence and a prosody generation model into Japanese end-to-end speech synthesis, synthetic speech can be produced with natural prosody that more correctly reflects the speaker's intention.

To produce more natural synthetic speech, we will study the phrase structure of the relevant phrase boundary and the structure of the preceding phrases, parallel relations, etc., and use more detailed information about such segmentation as part-of-speech information.

## 6. ACKNOWLEDGMENTS

We thank Takehazuchi [11] for making the reading voices and texts we used in this research widely available and allowing us to use them for non-commercial purposes.

This research was partially supported by JSPS Grants-in-Aid for Scientific Research JP21H05054 and JP21H03467.

## 7. REFERENCES

[1] T. Fujimoto, K. Hashimoto, K. Oura, Y. Nankaku, and K. Tokuda, "Impacts of Input Linguistic Feature Representation on Japanese End-to-End Speech Synthesis," Proc. of 10th ISCA Speech Synthesis Workshop (SSW), pp. 166-171, Vienna, Austria, Sep. 2019.

[2] K. Kurihara, N. Seiyama, and T. Kumano, "Prosodic Features Control by Symbols as Input of Sequence-to-Sequence Acoustic Modeling for Neural TTS," IEICE Trans. Inf. & Syst., Vol. E104-D, No. 2, pp. 302-311, Feb. 2021.

[3] K. Hakoda and H. Sato, "Prosodic rules in connected speech synthesis," Trans. IECE Japan Vol. J63-D No. 9, pp. 715-722, Sept. 1980. (in Japanese)

[4] K. Hirose, H. Fujisaki, H. Kawai, and M. Yamaguchi, "Speech synthesis of sentences based on a model of fundamental frequency contour generation," Trans. IECE Japan vol. J72-A, No. 1, pp. 32-40, Jan. 1989. (in Japanese)

[5] M. Abe and H. Sato, "Two-Layer F0 control model using syllable based F0 units," Journal of the Acoustical Society of Japan, vol. 49, No. 10, pp. 682-690, Oct. 1993. (in Japanese)

[6] N. Kaiki and Y. Sagisaka, "F0 Control Based on Local Phrase Dependency Structure," Trans. IECE Japan vol. J83-D-II, No. 9, pp. 1853-1860, Sept. 2000. (in Japanese)

[7] N. Kaiki and Y. Sagisaka, "Study on Pause Insertion Rules Based on Local Phrase Dependency Structure," Trans. IECE Japan vol. J79-D-II, No. 9, pp. 1455-1463, Sept. 1996. (in Japanese)

[8] K. Kurihara, N. Seiyama, T. Kumano, and A. Imai, “Study of Japanese end-to-end speech synthesis method that inputting kana and prosodic symbols,” Proc. Autumn Meeting of Acoustical Society of Japan, pp. 1083–1084, 2018. (in Japanese)

[9] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, and Q. Le, "Tacotron: Towards end-to-end speech synthesis," in Proc. Interspeech, pp. 4006-4010. Aug. 2017.

[10] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, RJ Skerry-Ryan, Rif A. Saurous, Y. Agiomyrgiannakis, and Y. Wu, "Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions," Proc. of ICASSP 2018.

[11] Takehazuchi, "On the reading of relieving stories Arabian Nights oral translation," https://o-keil.com/okinu-ba-ba/wordpress/?p=818 (in Japanese)

[12] "Open JTalk," http://open-jtalk.sourceforge.net/.

[13] "CaboCha/Nankai: Yet Another Japanese Dependency Structure Analyzer,"
http://taku910.github.io/cabocha/

[14] J. Nishitoba, "Introduction to CTC Segmentation," https://tech.retrieva.jp/entry/2020/10/02/143338 (in Japanese)

[15] S. Watanabe, T. Hori, S. Karita, T. Hayashi, J. Nishitoba, Y. Unno, N.E.Y. Soplin, J. Heymann, M. Wiesner, N. Chen, A. Rendouchintala, and T. Ochiai, "ESPnet: End-to-End Speech Processing Toolkit," Proc. Interspeech, pp. 2207-2211, 2018, https://github.com/espnet/espnet.

[16] "Montreal Forced Aligner," https://montrealcorpustools.github.io/Montreal-Forced-Aligner/

[17] R. Yamamoto, E. Song, K. Eunwoo, and J. Kim, “Parallel WaveGAN: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram,” Proc. of ICASSP, pp. 6199-6203, Feb. 2020.