Prosodic Features Control by Symbols as Input of Sequence-to-Sequence Acoustic Modeling for Neural TTS

Kiyoshi KURIHARA†, Nonmember, Nobumasa SEIYAMA†, Member, and Tadashi KUMANO†, Nonmember

SUMMARY This paper describes a method to control prosodic features using phonetic and prosodic symbols as input of attention-based sequence-to-sequence (seq2seq) acoustic modeling (AM) for neural text-to-speech (TTS). The method involves inserting a sequence of prosodic symbols between phonetic symbols that are then used to reproduce prosodic acoustic features, i.e., accents, pauses, accent breaks, and sentence endings, in several seq2seq AM methods. The proposed phonetic and prosodic labels have simple descriptions and a low production cost. By contrast, the labels of conventional statistical parametric speech synthesis methods are complicated, and the cost of time alignments such as aligning the boundaries of phonemes is high. The proposed method does not need the boundary positions of phonemes. We propose an automatic conversion method for conventional labels and show how to automatically reproduce pitch accents and phonemes. The results of objective and subjective evaluations show the effectiveness of our method.

key words: statistical parametric speech synthesis, end-to-end speech synthesis, phonetic and prosodic symbols, sequence-to-sequence acoustic modeling

1. Introduction

Deep neural network (DNN) based text-to-speech (TTS) [1] was developed using hidden Markov model (HMM)-statistical parametric speech synthesis (SPSS) [2]. End-to-end speech synthesis has been actively researched ever since Tacotron 2 [3] first produced English speech comparable in quality to that of human speech. Tacotron 2 uses a sequence-to-sequence (seq2seq) model, as do similar methods such as Tacotron [4], Char2Wav [5], Voiceloop [6], Deep Voice 3 [7], and Transformer-based TTS [8].

The work presented here focuses on the control of prosodic features which represent accent, etc., by adding a sequence of prosodic symbols between the phonetic symbols of the input sequence and reproducing the pitch-accent accurately in seq2seq acoustic modeling (AM) [9].

In related research, the system of Yasuda [10] considered input pitch-accent information in Tacotron, but the system controlled only accent information, and the quality of the synthesized speech was worse than that of conventional SPSS. Moreover, it was only compatible with Tacotron. Yasuda’s method inputs a phoneme sequence and accentual-type sequence separately and embeds them separately. Our method differs from Yasuda’s in terms of the accent sequence format. In our method, the phonetic and prosodic symbols are merged into one sequence. Our method is thus more readable than their method that has to separately input sequences. In addition, our method does not require changes to be made to the model. As well, the experimental results reported in Sect. 3.3.1 indicate that multiple seq2seq AM methods are effective for inputting the proposed notation. In addition, Fujimoto [11] conducted an experiment comparing the suitability of phoneme and mora as units of the input sequence, which is one-hot vectors or linguistic features, of seq2seq AM. The sequence in the phoneme method is about twice as long as that in the mora method, but long sequences tend to occur alignment error [12]. There is a possibility that the phoneme method causes more alignment errors than the mora method even though there is no difference between phoneme and mora in terms of pronunciation. To examine is issue, we performed an experimental comparison between the phoneme method and mora method using the JSUT corpus [13] (see Sect. 3.3.1). The system of Okamoto [14] inputted full-context (FC) labels [15] to Tacotron 2, but it could not input symbols directly and was only compatible with FC labels that have poor readability. Although seq2seq AM can handle symbols directly, the system of Okamoto does not take advantage of the benefits of simple input symbols. Furthermore, it is only compatible with Tacotron 2. Shechtman [16] proposed a method that reproduces prosody information. The attention section in the seq2seq AM of this method has a recurrent architecture. The method produces expressive speech and can control the duration, but it cannot replicate prosodic features. While the present study is related to these recent approaches in seq2seq AM, it is compatible with multiple seq2seq AM methods, controls prosodic features, and has directly readable and writable data descriptions, aspects that were not considered in the earlier studies.

We developed a method of controlling the prosodic features that works by inputting a symbol between phonetic symbols. Seq2seq AM corresponds to inputs of phonetic symbols [4], but it has been confirmed that the accents cannot be reproduced by learning input consisting of Japanese phonetic symbols [17] due to the lack of accent information. Furthermore, normal Japanese text contains multiple readings of characters, which makes it unsuitable for seq2seq AM training. To solve these problems, we propose a notation method that represents the uttered speech uniquely by using phonetic and prosodic (PP) labels. The prosodic symbols complement the acoustic features between phonetic symbols and can reproduce accurate utterances. We
conducted experiments showing that the proposed method is effective for pitch-accent languages such as Japanese.

PP labels can be generated automatically, and using them to make annotations is simple. The conventional SPSS with a DNN [1] uses time-aligned FC labels. A high-quality SPSS method requires accurate time alignments, and the cost of accurately determining time alignments is high. Our method does not require time alignment because the seq2seq model-based TTS can learn directly from the input symbols. We propose a way of automatically converting FC labels into PP labels. FC labels contain linguistic features such as phoneme, accent information, accentual phrase boundary, end-of-sentence (EOS), and pause. This information is rearranged according to the proposed rules, and the results are natural utterances produced in seq2seq AM. Evaluations have shown that using PP labels yields more natural speech than conventional SPSS. This shows the possibility of their general use with seq2seq AM methods. We confirmed that PP labels can control prosodic features by using three seq2seq AM methods in TTS for Japanese.

2. Proposed Method

Seq2seq AM is a method that generates mel-spectrograms representing inputted symbols. Since the advent of WaveNet [18], it has enabled high-quality audio to be generated from mel-spectrograms. The advantage of this approach is that mel-spectrogram can be converted directly into a waveform. To reproduce prosodic-acoustic features such as accent for seq2seq AM, we insert prosodic symbols between phonetic symbols and use the result as the input of seq2seq AM. This method represents the uttered speech uniquely, especially pitch accents, and should be easily readable by people. It is highly readable, so the labels can be read and understood directly, and a front-end is not required, depending on how it is used.

Figure 1 shows an overview of a TTS system incorporating the proposed method. This diagram consists of three parts. The first part is the PP label generation. By combining the FC label conversion method with the front-end, the PP labels can be automatically generated from the text. In addition, the FC labels can be used and the information in the existing labels, can be discarded. The second part is the mel-spectrogram estimation with seq2seq AM. This part can be replaced with other seq2seq architectures. The third part is waveform generation using WaveNet or a Griffin-Lim vocoder [19].

2.1 Prosodic Symbols

Table 1 lists the prosodic features and symbols. We specified the prosodic symbols with reference to the Japanese ToBI label model [20] as follows. The prosodic symbols consist of initial rising (which denotes a rapid rising of F0 after the symbol), accent nucleus (which denotes a rapid falling of F0 after the symbol), accentual phrase boundary, EOS, and pause symbols. An accent-phrase is a unit that forms an accent during pronunciation. This description method uses arbitrary symbols, but has a simple representation. Because the symbols are arbitrary, this method can be applied to other languages besides Japanese.

2.2 Japanese Accent System and Our Notation Method

Figure 2 illustrates our notation method for representing high (H) and low (L) pitch between mora for seq2seq AM in Fig. 2. The Japanese language has a pitch-accen feature that can be represented as a sequence of binary F0 levels in mora units [21], [22]. Japanese pitch accents in Tokyo dialect have the following rules.

- A rapid rise or fall in F0 must take place between the first mora and the second mora.
- The maximum number of rapidly rising patterns of F0 between two consecutive morae in a word is one.
- The maximum number of rapidly falling patterns of F0

| Feature     | Prosodic symbols |
|-------------|------------------|
| Initial rising | ^                |
| Accent nucleus | !               |
| Accentual phrase boundary | #               |
| EOS (Declarative) | (               |
| EOS (Interrogative) | ?               |

Table 1 Examples of prosodic symbols.
between two consecutive morae in a word is one.

We have added these rules to the notation method for seq2seq AM and made up the label format for seq2seq AM. Figure 2 shows the procedure of producing phonetic and prosodic symbols. First, we conduct morphological analysis [23] for obtaining PP labels to represent phonetic symbols and accent prediction [24] for obtaining accent information. Second, we combine these sequences and the representations for the phonetic symbols and prosodic features. Finally, we follow the accent rules in Table 1; we place the prosodic symbols between the phonetic symbols and obtain phonetic and prosodic symbols. In the inputted PP labels, the initial rising “ˆ” and accent nucleus “!” prosodic symbols make the prosodic-acoustic feature appropriate high or low pitch until the next prosodic symbol appears in the input of seq2seq AM.

2.3 Text Analysis and FC-Label-to-PP-Label Conversion

Open JTalk [24] is a TTS system that contains a Japanese front-end, and it converts Japanese text into FC labels [15]. Open JTalk contains Japanese text analysis functions, including grapheme-to-phoneme conversion, and a morphological analysis called Mecab [23]. Figures 3 and 4, and Tables 2 and 3 explain the FC label converter to PP label. The FC label has linguistic and acoustic information; we pick up the phonemes and features listed in Table 1 and convert the sequences into PP symbols. On the other hand, Japanese kanji (Chinese) characters potentially contain mismatches between input symbols and acoustic features, which can potentially cause alignment errors in the encoder outputs and decoder inputs due to the kanji having multiple readings. These mismatches may in turn cause alignment errors depending on conversion errors in the text of the corpus.

2.4 Hand-Editing and Direct Description

The PP labels are simple and can be edited by hand. This enables direct annotation without an FC label converter. Compared with time-aligned FC labels, PP labels do

\[ PP_1 = p_{1,1} \land p_{2,3} \land p_{3,3} \land p_{4,4} \land p_{5,5} \land \ldots \land p_{n-1,3} \land p_{n,3} \]
not require the boundary positions of the phonemes, the identification of which is a time-consuming task because the boundaries between phonemes are ambiguous. In particular, with PP labels, we can read and annotate them directly. Meanwhile, it is not possible to read FC labels because of their complicated expression.

3. Experiments

We conducted objective evaluations on the encoder-decoder alignments, synthesized mel-spectrograms and F0 (Sect. 3.2). We also conducted subjective evaluations of the naturalness of the synthesized speech (Sect. 3.3).

3.1 Datasets and Experimental Conditions

We used the JSUT [13] corpus, which is a large-scale open Japanese speech corpus. The whole corpus contains 10 hours of speech and corresponding normal Japanese text. It contains 7,696 utterances. The corpus was split into 7,596 samples for training and 100 samples for testing. The test set included 30 samples for evaluations and four samples for the evaluators’ training. FC labels with time alignments were generated by Open JTalk and Julius [25], and PP labels were generated by Open JTalk and the FC label conversion method (see Sect. 2.3). The number of training iterations for the WaveNet vocoder was 860,000. The sampling rate was 22 kHz, 16-bit.

We conducted experiments comparing four types of input sequences consisting of commonly used Japanese characters and three types of seq2seq AM method and implementations (see Table 4). We automatically prepared the Japanese phonetic symbols and prosodic (PP) labels by using the method described in Sect. 2.3. The first type was normal Japanese text consisting of kanji and hiragana (KH). Hiragana has the same reading as katakana. The second type was automatically generated plain katakana (KT) which represented readings without accent information; the characters were represented by mora. The third type was phonetic symbols consisting of Roman alphabet (phoneme) and prosodic symbols (PP (phon.)). The fourth type was phonetic symbols consisting of katakana (mora) and prosodic symbols (PP (mora)). The PP labels and KT included misread kanji converted by Open JTalk, but they were used in the training without making any hand-edited corrections to them. In Sects. 3.2.2 and 3.3.1, we also used the PP (phon.) for comparison with PP (mora). These types of sequences composed the input of the proposed method. Table 4 lists the systems that were tested in the experiment, and Table 5 shows an example of input symbols.

Mel-spectrograms generated from seq2seq AM were converted into waveforms by using WaveNet or a Griffin-Lim vocoder. The mel-spectrogram features of all methods had 80 dimensions, a 125 – 7600-Hz frequency band, and a 46-ms window size.

3.2 Objective Evaluation

3.2.1 Visualization of the Encoder-Decoder Alignment

Figure 5 shows the alignments of T2PP (mora), T2KT, and T2KH. The iterations and batch size of Tacotron 2 were 600,000 and 48, respectively. Figure 5 (a) shows a slightly non-smooth alignment. In this case, the combination of a phonetic and a prosodic symbol expressed an acoustic information of a mora as decoder timesteps.

The figure indicates that the alignment monotonically increased and was continuous. Figure 5 (b) shows a smoother alignment; there was a one-on-one correspondence between the encoder which represented a mora character and the decoder timestep which represented the mora mel-spectrogram. Figure 5 (c) shows unclear and discontinuous alignments in some timesteps, suggesting that T2KH had difficulty training the model.
3.2.2 Effectiveness of Prosodic Symbols

Figures 6 and 7 compare the mel-spectrograms and F0 of T2PP (mora), T2KT, and T2KH. The iterations and batch size of Tacotron 2 were 600,000 and 48, respectively. The proposed method (T2PP (mora)) was better at reconstructing the details in the red rectangles, and they replicated pauses and the falling and rising of tone and pitch. The ground truth and T2PP (mora) had similar features both overall and in detail. In contrast, T2KT produced overall flat speech; it did not reproduce the pitch-accent and pauses in the red rectangles. The T2KH was not similar in shape to the ground truth.

Moreover, the results in Figs. 6 and 7 confirmed that the accent nucleus “!” and initial rising “ ˆ ” symbols are effective. In the T2PP (mora) results of Fig. 7, we can see a rapid falling of F0 corresponding to “ηΩפח” and a rapid rising of F0 corresponding to “Ξי”. In T2PP (mora) results of Fig. 6, we can see a reproduced silence corresponding to a pause “_”. Overall, the results shown in these figures indicate that prosodic symbols can control acoustic features such as accent and pause information.

Table 6 compares the F0 correlations of T2PP (phon.), T2PP (mora), T2KT, and T2KH. The iterations and batch size of Tacotron 2 were 600,000 and 48, respectively. We used the WaveNet implementation [28] as the vocoder; the resynthesized results of the WaveNet vocoder are indicated as ReWN. The results of this experiment showed that the F0s of the two T2PPs were more similar than the F0s of the other systems. As well, T2PP (phon.) and T2PP (mora) with prosodic symbols had a higher evaluation value than T2KT. This result suggested that prosodic symbols replicate F0 and accentual features.

3.2.3 Comparing Manual and Automatic Generated Labels

Table 7 lists the matching rates of the manual and automatically generated labels. We prepared automatically generated and hand-edited PP labels made from 5,000 hand-edited FC labels [29]. They were made from 5,000 utterances of the JSUT corpus. We compared these labels and used the python difflib SequenceMatcher function [30] to calculate the strings’ similarity. Table 7 shows that the similarity in the case of using automatically generated PP labels is lower than that of kana. This experiment was conducted because it is known that the accuracy of the labels affects the evaluation results [31].

Moreover, the matching rates were significantly different. The matching rate in the PP case was 1.68%, while that
of kana was 22.6%. These results suggest that it is difficult to generate strings that perfectly match the hand-edited labels. Considering the difference in matching rates, predicting prosodic features is considered more difficult than kana. The prosodic features of Japanese depend on the context and potentially contain multiple patterns, so it is difficult to estimate them only from sentences without any acoustic features of speech.

### 3.2.4 Counting Errors in Miss-Synthesized Speech

Table 8 lists the number of synthesized utterances that were not generated correctly, comparing manually and automatically generated labels with synthesized speech. Seq2seq AMs have been reported to have synthesis errors that include deletions and repetitions of words [17]. These errors often occur at the end of a sentence. Even though a different sound is produced compared with the input sentence and it is a natural mistake for a sentence, it cannot be judged in subjective evaluations or from alignment errors. In this experiment, we prepared automatically generated PP labels (AUTO) and hand-corrected PP labels (HAND) made from the JSUT corpus 4,900 corrected FC labels [27]. The experiment was conducted using the ESPnet-Tacotron 2 [32] implementation and using the Griffin-Lim vocoder. There were 200 epochs, and we selected 50 texts and compared the input sentences with the synthesized speech by listening to them. We manually counted the errors in the synthesized speech. Table 8 shows that inputting the hand-corrected PP labels did not cause mis-synthesized speech. This means that PP labels are suitable input for seq2seq AM. Kanji conversion seemed to be the cause of the errors in the automatically generated PP labels; that is, PP labels did not cause the errors.

### 3.3 Subjective Evaluation

Four subjective evaluations were conducted using the 100 test samples as the evaluation stimuli. The evaluators were 200 speakers of standard Japanese (Tokyo dialect). The evaluated speech stimuli did not use any of the training data for the model. One audio sample was evaluated 20 times. Mean opinion scores (MOS) on a scale of 1-to-5 (1: bad, 5: excellent) and 95% confidence intervals were obtained from all the evaluators.

#### 3.3.1 Effectiveness of Prosodic Symbols

The systems listed in Table 4 were compared in terms of the naturalness of the speech they produced. We estimated the mel-spectrograms (MELSPC) by using three seq2seq AM models and generated the audio by using the Griffin-Lim vocoder (60 iterations). ReGL in the figure means re-synthesized audio created using the Griffin-Lim vocoder. The number of iterations and batch size of Tacotron 2 and Deep Voice 3 were 600,000 and 48, respectively. The number of iterations and batch size of Transformer-based TTS were 75,900 and 12, respectively. The results are summarized in Fig. 8. All methods that inputted PP labels had evaluation values greater than those of KT and KH. It can be seen that the PP labels worked effectively on all of the seq2seq AM methods. These results suggest that prosodic symbols can be applied to various architectures with attention-based seq2seq AM.

In addition, we conducted an experiment with different types of phonetic symbols to evaluate the effectiveness of prosodic symbols in Fig. 9. The systems listed in Table 4 and T2PP contained two types of phonetic symbol: Roman alphabet (phoneme) as T2PP (phon.) and katakana (mora) as T2PP (mora). The number of iterations and batch size of Tacotron 2 were 600,000 and 48. We generated the audio by using the WaveNet vocoder. The number of iterations and batch size for the vocoder were 200,000 and 48. The results are summarized in Fig. 9. T2PP (phon.) and T2PP (mora) had evaluation values greater than T2KT and T2KH. It can be seen that the PP labels worked well for every type of phonetic symbol. In this experiment, the comparison between T2PP (phon.) and T2PP (mora) showed a significant difference in T2PP (mora). These results suggest that prosodic symbols can be applied to various phonetic symbols with seq2seq AM and that T2PP (mora) is more effective for naturalness than T2PP (phon.) in this method.
### Table 9 Conventional systems used in the experiments.

| System | AM method     | Acoustic feature | Waveform synthesis |
|--------|---------------|------------------|--------------------|
| ReWD   | Re-synthesis  | Vocoder          | WORLD [33]         |
| ReWN   | Re-synthesis  | MELSPE           | WaveNet [28]       |
| MWD    | Merlin [34]   | Vocoder          | WORLD [33]         |
| MWN    | Merlin [34]   | Vocoder          | WaveNet [35]       |
| T2PP   | Tacotron 2 [26] | MELSPE           | WaveNet [28]       |

#### 3.3.2 Comparison with Conventional SPSS

We built the TTS systems listed in Table 9. All of them used automatically generated labels as training data. The conventional SPSS [1], [34] employed Open JTalk and Julius as a front-end. There were 281 errors due to forced alignment by Julius. Consequently, the training set had 7,324 sentences instead of 7,596, the test set had 91 instead of 100, and the evaluation object had 30. We used WORLD [33] as the vocoder, 60-dimensional mel-cepstral coefficients (MCCs), 2-dimensional band periodicities (BAPS), log F0 at 5 msec frame intervals (the acoustic features of Vocoder), and three recurrent hidden layers; each hidden layer had 512 LSTM (long short-term memory) units as duration and acoustic models.

The results are summarized in Fig. 10. Our method scored significantly higher than MWN and MWD, which are conventional SPSS methods. This result confirms the effectiveness of the proposed method without time alignment information. The reason for ReWN's score being lower than that of ReWD is that synthesized speech rarely contains unnatural sounds. A similar issue has been reported in experiments with neural vocoders [16].

#### 3.3.3 Effect of Changing the Volume of Data

We subjectively evaluated the effect of changing the volume of the training data of T2PP (mora). We generated the audio by using the WaveNet vocoder. The number of iterations and the batch size in this experiment were 200,000 and 48. Figure 11 shows that the evaluation values gradually increased with the number of hours of training. The systems trained on sets with less data tended to yield sentences that were perceived as mis-phrased and poorly accented. As the learning data was increased, the reproducibility of the phonetic and prosodic features increased, as did their naturalness.

#### 3.3.4 Comparison of Using Automatically Generated Labels and Hand-Corrected Labels

We subjectively evaluated the effectiveness of automatically generated PP labels in Figs. 12 and 13. For this experiment, we prepared automatically generated (AUTO) and hand corrected (HAND) labels made from 5,000 hand-corrected FC labels [29]. We generated the audio from the predicted mel-spectrograms by using the Griffin-Lim vocoder. It is logical to think that this experiment should have no effect on the evaluation result for any type of vocoder, because it does not affect high-low accent information. For this reason, we used the Griffin-Lim vocoder. The corpus contains 5,000 utterances of the JSUT corpus; the training set includes 4,900 utterances, the test set 100. We randomly selected 30 sentences. The evaluators were 60 speakers of standard Japanese (Tokyo dialect). The evaluated speech stimuli did not use any of the training data for the model. One audio sample was evaluated 30 times. There were 200 epochs, and we conducted a 1-to-5 evaluation (MOS score) and a pairwise comparison of the naturalness of the synthesized speech.

Figures 12 and 13 show that hand corrected PP labels was rated higher than automatically generated ones. These results suggest that the correctness of the labels affects the results of the evaluation. As described in Sect. 3.2.3, automatic labels contain errors due to miss-conversions, while prosodic features, which contain accent information, are difficult to estimate; these issues caused the low evaluations.
Mistakes in the prosodic features might have affected the reproducibility of the accents, etc., and reduced their naturalness.

4. Discussion

4.1 Experimental Findings

We confirmed the effectiveness of automatic generated labels for pitch-accent language in Japanese. The experiments showed that symbols can be used to control the accentual acoustic feature. Figures 6 and 7 confirm the controllability of the prosodic feature by comparing mel-spectrograms and F0. Until now, there has been no generic accent control method for Japanese seq2seq AM. The proposed method potentially solves the accentual control problems. Moreover, the prosodic features work with multiple seq2seq methods; this is in contrast to conventional SPSS, which requires extensive manual labor for correcting pronunciations by using labels describing the boundaries of phonemes. Although, the automatically generated labels were evaluated as worse than the hand-corrected ones, this could be attributed to miss-converted labels. Because the overall similarity is high, the phoneme alignments were correct, but the reproduced accents contained errors.

In Sect. 3.2.4, it was shown that PP labels are suitable for the input of seq2seq AMs. Seq2seq AMs rarely cause miss-synthesizes that are different from the sentence input at the end of the word. These sounds are often linguistically correct, but they do not express the same sound as the input text [17]. We experimentally confirmed how often errors occurred through the contribution of hand-edited labels [29] and found that no such errors occurred in the case of PP labels in the experiment described in Sect. 3.2.4. Prosodic features affect tone and pitch, and in Sect. 3.2.2, we showed it is possible for these acoustic features to be accurately represented by putting them between phonemes. For this reason, it may be possible that problems that cannot be learned with phonemes only, such as accurate accents, can be learned accurately by using PP labels.

4.2 How to Improve the Automatically Generated Labels

The production cost of controllable TTS is high in the conventional SPSS method using manual labeling. As described in Sect. 3.3.2, we conducted subjective evaluation experiments confirming that seq2seq AM with our method produces more natural speech than SPSS does. However, while this would solve the cost problem, the quality of the automatically generated labels remained lower than that of the hand-corrected labels. If this problem in pitch-accent languages can be solved, high-quality TTS can be realized without the need to use hand-edited labels.

The results of the subjective evaluation suggested that the prosodic symbols are effective. In Sect. 3.2.3, despite that automatically generated labels contained miss-converted symbols, the PP labels estimated tones of accent, etc. The experimental results showed that the prosodic symbols had other functions besides controlling phonemes. They also can be used to control the prosodic features, which do not correspond to the phonemes. Moreover, as shown in Sect. 3.3.1, the prosodic features could be adapted to seq2seq AMs that have not been proposed yet, and this might mean that we do not need to develop any new seq2seq AM models to input accent information. Accordingly, we can avoid developing new structures for accent control to support new seq2seq AMs.

In Sect. 3.2.3, it was shown that the prosodic features can control the accent, but we could not estimate the labels correctly. It was found that the Tacotron 2 used in this experiment could produce relatively correct speech even though there were some miss-converted labels. Moreover, it was found that an increase in accent estimation accuracy improved the evaluation results. Increasing the accuracy of accent estimation will lead to higher quality in the future.

The prosodic features potentially contain multiple expressions, so finding methods of linguistic analysis for accurately determining phonemes and acoustic analysis for accurately estimating accent remains an issue. In this study, we used linguistic analysis only for automatically generating labels; in the future, we would like to confirm the effectiveness of adding a prosodic symbol estimation method using acoustic analysis. In summary, the results of this study suggest that the prosodic features are effective for Japanese seq2seq AM. Moreover, because the input of seq2seq AM is merely symbols, the prosodic features may work in languages other than Japanese.

4.3 Points to be Confirmed in Future Experiments

The evaluation of ReWN described in Sect. 3.3.2 gave an unclear result. In this regard, we used mel-spectrograms extracted from raw audio to train the ReWN model; nevertheless, the result was not good. The reproduced consonants in ReWN may not have been good. In particular, confusion in the perception of the reproduced consonants may have influenced the results of the evaluation. However, ReWN did not seem to affect the results of the proposed method, because ReWN seemed to have no effect on the experimental results of T2PP (mora), MWD or MWN.

5. Conclusion

We proposed a method to control prosodic features by inserting symbols representing these features between phonetic symbols in various architectures with attention-based seq2seq AM. The addition of prosodic symbols resulted in more accurate replication of accents, pauses, and sentence ending acoustic expressions and improved the evaluation value compared with methods inputting only plain phoneme sequences. Moreover, we found that the proposed method in combination with a front-end could automatically generate speech without imposing a large annotation workload. But
an evaluation of the automatically generated labels and hand-edited labels showed that the hand-edited labels were still better, so the automatic estimation method for labels should be improved. The naturalness of the speech indicated in the evaluations was higher than that of conventional SPSS. Our method has the potential for application to various languages and to work in as yet undeveloped seq2seq acoustic modeling methods.

References

[1] H. Zen and A. Senior, “Deep mixture density networks for acoustic modeling in statistical parametric speech synthesis,” Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp.3844–3848, 2014.
[2] H. Zen, K. Tokuda, and A.W. Black, “Statistical parametric speech synthesis,” Speech Communication, vol.51, no.11, pp.1039–1064, 2009.
[3] R.J. Skerry-Ryan, E. Battenberg, Y. Xiao, Y. Wang, D. Stanton, J. Shor, R.J. Weiss, R. Clark, and R.A. Saurous, “Towards end-to-end prosody transfer for expressive speech synthesis with Tacotron,” Proc. International Conference on Machine Learning (ICML), pp.7471–7480, 2018.
[4] Y. Wang, R.J. 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,” Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH), pp.4006–4010, 2017.
[5] J. Sotelo, S. Mehri, J.F. Santos, K. Kastner, A. Courville, and Y. Bengio, “Char2Wave: End-to-end speech synthesis,” Proc. International Conference on Learning Representations (ICLR), 2017.
[6] Y. Taigman, L. Wolf, A. Polyak, and E. Nachmani, “Voiceloop: Voice fitting and synthesis via a phonological loop,” Proc. International Conference on Learning Representations (ICLR), no.31, 2018.
[7] W. Ping, K. Peng, A. Gibiansky, S.O. Arik, A. Kannan, S. Narang, J. Raiman, and J. Miller, “Deep Voice 3: Scaling text-to-speech with convolutional sequence learning,” Proc. International Conference on Learning Representations (ICLR), no.41, 2018.
[8] N. Li, S. Liu, Y. Liu, S. Zhao, M. Liu, and M. Zhou, “Neural speech synthesis with Transformer network,” arXiv preprint arXiv:1809.08895, 2018.
[9] K. Kurihara, N. Seiyama, T. Kumanou, and A. Imai, “Study of Japanese end-to-end speech synthesis method that inputting kana and prosodic symbols (in Japanese),” Proc. Autumn Meeting of Acoustical Society of Japan, pp.1083–1084, 2018.
[10] Y. Yasuda, X. Wang, S. Takaki, and J. Yamagishi, “Investigation of enhanced tacotron text-to-speech synthesis systems with self-attention for pitch accent language,” Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp.6905–6909, 2019.
[11] 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. 10th ISCA Speech Synthesis Workshop (SSW), pp.166–171, 2019.
[12] Y. Yasuda, X. Wang, and J. Yamagishi, “Initial investigation of encoder-decoder end-to-end TTS using marginalization of monotonic hard alignments,” Proc. 10th ISCA Speech Synthesis Workshop (SSW), pp.211–216, 2019.
[13] R. Sonobe, S. Takamichi, and H. Saruwatari, “JSUT corpus: Free large-scale Japanese speech corpus for end-to-end speech synthesis,” arXiv preprint arXiv:1711.00354, 2017.
[14] T. Okamoto, T. Toda, Y. Shiga, and H. Kawai, “Real-time neural text-to-speech with sequence-to-sequence acoustic model and WaveGlow or single Gaussian WaveRNN vocoders,” Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH), pp.1308–1312, 2019.
[15] “An example of context-dependent label format for hmm-based speech synthesis in Japanese,” http://hts.sp.nitech.ac.jp/archives/2.3/HTS-demo_NIT-ATR503-M001.tar.bz2.
[16] S. Shechtman and A. Sorin, “Sequence to sequence neural speech synthesis with prosody modification capabilities,” Proc. 10th ISCA Speech Synthesis Workshop (SSW10), pp.275–280, 2019.
[17] T. Hayashi, R. Yamamoto, K. Inoue, T. Yoshimura, S. Watanabe, T. Toda, K. Takeda, Y. Zhang, and X. Tan, “Esnet-TTS: Unified, re-producible, and integratable open source end-to-end text-to-speech toolkit,” Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp.7654–7658, 2020.
[18] A.v.d. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, “WaveNet: A generative model for raw audio,” arXiv preprint arXiv:1609.03499, 2016.
[19] D. Griffin and L. Jae, “Signal estimation from modified short-time fourier transform,” IEEE Trans. Acoust., vol.32, no.2, pp.236–243, 1984.
[20] J.J. Venditti, “Japanese ToBI labelling guidelines,” 1997.
[21] H. Fujisaki and K. Hirose, “Analysis of voice fundamental frequency contours for declarative sentences of Japanese,” Journal of the Acoustical Society of Japan (E), vol.5, no.4, pp.233–242, 1984.
[22] N. Minematsu, S. Kobayashi, S. Shimizu, and K. Hirose, “Improved prediction of Japanese word accent sandhi using crf,” Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH), pp.2561–2564, 2012.
[23] T. Kudo, K. Yamamoto, and Y. Matsumoto, “Applying conditional random fields to Japanese morphological analysis,” Proc. Conference on Empirical Methods in Natural Language Processing (EMNLP), pp.230–237, 2004.
[24] “Open JTalk,” http://open-jtalk.sourceforge.net/.
[25] A. Lee, T. Kawahara, and K. Shikano, “Julius - an open source real-time large vocabulary recognition engine,” https://github.com/julius-speech/synthesis-kit.
[26] R. Mamah, “Deepmind’s Tacotron 2 tensorflow implementation,” https://github.com/Rayhane-mamah/Tacotron-2.
[27] R. Yamamoto, “Deepvoice3.pytorch,” https://github.com/ry9y9/deepvoice3.pytorch.
[28] R. Yamamoto, “WaveNet vocoder,” https://github.com/ry9y9/wavenet-vocoder.
[29] T. Koriyama, “JSUT-label,” https://github.com/OpenJTalk/segmentation-kit.
[30] PSF, “Diffihb — helpers for computing deltas,” The Python Software Foundation, https://docs.python.org/3/library/diffihb.html.
[31] H.-T. Luong, X. Wang, J. Yamagishi, and N. Nishizawa, “Investigating accuracy of pitch-accent annotations in neural network-based speech synthesis and denoising effects,” Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH), pp.37–41, 2018.
[32] S. Watanabe, T. Hori, S. Karita, T. Hayashi, J. Nishitoba, Y. Unno, N.E.Y. Soplin, J. Heymann, M. Wiesner, and N. Chen, “Esnet: End-to-end speech processing toolkit,” Proc. Interspeech, pp.2207–2211, 2018, https://github.com/esnet/esnet.
[33] M. Morise, F. Yokomori, and K. Ozawa, “World: A vocoder-based high-quality speech synthesis system for real-time applications,” IEICE Trans. Inf. & Syst., vol.E99-D, no.7, pp.1877–1884, 2016.
[34] Z. Wu, O. Watts, and S. King, “Merlin: An open source neural prosody modification capabilities,” Proc. 10th ISCA Speech Synthesis Workshop (SSW10), pp.275–280, 2019.
[35] T. Oyama, T. Hayashi, K. Kobayashi, K. Takeda, and T. Toda, “Speaker-dependent WaveNet vocoder,” Proc. Interspeech, pp.1118–1122, 2017, https://github.com/openjtalk/openjtalk.
[36] IEICE TRANS. INF. & SYST., VOL.E104–D, NO.2 FEBRUARY 2021
Kiyoshi Kurihara received his B.E. and M.E. from the Department of Electronics and Communications, Meiji University, Japan, in 2004 and 2006 respectively. After his graduation, he joined NHK (Japan Broadcasting Corporation) and worked as a master control & radio studio engineer at a broadcasting station. He joined the Science & Technology Research Laboratories in 2016. His current research interests include speech synthesis and automatic generation of audio descriptions in broadcasts.

Nobumasa Seiyama received his B.E. and M.E. degrees in Electrical Engineering from Waseda University, Japan, in 1987 and 1989, respectively. Originally joining NHK in 1989, since then, he has worked with the Science and Technology Research Laboratories, engaging in speech signal processing, voice conversion, speech synthesis and development of a high-quality speech-rate conversion system for the elderly. He belonged to NHK Engineering Services, INC. from 2007 to 2010.

Tadashi Kumano is a principal research engineer in Science and Technology Research Laboratories (STRL) of NHK (Japan Broadcasting Corporation). He joined NHK STRL after getting his master’s degree in Computer Science from the Graduate School of Information Science and Engineering of the Tokyo Institute of Technology in 1995. He was also a member of ATR (Advanced Telecommunication Research Institute International) Spoken Language Translation Laboratories from 2000 to 2004. He has been engaged in research on natural language processing and machine translation, and recently has been interested in spoken language generation such as computer-producing voice commentaries for sports programs.