

SINGING VOICE SYNTHESIS BASED ON A MUSICAL NOTE POSITION-AWARE ATTENTION MECHANISM

Yukiya Hono, Kei Hashimoto, Yoshihiko Nankaku, and Keiichi Tokuda

Nagoya Institute of Technology, Nagoya, Japan

ABSTRACT

This paper proposes a novel sequence-to-sequence (seq2seq) model with a musical note position-aware attention mechanism for singing voice synthesis (SVS). A seq2seq modeling approach that can simultaneously perform acoustic and temporal modeling is attractive. However, due to the difficulty of the temporal modeling of singing voices, many recent SVS systems with an encoder-decoder-based model still rely on explicitly on duration information generated by additional modules. Although some studies perform simultaneous modeling using seq2seq models with an attention mechanism, they have insufficient robustness against temporal modeling. The proposed attention mechanism is designed to estimate the attention weights by considering the rhythm given by the musical score. Furthermore, several techniques are also introduced to improve the modeling performance of the singing voice. Experimental results indicated that the proposed model is effective in terms of both naturalness and robustness of timing.

Index Terms— Singing voice synthesis, sequence-to-sequence model, attention mechanism, temporal modeling

1. INTRODUCTION

Statistical parametric singing voice synthesis (SVS) have been evolving with the spread of machine learning techniques [1–6]. The essence of SVS is a sequence transform to generate an acoustic feature sequence from a score feature sequence obtained from musical scores. Since the singing voice needs to be strictly synchronized with the given musical score, how to model the temporal structure of the singing voice is one of the important issues in the SVS.

Typical deep neural network (DNN)-based SVS systems [2–5] model the acoustic feature and its temporal structure (specifically as phoneme duration and vocal timing deviation) independently using acoustic, duration, and time-lag models. The acoustic model works as a mapping function that generates the acoustic feature sequence from the time-aligned score feature sequence. Although these pipeline systems are stable when running in both training and synthesis, they suffer from alignment-related issues: 1) the modeling performance of acoustic features is affected by alignment errors, and 2) they do not have the sufficient ability to adequately model the correlation between the acoustic feature and its temporal structure. These errors may cause a lack of naturalness and expressiveness of synthesized singing voices.

Inspired by the success of autoregressive (AR) and non-AR generation models for modern text-to-speech (TTS) systems [7–14], several neural SVS systems have been developed [15–23]. Unlike TTS, it is not easy to model the temporal structure of a singing

voice because the phoneme duration of singing voices varies greatly depending on the corresponding note duration, even for the same phoneme. Thus, these kinds of SVS systems [15–21] mainly adopt encoder-decoder models with an explicit length regulator. This approach has the advantage of robustness in terms of alignment while it makes the model sensitive to the performance of external duration information, similar to conventional pipeline systems. A number of studies [22, 23] use an AR sequence-to-sequence (seq2seq) model with an attention mechanism to model the acoustic feature and its temporal structure simultaneously; however, they suffer from timing mismatches between the target musical score and synthesized singing voice. This is a critical issue for SVS applications because manually collecting the timing in these attention-based systems is both difficult and impractical.

In this paper, we propose a novel seq2seq model for SVS with a musical note position-aware attention mechanism. The proposed attention mechanism calculates the attention weights based on the note position informed by the musical score. Moreover, we also introduce additional techniques to help obtain robust alignment and improve naturalness: auxiliary note feature embedding, guided attention loss with a penalty matrix designed for singing voice alignments, and pitch normalization for the seq2seq model. The proposed seq2seq model can synthesize singing voices with proper vocal timing without extra supplementary temporal modeling.

2. RELATED WORK

Seq2seq modeling using an attention mechanism [24] is a key technique for modeling the correspondence between sequences with different lengths. In the TTS scenario, since the monotonicity and locality properties of TTS alignment can be exploited, hybrid location-sensitive attention combining content-based and location-based attention mechanisms [9], a forward attention mechanism computed recursively using a recursively forward algorithm [10], and monotonic attention mechanisms [25, 26] have been developed.

One of the noticeable differences between SVS and other typical seq2seq tasks is that the alignment path is strongly related to the note timing. The authors [23] have been proposed a global duration control attention, introducing a global transition token into a forward attention mechanism. Although the tempo of singing voices can be controlled through its token, a time-invariant token has the insufficient ability for duration control and cannot prevent the mismatch between the timing of synthesized singing voices and musical scores. The proposed musical note position-aware attention mechanism is based on the generalized form of forward attention with phoneme-dependent and time-variant transition probabilities, and the attention probability is calculated taking into account note position information via note position embeddings. This achieves appropriate modeling of the temporal structure of the singing voice by the attention mechanism.

This work was supported by JSPS KAKENHI Grant Number JP22H03614, CASIO SCIENCE PROMOTION FOUNDATION, and FOUNDATION OF PUBLIC INTEREST OF TATEMATSU.

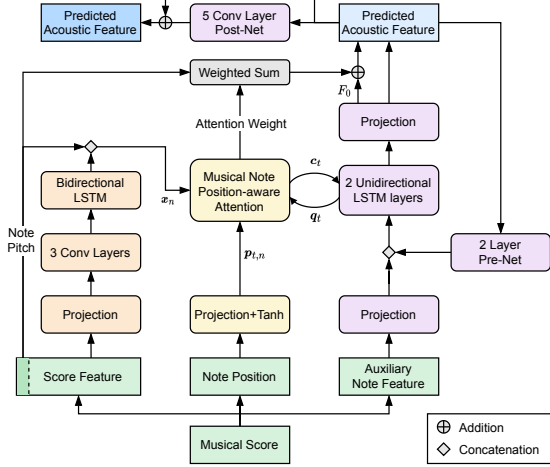


Fig. 1: Overview of the proposed model.

3. PROPOSED MODEL

The proposed model is based on an encoder-decoder model with an attention mechanism and can generate frame-level acoustic feature sequences directly from phoneme-level score feature sequences. The score feature consists of not only the phone, note pitch, and note length but also other rich musical contexts such as beat, key, tempo, dynamics, and staccato [1]. An overview of the proposed model is shown in Fig. 1. We describe four techniques in this section for seq2seq SVS systems to satisfy the singing-specific requirements, such as high controllability and robustness against tempo and pitch.

3.1. Musical note position-aware attention mechanism

The attention mechanism [24] calculates an attention weight at each decoder time-step to perform a soft-selection of encoder hidden state $\mathbf{x} = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N]$, which is obtained by processing the input score features by the encoder. The context vector \mathbf{c}_t can be obtained by $\mathbf{c}_t = \sum_{n=1}^N \alpha_t(n) \mathbf{x}_n$, where $\alpha_t(n)$ is the attention weight representing the degree of attention on the n -th hidden state at the t -th decoder time-step. Finally, the output vector \mathbf{o}_t can be the predicted conditioning of the context vector \mathbf{c}_t by the decoder.

The alignment obtained by the attention mechanism must be monotonic and continuous without skipping any encoder states to synthesize a singing voice following a given musical score. To satisfy this assumption, a current attention weight $\alpha_t(n)$ is calculated recursively using the previous alignment as follows:

$$\alpha'_t(n) = ((1 - u_{t-1}(n))\alpha_{t-1}(n) + u_{t-1}(n-1)\alpha_{t-1}(n-1)) \cdot y_t(n), \quad (1)$$

$$\alpha_t(n) = \alpha'_t(n) / \sum_{m=1}^N \alpha'_t(m), \quad (2)$$

where $y_t(n)$ is the output probability, and $u_t(n)$ is the transition probability where the attention mechanism notices the n -th phoneme at the t -th decoder step and moves to the $n+1$ -th phoneme at the $t+1$ -th decoder steps. Note that $u_t(n)$ in Eq. (1) is the phoneme-dependent time-variant transition probability; thereby this formula can be regarded as a generalized form of forward attention with a transition agent [10].

The alignment of SVS should be obtained by considering the temporal structure of musical notes, specifically note duration and

tempo. Therefore, we introduce musical note positional features to compute $y_t(n)$ and $u_t(n)$.

In the proposed method, the output probability $y_t(n)$ is calculated by extended content-based attention with a musical note position-aware additional term $\mathbf{U}^{(\cdot)} \mathbf{p}_{t,n}$ as:

$$e_t(n) = \mathbf{v}^{(e)\top} \tanh(\mathbf{W}^{(e)} \mathbf{q}_t + \mathbf{V}^{(e)} \mathbf{x}_n + \mathbf{U}^{(e)} \mathbf{p}_{t,n} + \mathbf{b}^{(e)}), \quad (3)$$

$$y_t(n) = \exp(e_t(n)) / \sum_{m=1}^N \exp(e_t(m)), \quad (4)$$

where \mathbf{q}_t denotes the t -th time-step decoder hidden state, and $\mathbf{W}^{(\cdot)} \mathbf{q}_t$ and $\mathbf{V}^{(\cdot)} \mathbf{x}_n$ are the content-based terms that represent query/key comparisons in the attention mechanism, and $\mathbf{b}^{(\cdot)}$ is the bias term. The note position embedded feature $\mathbf{p}_{t,n}$ is obtained by feeding the note position representation $[p_{t,n}^1, p_{t,n}^2, p_{t,n}^3]$ with a single tanh hidden layer. Each note position representation is computed from the note lengths of the given musical score as follows:

$$p_{t,n}^1 = t - s_n, \quad (5)$$

$$p_{t,n}^2 = e_n - t, \quad (6)$$

$$p_{t,n}^3 = \begin{cases} s_n - t, & (t < s_n) \\ 0, & (s_n \leq t \leq e_n) \\ t - e_n, & (e_n < t) \end{cases} \quad (7)$$

where s_n and e_n denote the start and end positions of the n -th musical note, respectively.

Since transition probabilities should be explicitly derived from past alignments, we adopt a location-sensitive attention [9]-like formula to calculate the transition probability as follows:

$$u_t(n) = \sigma(\mathbf{v}^{(u)\top} \tanh(\mathbf{W}^{(u)} \mathbf{q}_t + \mathbf{V}^{(u)} \mathbf{x}_n + \mathbf{U}^{(u)} \mathbf{p}_{t,n} + \mathbf{T}^{(u)} \mathbf{f}_{t,n} + \mathbf{b}^{(u)})), \quad (8)$$

where $\sigma(\cdot)$ is a logistic sigmoid function, and $\mathbf{T}^{(u)} \mathbf{f}_{t,n}$ denotes the location-sensitive term that uses convolutional features computed from the previous cumulative alignments [27].

3.2. Auxiliary note feature embedding

As the temporal structure of the singing voices depends on the context of the notes in the input score, the alignment of singing voices should also be predicted on the basis of it. To encourage this, we embed musical note context associated with the current note as an auxiliary note feature into the attention query.

The auxiliary note features contained musical note-related contexts, which were obtained by removing phoneme and mora-related contexts from the score features, and were expanded to the frame-level sequence using note length. This upsampled feature is then fed to a single dense layer, which is concatenated with the output of Pre-Net to form the decoder input. Since the auxiliary note feature delivers the attention mechanism to the current frame position and note context in the corresponding musical note via the attention query, it is expected to enable the attention mechanism to adjust the alignment to fit the rhythm provided by the musical score.

3.3. Guided attention loss for SVS

Since singing voices are generally sung to follow the rhythm of a musical score, it is natural to assume that the alignment of the singing voice should be close to the path determined by the note timing in the score. On the basis of this idea, we customize the guided attention loss [28] for SVS systems.

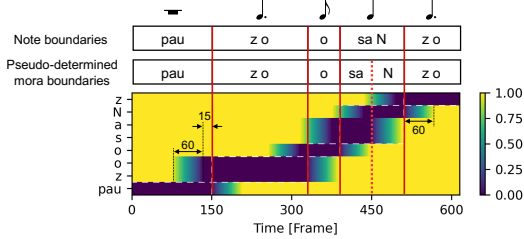


Fig. 2: Penalty matrix examples. This matrix is generated on the basis of pseudo-determined mora boundaries from note durations described in the musical score.

The difference from the original guided attention loss for TTS is how to construct the penalty matrix. We generate a penalty matrix based on the note boundaries, unlike the one for the TTS whose diagonal elements are zero. Specifically, we design a penalty matrix $\mathbf{G} \in \mathbb{R}^{N \times T}$ based on the pseudo-determined mora boundary so that alignment estimation is robust even when multiple morae are included in the same note, as shown in Fig. 2. These pseudo-boundaries are obtained by equally dividing note duration in accordance with the number of morae in each note. The soft matrix is generated by linearly decaying over a width of 60 frames. Note that the start positions of boundaries for constructing the penalty matrix are shifted 15 frames earlier to consider vocal timing deviation.

Let the alignment matrix $\mathbf{A} \in \mathbb{R}^{N \times T}$ be a matrix, where the element at (i, j) corresponds to $\alpha_j(i)$ and T is the total number of frames. The guided attention loss is defined as follows:

$$\mathcal{L}_{\text{att}}(\mathbf{G}, \mathbf{A}) = \frac{1}{NT} \|\mathbf{G} \odot \mathbf{A}\|_1, \quad (9)$$

where \odot denotes an element-wise product. The final loss function \mathcal{L} of the proposed model is given by

$$\mathcal{L} = \mathcal{L}_{\text{feat}}(\mathbf{o}, \hat{\mathbf{o}}) + \mathcal{L}_{\text{feat}}(\mathbf{o}, \hat{\mathbf{o}}') + \lambda \mathcal{L}_{\text{att}}(\mathbf{G}, \mathbf{A}), \quad (10)$$

$$\mathcal{L}_{\text{feat}}(\mathbf{o}, \hat{\mathbf{o}}) = \frac{1}{TD} \sum_{t=1}^T \|\mathbf{o} - \hat{\mathbf{o}}\|_2^2, \quad (11)$$

where $\hat{\mathbf{o}}$ and $\hat{\mathbf{o}}'$ represent the acoustic features predicted by the decoder and Post-Net, respectively, and D is the number of dimensions in the acoustic features. In Eq. (10), λ represents an adjustment parameter for guided attention loss.

3.4. Pitch normalization

The pitch of the synthesized singing voice must accurately follow the note pitch of the musical score. Thus, following our previous work [2], we integrate pitch normalization, where the log fundamental frequency (F_0) is modeled as the difference from the log F_0 determined by the musical score (note pitch), into the proposed model.

Pitch normalization requires a time-aligned frame-level note pitch sequence to process the generated F_0 sequence frame-by-frame. In the proposed system, the frame-level note pitch sequence can be obtained by weighting the phone-level input note pitch sequence using the attention weight at each decoder time-step.

4. EXPERIMENTS

4.1. Experimental conditions

To evaluate the proposed models, we conducted experiments using 70 Japanese children’s songs (total: 70 min) performed by one female singer. Sixty songs were used for training, and the rest were used for testing. Singing voice signals were sampled at 48 kHz,

Table 1: Results of experiment 1 with 95% confidence intervals.

| Systems | $\mathbf{p}_{t,n}$ (Sec. 3.1) | Aux. note feat. (Sec. 3.2) | \mathcal{L}_{att} (Sec. 3.3) | MOS |
|--------------|----------------------------------|-------------------------------|--|-----------------------------------|
| Base | | | | failed |
| NF | | ✓ | | failed |
| NP | ✓ | | | 3.12 ± 0.13 |
| NP+NF | ✓ | ✓ | | 3.81 ± 0.12 |
| Prop | ✓ | ✓ | ✓ | 3.95 ± 0.12 |

and each sample was quantized by 16 bits. The acoustic feature consisted of 0-th through 49-th mel-cepstral coefficients, a continuous log F_0 value, 25-dimensional analysis aperiodicity measures, 1-dimensional vibrato component, and a voiced/unvoiced binary flag. Mel-cepstral coefficients were extracted by WORLD [29]. The difference between the original log F_0 and the median-smoothed log F_0 were used as the vibrato component.

The model architectures of encoder, decoder, Pre-Net, and Post-Net are the same as those of [9]. A linear projection layer was used instead of the embedding layer. We added an extra linear projection layer to process the frame-level auxiliary note feature sequence. The score feature was a 267-dimensional feature vector. The auxiliary note feature was an 87-dimensional feature vector. The frame position in the current note and the note duration were concatenated with an expanded frame-level auxiliary note feature. For all systems, the reduction factor was set to 3, and pitch normalization was applied for F_0 modeling. The hyperparameter λ in Eq. (10) was set to 10.0. All systems were combined with the same pre-trained PeriodNet [30], a non-AR neural vocoder with a parallel structure, to reconstruct waveforms from predicted acoustic features.

4.2. Subjective evaluation

We performed 5-scale mean opinion score (MOS) tests¹ to evaluate the naturalness of the synthesized singing voices. Each of the 15 native Japanese-speaking participants evaluated ten phrases randomly selected from the test songs. A click sound generated based on the basis of the tempo of the score was superimposed on the synthesized singing voice to evaluate the overall naturalness considering the vocal timing.

4.2.1. Experiment 1

We first evaluated the effectiveness of the proposed techniques for modeling the temporal structure of the singing voice described in Section 3. We used five systems as shown in Table 1. In this experiment, the attention weights were calculated following Eqs. (1) and (2). In **Base** and **NF**, we calculated the output and transition probabilities without using $\mathbf{U}^{(\cdot)} \mathbf{p}_{t,n}$ in Eqs. (4) and (8).

Table 1 shows the subjective evaluation results, and Fig. 3 shows a number of examples of the alignments predicted by each system. From Fig. 3, **Base** and **NF** cannot obtain an appropriate alignment of the singing voice, which indicates the proposed musical note position-aware attention mechanism is effective. Since synthesized samples by **Base** and **NF** contained many timing and lyrics errors, they were excluded from the MOS test. **NP+NF** achieved much higher MOS scores than **NP**. This shows that the use of auxiliary note features is effective. The figure also shows that **Prop** can obtain a more monotonic alignment and better MOS scores than **NP+NF** and thus demonstrating the effectiveness of the proposed guided attention loss described in Section 3.3. Note that a previous study [22]

¹Audio samples are available at the following URL: <https://www.sp.nitech.ac.jp/~hono/demos/icassp2023/>

7. REFERENCES

- [1] K. Oura, A. Mase, T. Yamada, S. Muto, Y. Nankaku, and K. Tokuda, "Recent development of the HMM-based singing voice synthesis system—Sinsy," in *Proc. ISCA SSW7*, 2010, pp. 211–216.
- [2] Y. Hono, K. Hashimoto, K. Oura, Y. Nankaku, and K. Tokuda, "Sinsy: A deep neural network-based singing voice synthesis system," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 2803–2815, 2021.
- [3] M. Blaauw and J. Bonada, "A neural parametric singing synthesizer," in *Proc. Interspeech*, 2017, pp. 4001–4005.
- [4] Y. Hono, K. Hashimoto, K. Oura, Y. Nankaku, and K. Tokuda, "Singing voice synthesis based on generative adversarial networks," in *Proc. ICASSP*, 2019, pp. 6955–6959.
- [5] K. Nakamura, S. Takaki, K. Hashimoto, K. Oura, Y. Nankaku, and K. Tokuda, "Fast and high-quality singing voice synthesis system based on convolutional neural networks," in *Proc. ICASSP*, 2020, pp. 7239–7243.
- [6] Y.-H. Yi, Y. Ai, Z.-H. Ling, and L.-R. Dai, "Singing voice synthesis using deep autoregressive neural networks for acoustic modeling," in *Proc. Interspeech*, 2019, pp. 2593–2597.
- [7] J. Sotelo, S. Mehri, K. Kumar, J. F. Santos, K. Kastner, A. Courville, and Y. Bengio, "Char2wav: End-to-end speech synthesis," in *Proc. ICLR Workshop Track*, 2017.
- [8] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio *et al.*, "Tacotron: Towards end-to-end speech synthesis," in *Proc. Interspeech*, 2017, pp. 4004–4010.
- [9] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerry-Ryan *et al.*, "Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions," in *Proc. ICASSP*, 2018, pp. 4779–4783.
- [10] J.-X. Zhang, Z.-H. Ling, and L.-R. Dai, "Forward attention in sequence-to-sequence acoustic modeling for speech synthesis," in *Proc. ICASSP*, 2018, pp. 4789–4793.
- [11] N. Li, S. Liu, Y. Liu, S. Zhao, and M. Liu, "Neural speech synthesis with transformer network," in *Proc. AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 6706–6713.
- [12] Y. Ren, C. Hu, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, "FastSpeech 2: Fast and high-quality end-to-end text to speech," in *Proc. ICLR*, 2021.
- [13] J. Shen, Y. Jia, M. Chrzanowski, Y. Zhang, I. Elias, H. Zen, and Y. Wu, "Non-attentive Tacotron: Robust and controllable neural tts synthesis including unsupervised duration modeling," *arXiv preprint arXiv:2010.04301*, 2020.
- [14] I. Elias, H. Zen, J. Shen, Y. Zhang, Y. Jia, R. J. Weiss, and Y. Wu, "Parallel Tacotron: Non-autoregressive and controllable tts," in *Proc. ICASSP*, 2021, pp. 5709–5713.
- [15] J. Lee, H.-S. Choi, C.-B. Jeon, J. Koo, and K. Lee, "Adversarially trained end-to-end korean singing voice synthesis system," in *Proc. Interspeech*, 2019, pp. 2588–2592.
- [16] M. Blaauw and J. Bonada, "Sequence-to-sequence singing synthesis using the feed-forward transformer," in *Proc. ICASSP*, 2020, pp. 7229–7233.
- [17] P. Lu, J. Wu, J. Luan, X. Tan, and L. Zhou, "XiaoiceSing: A high-quality and integrated singing voice synthesis system," in *Proc. Interspeech*, 2020, pp. 1306–1310.
- [18] J. Chen, X. Tan, J. Luan, T. Qin, and T.-Y. Liu, "HiFiSinger: Towards high-fidelity neural singing voice synthesis," *arXiv preprint arXiv:2009.01776*, 2020.
- [19] J. Wu and J. Luan, "Adversarially trained multi-singer sequence-to-sequence singing synthesizer," in *Proc. Interspeech*, 2020, pp. 1296–1300.
- [20] Y. Gu, X. Yin, Y. Rao, Y. Wan, B. Tang, Y. Zhang, J. Chen, Y. Wang, and Z. Ma, "ByteSing: A chinese singing voice synthesis system using duration allocated encoder-decoder acoustic models and wavrn vocoders," in *Proc. ISCSLP*, 2021, pp. 1–5.
- [21] J. Shi, S. Guo, N. Huo, Y. Zhang, and Q. Jin, "Sequence-to-sequence singing voice synthesis with perceptual entropy loss," in *Proc. ICASSP*, 2021, pp. 76–80.
- [22] O. Angelini, A. Moinet, K. Yanagisawa, and T. Drugman, "Singing synthesis: with a little help from my attention," in *Proc. Interspeech*, 2020, pp. 1221–1225.
- [23] T. Wang, R. Fu, J. Yi, J. Tao, and Z. Wen, "Singing-Tacotron: Global duration control attention and dynamic filter for end-to-end singing voice synthesis," *arXiv preprint arXiv:2202.07907*, 2022.
- [24] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *Proc. ICLR*, 2015.
- [25] M. He, Y. Deng, and L. He, "Robust sequence-to-sequence acoustic modeling with stepwise monotonic attention for neural TTS," in *Proc. Interspeech*, 2019, pp. 1293–1297.
- [26] Y. Yasuda, X. Wang, and J. Yamagishi, "Initial investigation of encoder-decoder end-to-end tts using marginalization of monotonic hard alignments," in *Proc. ISCA SSW10*, 2019, pp. 211–216.
- [27] J. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, "Attention-based models for speech recognition," in *Advances in Neural Information Processing Systems*, 2015, pp. 577–585.
- [28] H. Tachibana, K. Uenoyama, and S. Aihara, "Efficiently trainable text-to-speech system based on deep convolutional networks with guided attention," in *Proc. ICASSP*, 2018, pp. 4784–4788.
- [29] M. Morise, F. Yokomori, and K. Ozawa, "WORLD: A vocoder-based high-quality speech synthesis system for real-time applications," *IEICE Transactions on Information and Systems*, vol. 99, no. 7, pp. 1877–1884, 2016.
- [30] Y. Hono, S. Takaki, K. Hashimoto, K. Oura, Y. Nankaku, and K. Tokuda, "PeriodNet: A non-autoregressive raw waveform generative model with a structure separating periodic and aperiodic components," *IEEE Access*, vol. 9, pp. 137 599–137 612, 2021.
- [31] H. Zen, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "A hidden semi-Markov model-based speech synthesis system," *IEICE Transactions on Information and Systems*, vol. E90-D, no. 5, pp. 825–834, 2007.
- [32] Y. Nankaku, K. Sumiya, T. Yoshimura, S. Takaki, K. Hashimoto, K. Oura, and K. Tokuda, "Neural sequence-to-sequence speech synthesis using a hidden semi-markov model based structured attention mechanism," *arXiv preprint arXiv:2108.13985*, 2021.