INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

Enhancing Monotonicity for Robust Autoregressive Transformer TTS

Xiangyu Liang[1,2], *Zhiyong Wu*[1,2,4,*], *Runnan Li*[3], *Yanqing Liu*[3], *Sheng Zhao*[3], *Helen Meng*[1,4]

[1]Tsinghua-CUHK Joint Research Center for Media Sciences, Technologies and Systems, Shenzhen International Graduate School, Tsinghua University, Shenzhen, China
[2] Department of Computer Science and Technology, Tsinghua University, Beijing, China
[3] Search Technology Center Asia (STCA), Microsoft
[4] Department of Systems Engineering and Engineering Management, The Chinese University of Hong Kong, Shatin, N.T., Hong Kong SAR, China

liangxy18@mails.tsinghua.edu.cn, {Runnan.Li, yanqliu, Sheng.Zhao}@microsoft.com, {zywu,hmmeng}@se.cuhk.edu.hk

Abstract

With the development of sequence-to-sequence modeling algorithms, Text-to-Speech (TTS) techniques have achieved significant improvement in speech quality and naturalness. These deep learning algorithms, such as recurrent neural networks (RNNs) and its memory enhanced variations, have shown strong reconstruction ability from input linguistic features to acoustic features. However, the efficiency of these algorithms is limited for its sequential process in both training and inference. Recently, Transformer with superiority in parallelism is proposed to TTS. It employs the positional embedding instead of recurrent mechanism for position modeling and significantly boosts training speed. However, this approach lacks monotonic constraint and is deficient with issues like pronunciation skipping. Therefore, in this paper, we propose a monotonicity enhancing approach with the combining use of Stepwise Monotonic Attention (SMA) and multi-head attention for Transformer based TTS system. Experiments show the proposed approach can reduce bad cases from 53 of 500 sentences to 1, together with an improvement on MOS from 4.09 to 4.17 in the naturalness test.

Index Terms: Text-to-Speech, Transformer, Monotonic attention

1. Introduction

Text-to-speech (TTS), as one essential component in human-computer speech interaction, aiming to synthesize speech on human-parity quality, has attracted increasing research interests. Conventional statistical parametric speech synthesis approaches, such as hidden Markov models (HMMs) [1], can produce speech with good naturalness and controllability, but are suffered from low quality, complex training pipelines and huge requirement for expertly labeled data.

Recently, neural end-to-end approaches are proposed to lower the barriers for developing high-quality text-to-speech systems [2, 3, 4, 5]. Most of these approaches are constructed with recurrent neural networks (RNNs) and its memory enhanced variations following sequence-to-sequence structures [3, 6]. These approaches generally encode the input to intermediate states for linguistic information extraction, then decode the encoded states to targeted acoustic parameters. An additional vocoder [7, 8, 9] is then employed to produce waveforms with the generated features. An attention mechanism is generally employed to align the encoder and decoder states for source-target correspondence, contributes crucial influence on the naturalness and robustness of the synthetic speech.

To achieve better performance, many researches have been done to explore different attention mechanisms for RNN based TTS systems. Tacotron2 [3] adopts a location-sensitive attention [10] to exploit alignment information from previous steps to current inference, which could mitigate repetition and omitting problems compared with conventional additive attention mechanism [11]. Forward attention [12], monotonic attention [13] and dynamic convolution attention [14] are also proposed with their properties in modeling monotonic information in TTS. However, being proposed with RNN structure, these mechanisms are based on bringing in additional sequential dependence and thus have limited computing efficiency.

To optimize the efficiency of TTS system, a none-recurrent Transformer based TTS system is proposed [15]. Transformer [16] is solely developed with attention mechanisms and dispensed with recurrences and convolutions entirely and shows extraordinary performance on neural machine translation (NMT) task compared with conventional RNN-based approaches. Being free from recurrent structure, Transformer based TTS benefits from parallel computing in training, and with multi-head attention mechanism it can learn long-distance dependency to produce speech with natural prosody [17, 18]. However, different from the source-target mapping in NMT tasks, the modeling from linguistic feature to acoustic feature in TTS is of monotonic essence, which is weak in the original design of Transformer. This weakness makes the system suffer from robust issues, such as word skipping and repeating, which further results in lower perception satisfaction to users.

Alleviating these issues requires the monotonicity enhancement for Transformer. However, considering the parallel computation property, conventional monotonicity enhancing mechanisms for RNN structures cannot be simply integrated with Transformer. In this paper, we propose a novel monotonicity enhanced attention with the combining usage of multi-head attention and Stepwise Monotonic Attention (SMA) [19]. In the proposed method, we detect those heads with diagonal patterns in pre-trained multi-head attention and then employ SMA to tune these heads to get more focused and accurate monotonic alignments. This helps to improve the system's robustness while retaining other heads with scattered alignments for improving naturalness of speech. With the proposed attention mechanism, experimental results show the proposed approach is more robust compared with baseline Transformer-TTS, with

* Corresponding author

http://dx.doi.org/10.21437/Interspeech.2020-1751

bad case counts from 53 of 500 sentences to 1. Furthermore, the proposed model has also achieved better evaluation results in MOS test, from 4.09 to 4.17 in naturalness with the baseline model.

The contributions of this paper can be summarized as:

1. Propose a novel monotonicity enhanced attention approach with the combining use of multi-head attention and Stepwise Monotonic Attention for Transformer.
2. Significantly alleviate the robust issues in Transformer based TTS systems.
3. Further improve the naturalness of synthetic speech.
4. Provide more precise alignments for knowledge distillation models, such as FastSpeech [20] student models.

2. Methodology

2.1. Multi-head Attention

Transformer[16] is a transduction model entirely constructed with multi-head self-attention for input representations extraction and target features reconstruction without any recurrence or convolution. For the encoder in Transformer based TTS system, the core self-attention mechanism is employed to extract the long-time dependency between input features with the consideration of the sentence-level contextual information. For the decoder, the self-attention mechanism can model the dependency between any frame pair to enable high-quality acoustic features prediction. These properties are essential for natural speech production, especially for long speeches. To align unparalleled encoded input and targeted output, multi-head attention is also employed for encoder-decoder alignment.

For given query matrix $Q \in \mathbb{R}^{n\times d}$, key matrix $K \in \mathbb{R}^{l\times d}$, and value matrix $V \in \mathbb{R}^{l\times d}$, where n, l, d represents the length of target sequence, the length of source sequence, and the dimension of each sequence respectively, the muti-head attention algorithm computes the hidden output H_i as following:

$$H_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V) \tag{1}$$

where Q, K, V are projected to h sub-spaces for computation by multiplying trainable projection matrix $W_i^Q, W_i^K, W_i^V (i = 1, 2, ..., h)$. For each sub-space, Scaled Dot-Product Attention (SDA) is employed to compute the attention output:

$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d}})V \tag{2}$$

By concatenating outputs from all sub-spaces and passing through a linear projection with weight W^O, the final output is computed as:

$$\text{MultiHead}(Q, K, V) = \text{Concat}(H_1, ..., H_h)W^O \tag{3}$$

In Transformer TTS, stacked multi-head attention layers are employed to extract input linguistic representation in encoder progressively, and to reconstruct target acoustic features in the decoder gradually. The model can also benefit from multi-head attention since each sub-space can represent a different aspect for the modeling. This mechanism can enhance the robustness and naturalness of the synthetic speech.

However, limitations still exist in these Transformer based TTS approaches: the weakness in modeling location sensitive information and the lacking of monotonic constraints. Unlike RNN based approaches using recurrent attention to produce consistent attention, the recurrence and convolution free Transformer cannot model the position information, results in ignoring the orders for both input and output sequence. However, for TTS task, the order information of input linguistic features and output acoustic features are vital for speech generation. To alleviate this issue, an additional scaled positional encoding is employed to import position information:

$$PE(pos, 2c) = sin(\frac{pos}{10000^{\frac{2c}{d}}}) \tag{4}$$

$$PE(pos, 2c+1) = cos(\frac{pos}{10000^{\frac{2c}{d}}}) \tag{5}$$

$$x'_{pos} = x_{pos} + \beta PE(pos) \tag{6}$$

where pos is the time step index, $2c$ and $2c + 1$ is the channel index, d is the dimension and β is a trainable scale factor. The position embedding is employed to both encoder and decoder by adding to origin input x_{pos}, providing absolute position information to the model. However, this mechanism is still weak in modeling location dependency and monotonicity of speech, result in word skipping and repeating in speech generation. Therefore, location sensitive monotonicity enhancement approach is required.

2.2. Stepwise Monotonic Attention

Stepwise Monotonic Attention (SMA) [19] is proposed to enhance robustness for RNN based TTS approaches by introducing additional monotonic constraint and has proven with effective robustness enhancing performance.

Suppose the query of previous time step $i - 1$ attend to key j, by sampling $z_{i,j} \sim \text{Bernoulli}(p_{i,j})$, the query of current step i will either stop and attend to j if $z_{i,j} = 1$, or one step forward to attend to $j + 1$ if $z_{i,j} = 0$. For query i the process only keeps j unmoved or one step forward to stop at $j = j + 1$, the stepwise monotonicity constraint is thus imposed.

The sampling probability $p_{i,j}$ for given query at time step i to stop at key j is:

$$p_{i,j} = \text{sigmoid}(e_{i,j} + \mathcal{N}(0, 1)) \tag{7}$$

$$e_{i,j} = \text{Energy}(\boldsymbol{q}_i, \boldsymbol{k}_j, \theta) \tag{8}$$

where $\text{sigmoid}()$ is the non-linear sigmoid activation function, $e_{i,j}$ is the energy function to score how well the input query $\boldsymbol{q}_i$ and key $\boldsymbol{k}_j$ matches, θ denotes additional parameters for energy functions. The $\mathcal{N}(0, 1)$ is the Gaussian noise employed to bridge the gap between sampling and expectation. The expectation is employed in training since the sampling operation will block the gradient back propagation in network. The expectation value $\alpha_{i,j}$ can be calculated in a recursive formula according to the tactic of sampling:

$$\alpha_{i,j} = \alpha_{i-1,j-1}(1 - p_{i,j-1}) + \alpha_{i-1,j}p_{i,j} \tag{9}$$

with paralleled computing formula as:

$$\boldsymbol{\alpha}_i = \boldsymbol{\alpha}_{i-1} \cdot \boldsymbol{p}_i + [0; \alpha_{i-1,:-1} \cdot (1 - p_{i,:-1})] \tag{10}$$

Two different approaches named soft and hard decoding can be employed in inference. The soft decoding directly use the value of $\alpha_{i,j}$ as alignment weight, and the hard one will sample from $\alpha_{i,j}$ with 0 or 1 under Bernoulli distribution. In the TTS task, the soft inference approach could reduce the influence from context mismatching and provide more robust performance comparing with the hard one.

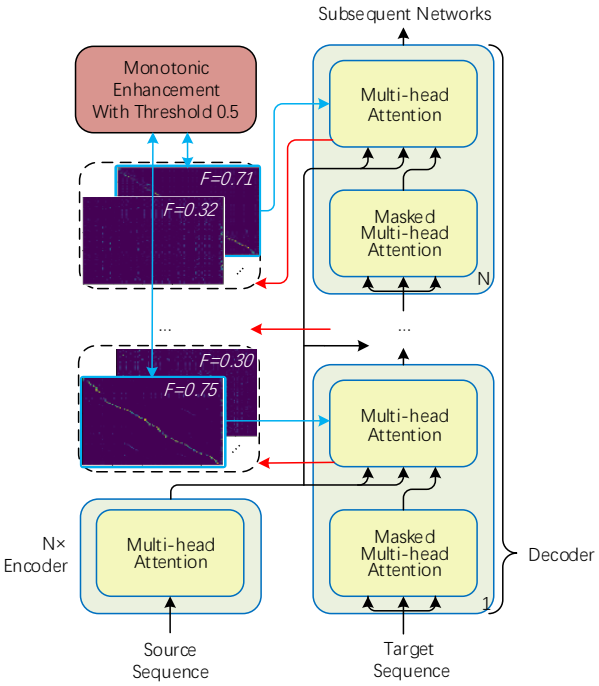

Figure 1: *The proposed multi-head attention with SMA. The encoder-decoder alignments are sent to SMA blocks for focus rates computing. Those alignment results with focus rates greater than prefixed threshold are then sent to SMA for imposing monotonicity constraints.*

2.3. Stepwise Monotonic Attention Tuned Process

Encoder-decoder alignments are of great importance in end-to-end TTS systems. In alignments from Transformer TTS, those with focused diagonal pattern may directly contribute to the speech robustness and naturalness. We proposed a SMA based alignment tuning approach to optimize these alignment results, to improve the robustness and naturalness of synthetic speech by introducing monotonic constraint and location sensitive information modeling.

SMA can be simply introduced to Transformer by replacing the attention computation Eq.(2). To produce SMA attention, probability sampling matrix P is firstly computed as:

$$\mathrm{P} = \mathrm{sigmoid}(\frac{QK^T}{\sqrt{d}} + \mathcal{N}(0,1)) \tag{11}$$

where $\mathrm{sigmoid}()$ is the non-linear sigmoid activation function, $\mathcal{N}(0,1)$ is the Gaussian noise employed to bridge the gap between sampling and expectation. Then P is split into probability vector for steps from 1 to n, $[\boldsymbol{p}_1, \boldsymbol{p}_2, ..., \boldsymbol{p}_n]$, and expectation vectors $[\boldsymbol{\alpha_1}, \boldsymbol{\alpha_2}, ..., \boldsymbol{\alpha_n}]$ are computed following Eq.(10). Attention matrix A is then produced by concatenating the expectation vectors. Thus Eq.(2) can be modified into:

$$\mathrm{Attention}(Q, K, V) = AV \tag{12}$$

Considering not all the alignments produced from multi-head attention are following the diagonal pattern, and scattered ones may also represent one mapping dependency across encoded sequence and targeted decoding sequence, it is recommended to select the diagonal alignments for monotonicity enhancing while retaining the scattered alignments. We propose the usage of focus rate [20] to auto-select the alignment heads for SMA tuning. The focus rate is computed as:

$$F = \frac{1}{n}\sum_{i=1}^{n} \max_{1 \le j \le l} \hat{\alpha}_{i,j} \tag{13}$$

where $\hat{\alpha}_{i,j}$ is the element in the attention matrix. A prefixed threshold is defined and those alignment heads with greater focus rate will be sent for SMA optimization, as shown in Fig.(1). Empirically, we set the threshold to 0.5.

Since SMA is proposed as a tuning approach, it is suggested to pre-train a basic Transformer model. Thus we can conclude the training process in two steps: 1) train an original Transformer TTS network until it converges to a certain step when all diagonal alignments stably appear; 2) enable SMA instead of SDA in Transformer and continue training until the model converges. Following this process, we can donate benefits as:

- With a pre-trained source Transformer model, SMA can be stably adopted within less training epochs, and significantly reduce the extra training time cost caused by imposing SMA.
- The monotonicity enhancement is only adopted to diagonal alignments, with other scattered alignments retained. This helps the proposed approach maintain original modeling ability in exploiting contextual dependencies across the input and contribute to the naturalness and quality of synthetic speech.

3. Experiments

3.1. Experiment Setup

In experiments, LJSpeech [21] is employed for evaluation, which contains 13,100 recordings from a native female speaker. 12600, 250, 250 instances split from the corpus are used as the training, validation, test sets, respectively. Phoneme sequences extracted from the text are employed as the source input, and 80-dimension of mel-spectrogram is employed as the target acoustic parameters.

To better evaluate the performance of the proposed system, three different systems with different set-up are constructed, including an original Transformer baseline, a monotonicity enhanced SMA approach with hard decoding, and the proposed monotonicity enhancing SMA approach with soft decoding, each has the setting as follows:

- **Baseline**: The baseline Transformer-TTS, which has two fully-connected layers as decoder prenets, each with 64 hidden units. Both encoder and decoder are constructed with 4 multi-head attention blocks, each employs 512 hidden units, and the following feed-forward networks employ 1024 hidden units. Only Scaled Dot-Product Attention is employed.
- **SMA (hard)**: Approach with SMA hard decoding. Sharing the same hyper-parameter settings with the baseline approach, but with enabled monotonicity enhancing mechanism in training and with SMA hard decoding.
- **SMA (soft)**: Approach with SMA soft decoding. Sharing the same hyper-parameter settings with the baseline approach, but with enabled monotonicity enhancing mechanism in training and with SMA soft decoding.

These comparison systems are implemented with ESPNET development framework [22], and all models are trained with 4 cards of Tesla V100 GPU. As listed in Table.1, comparison models have similar scale and share similar time consuming in training. SMA tuned approaches employ 600 epoch pre-trained Transformer-TTS as the source model, and processes another 50 epochs for fine-tuning. For the vocoder, we use Griffin-Lim in the Robustness test for simplicity and use WaveNet in the naturalness test. Example audios are available at [1].

Table 1: *Quantitative description of models.*

	Total Parameters	Time Cost	Epochs
Baseline	22.34 m	18 h	1000
SMA Tuned	22.34 m	19 h	600+50

3.2. Robustness Test

The robustness test is firstly conducted to evaluate the stability of different comparisons in speech generation. In this test, the bad cases are counted to show how robust the system is in large-scale testing. The error cases are categorized as word-level repeating, skipping, and mispronunciation. All the reserved sentences in validation and test set are employed in this test, thus 500 sentences containing 8,320 words in total are included.

The evaluation result is shown in Table.2. The baseline Transformer-TTS is suffered from bad repeating and skipping issues, produces 115 word-level bad cases in total. With the monotonicity enhanced approach, these issues are significantly alleviated: only 7 word-level bad cases are observed in SMA (hard) approach and only 1 word-level bad case is observed in SMA (soft). This helps sentence-level bad case rate drop from 10.6% in Transformer baseline to 1.2% in SMA (hard), and 0.2% in SMA (soft).

Table 2: *Evaluation result on 500 sentences in the robustness test.* ***Words*** *denotes the word-level bad case numbers, and* ***Sentences*** *counts the sentences with word-level bad cases.*

	Baseline	SMA (hard)	SMA (soft)
Repeat	55	1	0
Skipping	57	0	0
Mispro	3	6	1
Words	115	7	1
Sentences	53	6	1

3.3. Naturalness Evaluation

Naturalness evaluation is then conducted to evaluate the perception quality of the synthetic speech. In this test, 50 out-domain sentences are used to generate speech samples. And 20 native language experts are invited to score the recordings as well as the samples generated by different approaches, in which samples are scored as 1=Poor, 2=Bad, 3=Fair, 4=Good, 5=Excellent.

Table 3: *Results of the Mean Opinion Scores (MOS) of the speeches, with confidence interval as 95%.*

	MOS scores
Ground Truth	4.52 ± 0.11
Baseline	4.09 ± 0.06
SMA (hard)	4.09 ± 0.06
SMA (soft)	4.17 ± 0.06

[1] https://thuhcsi.github.io/interspeech2020-monotonicity-transformer-tts/

The evaluation result is shown in Table.3. When introducing hard decoding SMA to Transformer-TTS, the naturalness of synthetic speech has achieved on par performance with the Transformer baseline. And when using soft decoding SMA instead, the proposed approach has achieved a further improvement on the naturalness, achieving +0.08 MOS gain compared with the Transformer baseline.

3.4. Alignment Case Study

Fig.2 shows a typical sentence-level bad case in the evaluation. In this sentence, repeating issues happen in Transformer baseline system around "the privacy of the" where the two "the" are similarly pronounced, leading to misalignment and causing repeating bad case. With the monotonicity enhancement, this issue is not observed in the proposed system. And with SMA, the produced alignment also has a higher focus rate in alignment than the baseline, this may also help the system produce robust speech.

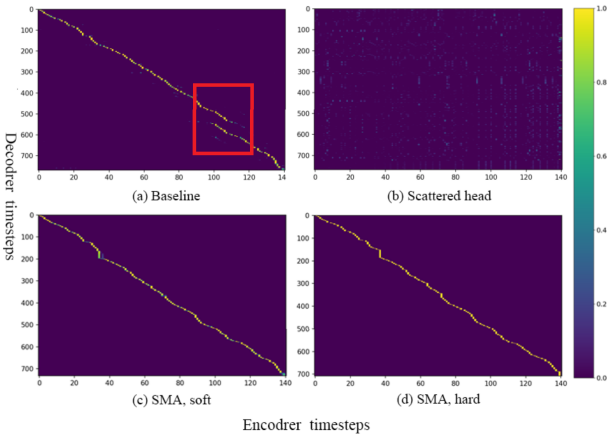

Figure 2: *An instance of "Presidents have made it clear, however, that they did not favor this or any other arrangement which interferes with the privacy of the president and his guests." Picture (a) (c) (d) show the diagonal alignments with highest focus rate in baseline, SMA(soft) and SMA(hard) respectively.(b) is an example of head with scattered content from baseline.*

4. Conclusions

In this paper, we proposed a monotonicity enhanced approach for Transformer TTS systems. By introducing Stepwise Monotonic Attention to tune alignment results from multi-head attention, the proposed approach can significantly improve the robustness of synthetic speech from Transformer-TTS and also improve the synthetic naturalness. We believe our methods could also be applied to other tasks applying Transformer structure but need monotonic constraint.

5. Acknowledgements

This work was conducted when the first author was an intern at Microsoft, and is supported by joint research fund of National Natural Science Foundation of China - Research Grant Council of Hong Kong (NSFC-RGC) (61531166002, N_CUHK404/15), National Natural Science Foundation of China (61433018, 61375027), the Major Project of National Social Science Foundation of China (13&ZD189).

6. References

[1] A. W. Black, H. Zen, and K. Tokuda, "Statistical parametric speech synthesis," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 4, April 2007, pp. IV–1229–IV–1232.

[2] Y. Wang, R. J. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, Q. V. Le, Y. Agiomyrgiannakis, R. Clark, and R. A. Saurous, "Tacotron: A fully end-to-end text-to-speech synthesis model," *Conference of the International Speech Communication Association (Interspeech)*, p. 4006–4010.

[3] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan, R. A. Saurous, Y. Agiomvrgiannakis, and Y. Wu, "Natural tts synthesis by conditioning wavenet on mel spectrogram predictions," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, April 2018, pp. 4779–4783.

[4] 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," *The International Conference on Learning Representations (ICLR)*, Oct 2018.

[5] Y. Taigman, L. Wolf, A. Polyak, and E. Nachmani, "VoiceLoop: Voice Fitting and Synthesis via a Phonological Loop," *The International Conference on Learning Representations (ICLR)*, Jul 2018.

[6] J. Sotelo, S. Mehri, K. Kumar, J. F. Santos, K. Kastner, A. Courville, and Y. Bengio, "Char2wav: End-to-end speech synthesis," *The International Conference on Learning Representations (ICLR)*.

[7] 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 e-prints*, Sep 2016.

[8] S. Mehri, K. Kumar, I. Gulrajani, R. Kumar, S. Jain, J. Sotelo, A. Courville, and Y. Bengio, "SampleRNN: An Unconditional End-to-End Neural Audio Generation Model," *The International Conference on Learning Representations (ICLR)*, Dec 2017.

[9] 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.

[10] J. K. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, "Attention-based models for speech recognition," in *Advances in neural information processing systems (NIPS)*, 2015, pp. 577–585.

[11] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *The International Conference on Learning Representations (ICLR)*, 2015.

[12] J.-X. Zhang, Z.-H. Ling, and L.-R. Dai, "Forward Attention in Sequence-to-sequence Acoustic Modelling for Speech Synthesis," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Jul. 2018.

[13] C. Raffel, M.-T. Luong, P. J. Liu, R. J. Weiss, and D. Eck, "Online and Linear-Time Attention by Enforcing Monotonic Alignments," *International Conference on Machine Learning (ICML)*, Apr. 2017.

[14] E. Battenberg, R. Skerry-Ryan, S. Mariooryad, D. Stanton, D. Kao, M. Shannon, and T. Bagby, "Location-Relative Attention Mechanisms For Robust Long-Form Speech Synthesis," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Oct. 2019.

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

[16] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention Is All You Need," *Neural Information Processing Systems (NIPS)*, Jun. 2017.

[17] Y. Shan, H. Lu, S. Kang, L. Xue, J. Xiao, D. Su, L. Xie, and D. Yu, "On the localness modeling for the self-attention based end-to-end speech synthesis," in *Neural Networks, Elsevier*, 2020.

[18] S. Yang, H. Lu, S. Kang, L. Xie, and D. Yu, "Enhancing Hybrid Self-Attention Structure With Relative-Position-Aware Bias For Speech Synthesis," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2019.

[19] M. He, Y. Deng, and L. He, "Robust Sequence-to-Sequence Acoustic Modeling with Stepwise Monotonic Attention for Neural TTS," *Conference of the International Speech Communication Association (Interspeech)*, Jun. 2019.

[20] Y. Ren, Y. Ruan, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, "FastSpeech: Fast, Robust and Controllable Text to Speech," *Neural Information Processing Systems (NIPS)*, May 2019.

[21] K. Ito, "The lj speech dataset," https://keithito.com/LJ-Speech-Dataset/, 2017.

[22] S. Watanabe, T. Hori, S. Karita, T. Hayashi, J. Nishitoba, Y. Unno, N. Enrique Yalta Soplin, J. Heymann, M. Wiesner, N. Chen, A. Renduchintala, and T. Ochiai, "Espnet: End-to-end speech processing toolkit," in *Conference of the International Speech Communication Association (Interspeech)*, 2018, pp. 2207–2211.