10th ISCA Speech Synthesis Workshop
20-22 September 2019, Vienna, Austria

Impacts of Input Linguistic Feature Representation on Japanese End-to-End Speech Synthesis

Takato Fujimoto, Kei Hashimoto, Keiichiro Oura, Yoshihiko Nankaku, and Keiichi Tokuda

Department of Computer Science, Nagoya Institute of Technology, Nagoya, Japan

{taka19, bonanza, uratec, nankaku, tokuda}@sp.nitech.ac.jp

Abstract

We investigate the impact of input linguistic feature representation on Japanese end-to-end speech synthesis. An end-to-end speech synthesis system, which directly generates natural speech from text, has recently been proposed. The English end-to-end system Tacotron 2 achieves sound quality close to that of natural speech. However, unlike alphabetic language that use stress accent, such as English and Spanish, it is difficult to achieve end-to-end speech synthesis with other non-alphabetic languages (e.g., Japanese and Chinese, which use pitch accent and tone, respectively, and use ideograms). We investigated the units of an input sequence, contexts, pause insertion, vowel devoicing, and pronunciation of particles for Japanese end-to-end speech synthesis. Experimental results indicate improvement in the naturalness of the synthesized speech using high or low accents. The results also indicate that the accent-phrase information can help to predict pause insertion, and an end-to-end text-to-speech model may be able to change the pronunciation for devoiced vowels and particles.

Index Terms: end-to-end model, Japanese speech synthesis, linguistic feature representation, units of an input sequence

1. Introduction

The performance of text-to-speech (TTS) synthesis systems has improved due to the development of deep neural networks (DNNs) [1]. Sequence-to-sequence models with attention mechanisms [2] have recently been widely adopted as DNN-based architectures. These sequence-to-sequence models have shown considerable success not only in TTS but also in various tasks such as machine translation [2, 3] and speech recognition [4]. A sequence-to-sequence model consists of an encoder and decoder. The encoder maps the input sequence to a fixed-sized vector, and the decoder maps the vector to the target sequence. Attention mechanisms allow the decoder to focus on different parts of the input sequence at different decoding steps.

Several end-to-end TTS systems based on sequence-to-sequence models have been proposed, including Char2Wav [5], VoiceLoop [6], Deep Voice 3 [7], ClariNet [8], Transformer-based TTS [9], Tacotron [10], and Tacotron 2 [11]. It has been reported that the quality of the synthesized speech of some end-to-end TTS systems is close to that of natural speech [9, 11]. However, these results are reported only for English, and end-to-end TTS systems for other non-alphabetic languages, such as Japanese and Chinese, have not been successful as they have issues. End-to-end speech synthesis for such languages has been investigated [12, 13, 14] but not insufficiently. This work focuses on Japanese speech synthesis and investigates similar and different approaches.

The English alphabet has 26 letters, each having an uppercase and a lowercase form. In addition, words are separated by a space character. However, it is difficult to input a character sequence in languages with many characters such as Japanese and Chinese. To address this problem, a phoneme sequence is used in a Japanese end-to-end speech synthesis [12]. The sequence of accent types, which represent the position of the accent nucleus, is also used because Japanese accents are essential to the prosody of Japanese utterances and difficult to predict from phoneme sequences. In [12], expanded Tacotron with self-attention has been proposed, and the effectiveness of self-attention to capture long-term dependencies has indicated. However, it does not outperform traditional pipeline systems for Japanese speech synthesis. [13] adds prosodic symbols representing prosodic information, such as accent nuclei, accent-phrase boundaries, and pauses, to a mora sequence, as input to an end-to-end TTS model. These Japanese TTS systems require text analyzers because of pitch accent and kanji, of which there is a large number, and difficult to predict pronunciation. Morphological analysis, estimation of reading and accent, pause insertion, and change in pronunciation such as voiceless vowels and long sounds are conducted in Japanese text analysis for speech synthesis. We investigate the impact on an end-to-end TTS model by comparing phoneme-level and mora-level input language feature representations. We also investigate whether part of the text processing can be replaced with an end-to-end TTS model. The processing we targeted in this study is the insertion of pauses and the change in pronunciation for devoicing and particles.

In [14], a Chinese end-to-end TTS system for modeling additional linguistic features has been proposed and contexts are used to improve the prosody of the synthesized speech. Although contexts including both prosodic word information and sentence-level information are effective, traditional full context labels cannot function. We investigate the impact of full context labels in more detail by using phoneme-level and mora-level full context labels.

We investigate whether the suitable units of the input sequence for Japanese end-to-end speech synthesis are phoneme or mora since such units are different in Japanese end-to-end TTS systems. We also verify the effectiveness of the accent information using high pitch/low pitch (H/L) accents and investigate the impact of input linguistic feature representation for pause insertion and change in pronunciation.

The rest of this paper is organized as follows. In sections 2 and 3, we describe Tacotron 2, which is an end-to-end speech synthesis system, and linguistic features, respectively. We present the experimental conditions and results in Section 4 and provide concluding remarks and future work in Section 5.

2. Tacotron 2

Tacotron 2 [11] is an end-to-end speech synthesis system that generates human-like speech. This system consists of two components: a sequence-to-sequence model for predicting a se-

10.21437/SSW.2019-30

quence of mel spectrograms from an input sequence and a vocoder based on WaveNet [15] for generating waveforms samples from predicted mel spectrograms.

2.1. Sequence-to-sequence model

The sequence-to-sequence model of Tacotron 2 is composed of an encoder and decoder with attention. The encoder converts a character sequence into a hidden feature representation for each encoder step. The hidden feature representation sequence is summarized in a context vector as the input to the decoder by the attention mechanism. The context vector represented as a weighted sum of the hidden feature representations is used to help predict the current output. The attention mechanism used in Tacotron 2 is the location-sensitive attention [16], which uses cumulative attention weights from previous decoder time steps. The decoder, which is an autoregressive recurrent neural network predicts the current mel spectrogram from the context vector and previous mel spectrogram.

2.2. WaveNet vocoder

A WaveNet vocoder is a vocoder based on neural networks that generate audio waveforms from acoustic features. The input of WaveNet is a sequence of predicted waveform samples in the past and auxiliary features. The joint probability of a sequence of waveform samples $\boldsymbol{x} = (x_1, ..., x_T)$ can be written as

$$P(\boldsymbol{x}|\boldsymbol{h}) = \prod_{t=1}^{T} P(x_t|x_1, ..., x_{t-1}, \boldsymbol{h}), \quad (1)$$

where $\boldsymbol{h}$ represents the auxiliary features. The auxiliary features are used to predict the waveform sample at gated activation units. The gated activation function is defined as follows:

$$\boldsymbol{z} = \tanh(\boldsymbol{W}_f * \boldsymbol{x} + \boldsymbol{V}_f * \boldsymbol{y}) \odot \sigma(\boldsymbol{W}_g * \boldsymbol{x} + \boldsymbol{V}_g * \boldsymbol{y}), \quad (2)$$

where $\boldsymbol{x}$ and $\boldsymbol{z}$ are the input and output of the activation units, $*$ is a convolution operator, $\odot$ is an element-wise product operator, $\sigma(\cdot)$ represents a sigmoid function, $\boldsymbol{W}$ and $\boldsymbol{V}$ represent learnable convolution filters for the input and auxiliary features, and f and g represent a filter and gate, respectively. The variable $\boldsymbol{y}$ is a time series of the original auxiliary features $\boldsymbol{h}$ transformed into the same resolution as $\boldsymbol{x}$.

3. Linguistic features in Japanese language

In this section, we describe the units of an input sequence, effectiveness of H/L accents, pause insertion, vowel devoicing, and Japanese particles.

3.1. Units of input sequence

In end-to-end TTS models, a character or phoneme sequence is widely used as the input sequence. English, which is an alphabetic language, has 26 letters, whereas Japanese has many graphemes, e.g., kanji, hiragana, and katakana. Unlike Chinese characters, Japanese kanji have two different readings, i.e., Chinese and Japanese. Consequently, it is difficult to represent all characters and predict the kanji reading. In [12], the phoneme sequence is input to the Japanese end-to-end TTS system. However, Japanese graphemes can be converted into hiragana or katakana, and a notation-level sentence can be converted into a pronunciation-level sentence, as shown in Figure 1. Japanese uses morae rather than syllables, and the Japanese writing system of kana (i.e., hiragana and katakana) is based on morae,

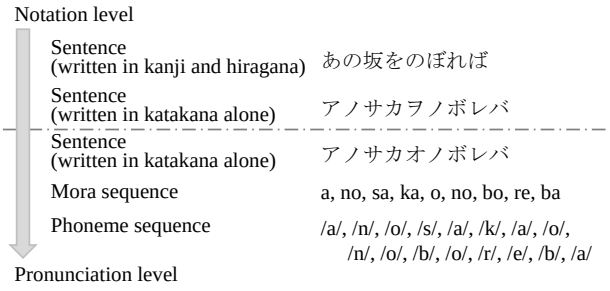

Figure 1: *Example of phoneme and mora sequence for "あの坂をのぼれば (If you go up the hill)".*

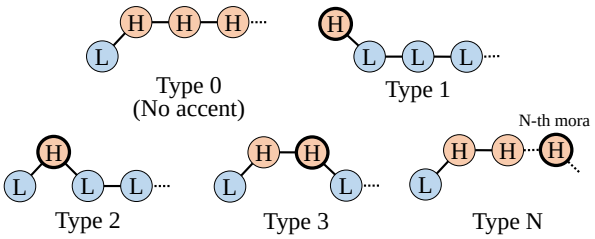

Figure 2: *High pitch/low pitch (H/L) accents and accent types. Thick circle represents accent nucleus.*

which are rhythmic units. Therefore, a Japanese end-to-end system with mora as the unit of the input sequence has also been proposed [13]. Since the units of these input sequence have not been compared, we compare phoneme and mora as suitable units of the input sequence for Japanese end-to-end speech synthesis.

3.2. Japanese accents

Japanese is a pitch-accent language, and the accents are essential to the prosody of Japanese utterances. Japanese accents are represented by accent type, which is the position of the accent nucleus in the accent phrase, and are represented by the accent, which is either high or low in the pitch of each mora, as shown in Figure 2. Every mora can either be pronounced with a high or low pitch. Also, there is a drastic fall in pitch after the accent nucleus in the accent phrase. Japanese pitch accents have the following features [17].

- There is a drastic rise or fall in pitch between the first mora and second one.
- The pitch of the first mora that is not an accent nucleus is low.
- The pitch of the mora after the accent nucleus is low.

3.3. Pauses, devoiced vowels, and particles

Pauses play important roles both for the intelligibility and naturalness of speech [18, 19]. An unnatural pause location is one cause of degradation in the naturalness of synthesized speech. Therefore, it is necessary to insert pauses at appropriate locations. Because words are separated by space characters in English, it is unlikely that pauses will be inserted in the words. Since word boundaries are unknown in Japanese and Chinese, pauses may be inserted into words. Therefore, pauses are inserted in advance in Japanese end-to-end speech synthesis.

Because Japanese has exceptional kana and phoneme associations, phonological changes are applied. We focus on vowel devoicing and pronunciation of particles. There is a common process of vowel devoicing in Japanese. Close vowels, i.e., /i/ and /u/, tend to become voiceless when placed between two voiceless consonants or between a voiceless consonant and silence, and other vowels become occasionally voiceless. Natural speech requires vowel devoicing, and an unnatural devoiced vowel decreases the naturalness of synthesized speech [20]. Also, pronunciation changes when は (“ha”) and へ (“he”) are used as particles; “ha” and “he” are usually pronounced as /ha/ and /he/, but pronounced as /wa/ and /e/ when used as particles, respectively.

4. Experiments

4.1. Experimental conditions

We conducted four subjective evaluation experiments to investigate the units of the input sequence and contexts, accent information, pause insertion, and the change in pronunciation for devoicing and particles for Japanese end-to-end speech synthesis. The Ximera [21] datasets as a Japanese speech database were used in the experiments. The database includes 12203 utterances uttered by a female speaker; 11592 utterances were used for a training set and the remaining 428 and 123 utterances were used for validation and a test set, respectively. These speech signals were sampled at 16kHz, and there was silence before and after the utterance. Natural speeches were analyzed using a short time Fourier transform with 12.5-ms frame shift and 50-ms frame length then transformed to the mel scale using 80-channel mel filter banks.

We used Tacotron 2 as the end-to-end speech synthesis system. The number of units or channels for each layer of this system was the same as [11], except embedding layer was not used. There are 30 residual blocks in the WaveNet vocoder. Specifically, dilation in 10 layers was set to $2^0, 2^1, 2^2, ..., 2^9$ and repeated three times to form a total of 30 dilated causal convolution layers. The number of channels for dilated causal convolutions, residual connection, and skip-connection was set to 256. The WaveNet vocoder output a 256-way categorical distribution to predict waveform sample values encoded using the μ-law algorithm [22]. We trained the WaveNet vocoder by using mel spectrograms extracted from training data as auxiliary features.

To evaluate the naturalness of synthesized speech, a subjective listening test was conducted. The naturalness of the synthesized speech was evaluated using the mean opinion score (MOS) test based on a five-point scale (5: natural – 1: poor). Twenty utterances were chosen at random from the test set and the subjects were fifteen Japanese.

4.2. Input sequence and contexts

We compared six inputs to investigate input units to the encoder and contexts on the end-to-end TTS model.

- **Phoneme_Onehot**: The input to the encoder was a phoneme sequence, where each phoneme was represented as a 41-dimensional one-hot vector. The phoneme label had 41 classes including silence and a pause, and a long pause and a short pause could not be distinguished.
- **Phoneme_Onehot+HL**: The input to the encoder was a sequence of the phoneme one-hot vectors of **Phoneme_Onehot** concatenated with two-dimensional vectors, which represent a high or low pitch. In silence or a pause, the accent vector was a zero vector.
- **Phoneme_Context**: The input to the encoder was a sequence of 528-dimensional linguistic features. These features were designed based on the context label adopted by Japanese HTS [23], except the word-level context was not included. These features included the phoneme-, mora-, accent-phrase-, breath-group-, and utterance-level contexts as shown in Table 1.
- **Mora_Onehot**: The input to the encoder was a mora sequence, where each mora was represented as a 131-dimensional one-hot vector. The mora label included silence and a pause, as in **Phoneme_Onehot**.
- **Mora_Onehot+HL**: The input to the encoder was a sequence of the mora one-hot vectors of **Mora_Onehot** concatenated with two-dimensional vectors, which represent a high or low pitch.
- **Mora_Context**: The input to the encoder was a sequence of mora-level 983-dimensional linguistic features. The mora-level linguistic features included the mora-, accent-phrase-, breath-group-, and utterance-level contexts in Table 1. Regarding the difference between **Phoneme_Context** and **Mora_Context**, the mora-level linguistic features did not include the phoneme-level context but included the phoneme identities of the consonant and vowel constituting the current and the two preceding and succeeding morae.

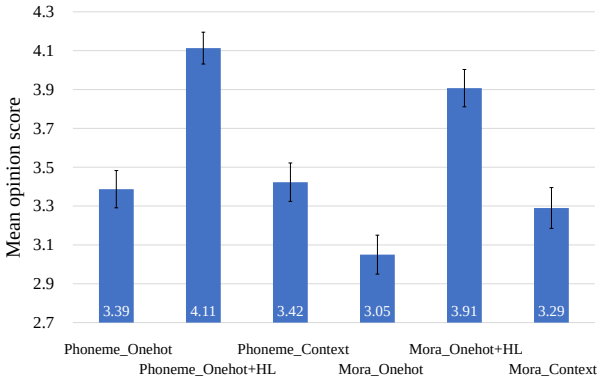

Figure 3: *Mean opinion score (MOS) with 95% confidence intervals for input sequence and contexts.*

Figure 3 shows the experimental results from the MOS test. **Phoneme_Onehot+HL** and **Mora_Onehot+HL** significantly outperformed the other inputs. These results indicate that inputting the accent information due to H/L accents is effective. However, **Phoneme_Context** and **Mora_Context** received lower scores than **Phoneme_Onehot+HL** and **Mora_Onehot+HL**, even though many contexts were input including accent information. It seems that useful contexts could not be used properly because they included redundant information, as in [14]. However, because there was not much difference between **Phoneme_Context** and **Phoneme_Onehot** and there was a significant difference between **Mora_Context** and **Mora_Onehot**, **Mora_Context** received a higher score than **Mora_Onehot** due to phoneme identities of the consonants and vowels. This was also supported by the results between **Phoneme_Onehot** and **Mora_Onehot** and between

Table 1: *Contexts included in linguistic features.*

Phoneme	Current and two each preceding and succeeding phoneme identities.
Mora	Difference between the accent nucleus and the current mora. Position of the current mora in the accent phrase.
Accent phrase	Number of morae in the previous, current, or next accent phrase. Accent type of the previous, current, or next accent phrase. Whether or not there is a pause in the accent phrase boundaries. Whether or not the previous, current, or next accent phrase is interrogative. Position of the current accent phrase in the breath group.
Breath group	Number of morae or accent phrases in the previous, current, or next breath group. The position of the current breath group in the utterance.
Utterance	Number of morae, accent phrases, or breath groups in the current utterance.

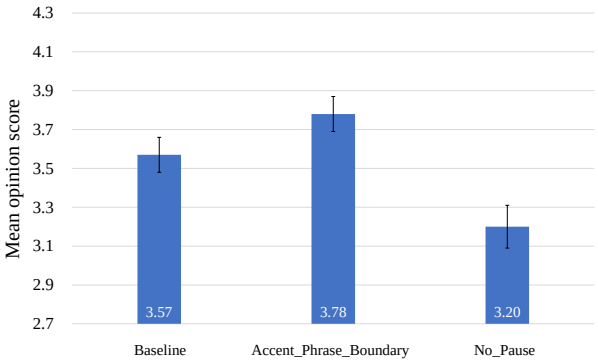

Figure 4: *MOS with 95% confidence intervals for pause insertion.*

Phoneme_Onehot+HL and **Mora_Onehot+HL** because **Phoneme_Onehot** and **Phoneme_Onehot+HL** with phoneme information received higher scores than **Mora_Onehot** and **Mora_Onehot+HL** without phoneme information. Therefore, in the end-to-end TTS model, since phonological information is preferentially extracted over prosodic information, it seems that prosodic information needs to be input with compact features such as distributed representations.

4.3. Pause insertion

We conducted the experiment on pause insertion. The pause symbol was represented as one of 41 classes.

- **Baseline**: **Phoneme_Onehot** of the first experiment. The pause symbols were inserted in the observed pause locations.
- **Accent_Phrase_Boundary**: The input with a pause symbol at every accent-phrase boundary, where the pause location is unknown. The pause symbols were inserted in the hypothesized pause locations. The input vector was 41 dimensions, the same as **Baseline**.
- **No_Pause**: The input without pause symbols, where the pause location is unknown. The pause symbols were not inserted. The input vector was 40 dimensions and a pause symbol was removed from the one-hot vector of **Baseline**.

Table 4 shows the experimental results from the MOS test. **Accent_Phrase_Boundary** significantly outperformed the other inputs. This indicates that even if the pause location is unknown, more natural synthesized speech can be generated than in **Baseline** in which the pause location is known. This is because a more natural accent was estimated based on the accent-phrase boundary, although the pauses were occasionally inserted in an unnatural location. However, **No_Pause** received the lowest score. These results indicate that the natural pause location cannot be estimated from the phoneme sequence alone.

Table 2: *Comparison of MOS on whether there were devoicing vowels in input sequence.*

Change in pronunciation for vowel devoicing	MOS	p-value
Yes (**Baseline**)	3.37 ± 0.10	0.097
No (**No_Devoiced_Vowel**)	$\mathbf{3.48} \pm 0.09$	

4.4. Vowel devoicing

The phoneme sequences with and without voiceless vowels were compared in this experiment.

- **Baseline**: **Phoneme_Onehot** of the first experiment. The input phoneme sequence included the devoiced vowels.
- **No_Devoiced_Vowel**: The input without the devoiced vowels. The input vector was the 38-dimensional vector, which did not include the devoiced vowels. The devoiced vowels were input as the corresponding voiced vowels.

Table 2 shows the experimental results from the MOS test. There was not much difference between **Baseline** and **No_Devoiced_Vowel**, indicating that the naturalness of the synthesized speech of **No_Devoiced_Vowel** was equivalent to that of **Baseline**. Although the end-to-end TTS model cannot always predict vowel devoicing, it is possible to achieve close naturalness to that of **Baseline**.

4.5. Particles ha/he

We evaluated whether the end-to-end TTS model could change the pronunciation for “ha” and “he” by conducting a MOS test for naturalness.

- **Baseline**: **Phoneme_Onehot** of the first experiment. “ha” and “he” were input as /w/ and /a/, and /e/, respectively.

Table 3: *MOS for change in pronunciation for particles ha/he.*

Change in pronunciation for particles	MOS	p-value
Yes (**Baseline**)	3.66 ± 0.09	0.398
No (**No_Particle**)	$\mathbf{3.68} \pm 0.09$	

- **No_Particle**: The input that does not change the pronunciation of the particles. The particles were input as they were.

Table 3 shows the experimental results from the MOS test. The MOS of **No_Particle** was equivalent to that of **Baseline**. This indicates there was no difference between the quality of **Baseline** and that of **No_Particle**. Figure 5 shows the mel spectrograms predicted from **Baseline** and **No_Particle** to compare the pronunciation of the particles. The top and bottom mel spectrograms of Figure 5 were predicted from different inputs but were similar. However, the middle and bottom mel spectrograms were predicted from the same inputs but were not similar. The middle mel spectrogram was different from the top one, with /h/ inserted and /w/ replaced with /h/. Note that the accent was also different. These results also indicate that **No_Particle** can usually change the pronunciation of the particles. It is also necessary to estimate that "ha" or "he" is a particle in order to change the pronunciation. If the end-to-end TTS model of **No_Particle** changes the pronunciation of "ha" and "he" regardless of whether they are particles, the naturalness of the synthesized speech decreases. Therefore, the end-to-end TTS model of **No_Particle** may have estimated the parts of speech from phoneme sequences covertly.

5. Conclusions

We investigated units of an input sequence, contexts, pause insertion, vowel devoicing, and the pronunciation for the particles on Japanese end-to-end speech synthesis. Experimental results indicate that a phoneme sequence and H/L accents are effective and accent-phrase boundaries help to insert a pause. The results also indicate that it is possible to change the pronunciation for the devoiced vowels and particles by using the Japanese end-to-end TTS model. Since the dimensions of the input vector are too large, the effectiveness of the traditional full context labels is not determined. The naturalness of the synthesized speech may not be improved unless the additional linguistic features for prosody are compact.

Future work includes investigating the linguistic features necessary to improve the prosodic naturalness and inputting more raw text in end-to-end speech synthesis for non-alphabetic languages such as Japanese and Chinese.

6. Acknowledgements

The MIC/SCOPE #162106106, JSPS KAKENHI Grant Number JP18H04128.

7. References

[1] H. Zen, A. Senior, and M. Schuster, "Statistical parametric speech synthesis using deep neural networks," in *Proceedings of ICASSP*, 2013, pp. 7962–7966.

[2] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *Proceedings of ICLR*, 2015.

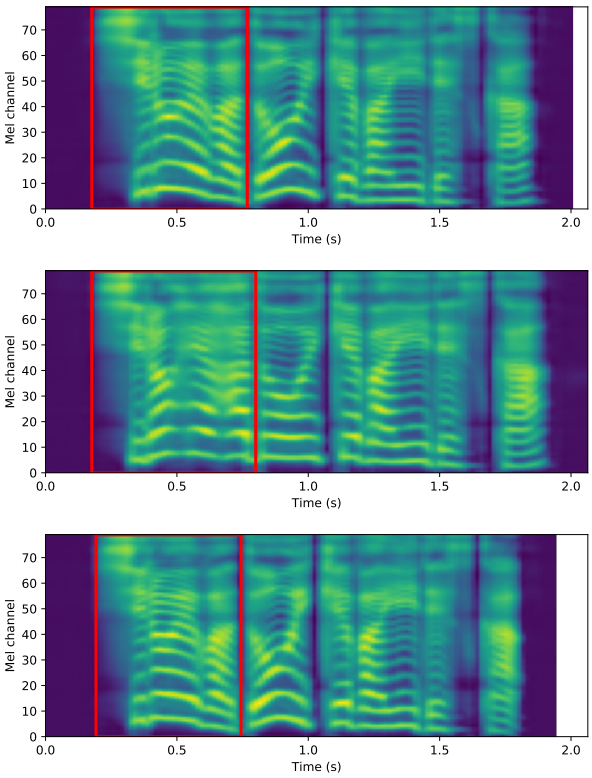

Figure 5: *Mel spectrograms predicted using the end-to-end TTS model for "しろへはどういけばいいですか。 (notation-level morae: shi, ro, he, ha, do, u, i, ke, ba, i, i, de, su, ka)". This sentence has both particles "ha" and "he". Top: Mel spectrogram predicted by inputting /e/, /w/, /a/, which are pronunciation-level phonemes of "he, ha", into* ***Baseline****. Middle: Mel spectrogram predicted by inputting /h/, /e/, /h/, /a/, which are notation-level phonemes of "he, ha", into* ***Baseline****. Bottom: Mel spectrogram predicted by inputting /h/, /e/, /h/, /a/, which are notation-level phonemes of "he, ha", into* ***No_Particle****. Red rectangles show mel spectrograms of "shi, ro, he, ha".*

[3] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, ukasz Kaiser, and I. Polosukhin, "Attention is all you need," in *Proceedings of NIPS*, 2017, pp. 6000–6010.

[4] W. Chan, N. Jaitly, Q. Le, and O. Vinyals, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," in *Proceedings of ICASSP*, 2016, pp. 4960–4964.

[5] J. Sotelo, S. Mehri, K. Kumar, J. F. Santos, K. Kastner, A. Courville, and Y. Bengio, "Char2Wav: End-to-end speech synthesis," in *ICLR2017 workshop submission*, 2017.

[6] Y. Taigman, L. Wolf, A. Polyak, and E. Nachmani, "Voiceloop: Voice fitting and synthesis via a phonological loop," in *Proceedings of ICLR*, 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," in *Proccedings of ICLR*, 2018.

[8] W. Ping, K. Peng, and J. Chen, "Clarinet: Parallel wave generation in end-to-end text-to-speech," in *Proceedings of ICLR*, 2019.

[9] N. Li, S. Liu, Y. Liu, S. Zhao, M. Liu, and M. Zhou, "Close to human quality TTS with transformer," *arXiv:1809.08895*, 2018.

[10] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, Q. Le,

Y. Agiomyrgiannakis, R. Clark, and R. A. Saurous, "Tacotron: Towards end-to-end speech synthesis," in *Proccedings of Interspeech*, 2017, pp. 4006–4010.

[11] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerry-Ryan, R. A. Saurous, Y. Agiomyrgiannakis, and Y. Wu, "Natural TTS synthesis by conditioning WaveNet on Mel spectrogram predictions," in *Proccedings of ICASSP*, 2018, pp. 4779–4783.

[12] 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," in *Proceedings of ICASSP*, 2019, pp. 6905–6909.

[13] K. Kurihara, N. Seiyama, T. Kumano, and A. Imai, "Study of Japanese End-to-End speech synthesis method that inputting kana and prosodic symbols," in *Autumn Meeting of the Acoustical Society of Japan*, 2018, pp. 1083–1084, (in Japanese).

[14] Y. Lu, M. Dong, and Y. Chen, "Implementing prosodic phrasing in chinese end-to-end speech synthesis," in *Proceedings of ICASSP*, 2019, pp. 7050–7054.

[15] A. van den 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:1609.03499*, 2016.

[16] J. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, "Attention-based models for speech recognition," in *Proceedings of NIPS*, 2015, pp. 577–585.

[17] N. Minematsu, S. Kobayashi, S. Shimizu, and K. Hirose, "Improved prediction of Japanese word accent sandhi using CRF," in *Proccedings of Interspeech*, 2012, pp. 2562–2565.

[18] H. Fujisaki, S. Ohno, and S. Yamada, "Analysis of occurrence of pauses and their durations in Japanese text reading," in *Proccedings of ICSLP*, 1998, pp. 1387–1390.

[19] H. Muto, Y. Ijima, N. Miyazaki, H. Mizuno, and S. Sakauchi, "Analysis and evaluation of factors relating pause location for natural text-to-speech synthesis," *Transactions of Information Processing Society of Japan*, pp. 993–1002, 2015.

[20] H. Kawai, N. Higuchi, T. Shimizu, and S. Yamamoto, "Devoicing rules for text-to-speech synthesis of Japanese," *Journal of the Acoustical Society of Japan*, pp. 698–705, 1995, (in Japanese).

[21] H. Kawai, T. Toda, J. Ni, M. Tsuzaki, and K. Tokuda, "XIMERA: A new TTS from ATR based on corpus-based technologies," in *Proceedings of SSW5*, 2004, pp. 179–184.

[22] "Pulse code modulation (pcm) of voice frequencies," *ITU-T. Recommendation G. 711.*, 1988.

[23] "HTS," http://hts.sp.nitech.ac.jp/.