ENHANCING HYBRID SELF-ATTENTION STRUCTURE WITH RELATIVE-POSITION-AWARE BIAS FOR SPEECH SYNTHESIS

Shan Yang[⋆†]*, Heng Lu*[†]*, Shiying Kang*[†]*, Lei Xie*[⋆]*, Dong Yu*[†]

[⋆]School of Computer Science, Northwestern Polytechnical University, Xi'an, China
[†]Tencent AI Lab

ABSTRACT

Compared with the conventional "front-end"–"back-end"–"vocoder" structure, based on the attention mechanism, end-to-end speech synthesis systems directly train and synthesize from text sequence to the acoustic feature sequence as a whole. Recently, a more calculation efficient end-to-end architecture named transformer, which is solely based on self-attention, was proposed to model global dependencies between the input and output sequences. However, although with many advantages, transformer lacks position information in its structure. Moreover, the weighted sum form in self-attention may disperse the attention to the whole input sequence other than focusing on the more important neighbouring positions. In order to solve the above problems, this paper introduces a hybrid self-attention structure which combines self-attention with the recurrent neural networks (RNNs). We further enhance the proposed structure with relative-position-aware biases. Mean opinion score (MOS) test results indicate that by enhancing hybrid self-attention structure with relative-position-aware biases, the proposed system achieves the best performance with only 0.11 MOS score lower than natural recording.

***Index Terms*—** text-to-speech synthesis, sequence-to-sequence model, self-attention, relative-position-aware representation

1. INTRODUCTION

Text-to-speech (TTS) technology advances as the we move from hidden Markov models (HMM) to neural networks (NN) approaches [1, 2, 3]. Generally speaking, the conventional TTS frameworks are composed of three major parts: 1) a complex "front-end" module to analyze the raw text into linguistic feature. 2) a "back-end" module, which learns and transforms linguistic features to acoustic features. And finally, 3) a "vocoder", to reconstruct waveform from the generated acoustic features. Normally, the "front-end" module is language dependent, and designing a good "front-end" requires expert knowledge on specific language as well as great time and effort. Recently, the successful applications of attention-based sequence-to-sequence (seq2seq) model [4, 5] are demonstrated to outperform conventional structure by several end-to-end speech synthesis systems [6, 7, 8, 9, 10]. The attention mechanism can jointly learn the alignments and the linguistic feature to acoustic feature mapping, so as to train and infer from the "text end" to the "speech end" as a whole.

The emergence of these seq2seq based speech synthesis models such as Tacotron [8] discard the independently designed text analyzer, and replace the traditional aligned linguistic frame to acoustic frame mapping networks by an encoder-decoder paradigm. The basic idea of Tacotron is to use recurrent neural networks (RNNs) to encode a text symbol sequence and then generate a variable length output acoustic feature sequence with a decoder RNNs. The encoder and decoder are connected through a soft attention mechanism [5]. Since such recurrent architecture relies on the entire past information during hidden state computations, fully convolutional neural networks (CNNs) based seq2seq framework was proposed to enable parallel training. Compared to RNNs, convolutions collect local contexts by kernels, and catch long-range dependencies by a large receptive field created by the stacking of CNN layers [11]. This kind of architecture is successfully used in speech synthesis area such as Deep Voice 3 [10] and DCTTS [9].

More recently, a more parallelizable architecture named transformer [12] was proposed to model global dependencies between input and output, while addressing the vanishing gradients problem of RNNs. The transformer is solely based on attention mechanisms to achieve seq2seq modeling, which requires minimum number of sequential operations during training. With self-attention, transformer can attend the whole sequence information at each position. Furthermore, it is proved that self-attention has a shorter path length between long-range dependencies than RNN and CNN, which make it easier to learn [12]. Almost at the same time of this work, the transformer architecture with CNN pre-net is successfully applied to speech synthesis task in [13], which shows a good

Work performed when Shan Yang was interning at Tencent AI Lab. Lei Xie is the corresponding author. The research work is supported by the National Key Research and Development Program of China (No.2017YFB1002102) and Tencent AI Lab Rhino-Bird Joint Research Program (No. JR201853).

capability of generating speech with self-attention.

However, there are obvious shortcomings in the self-attention mechanism. Firstly, [12] pointed out that there is no position guidance in self-attention. So we need to inject extra position encoding to handle the sequential information. Secondly, the weighted averaging operation in self-attention may lead to the dispersion of the distribution of attention, which results in overlooking the relation of neighboring signals [14]. Recent works have shown that self-attention networks benefit from locality modeling [15, 14]. According to our speech synthesis experiments, even by adding position encoding in the self-attention based TTS system, we can still find computer talking nonsense from time to time. And this issue is later fixed by introducing the relative-position-aware structure into self-attention [15]. At last, fully self-attention is proved not performing as good as RNNs when modeling underlying hierarchical structures in many tasks [16, 17].

In this paper, we introduce a hybrid architecture with relative-position-aware self-attention [15] to handle the above mentioned issues in speech synthesis. For the first two issues, we employ the relative-position-aware structure into self-attention to inject relative position relations among all positions. Thus it can enhance the local relations to avoid overlooking. For the third problem, inspired by [18], we investigate RNN and CNN based architectures with self-attention in the speech synthesis system, and then propose the multi-tower hybrid framework, which achieves the best Mean Opinion Score (MOS) among all systems.

The rest of the paper is organized as follows. Section 2 introduces the proposed hybrid architecture and its components in details. Section 3 introduces the experiments and the competing speech synthesis systems. We conclude this paper in Section 4.

2. PROPOSED HYBRID ARCHITECTURE WITH RELATIVE-POSITION-AWARE SELF-ATTENTION

Fig. 1 illustrates the architecture of the proposed hybrid architecture with relative-position-aware self-attention. It contains a mulit-tower hybrid encoder, an N-block self-attention based decoder, and a multi-head attention to connect the encoder and the decoder. Character embeddings are used as input to predict mel-scale spectrogram. WaveNet [19] conditioned on the mel-scale spectrogram is employed as vocoder to reconstruct audios. Details of the components are described below.

2.1. Multi-Head Attention Mechanism

Given a sequence as query Q, an attention function tends to output a weighted sum of the source sequence value V, where we use a key K as an index of V. Traditional attention mechanisms often conduct a single function among Q, K and V:

$$Attention(Q, K, V) = f(Q, K)V \tag{1}$$

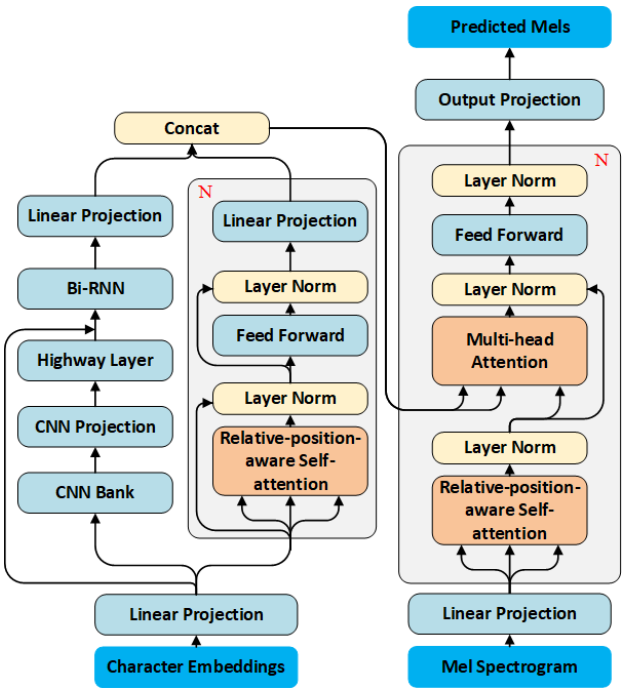

Fig. 1. System architecture.

where $f(Q, K)$ is a score function with softmax, such as Bahdanau [5] and Luong [20] score function. Here we use scaled dot-product score function [12]:

$$f(Q, K) = softmax(\frac{QK^T}{\sqrt{d_k}}) \tag{2}$$

where d_k is dimension of K.

Different from the single attention function, multi-head attention [12] tends to project Q, K and V into different subspace h times, where h is the number of heads. It allows the model to jointly learn from different representations of heads, which is proven to be beneficial for many tasks [18]. For each head $head_i$, the attention is computed as:

$$head_i = Attention(QW_i^Q, KW_i^K, VW_i^V) \tag{3}$$

where W_i^Q, W_i^K, W_i^V and W_i^O are different parameter matrices in linear projections. Then we can concatenate all the heads followed by an output projection to obtain the final attentions.

2.2. Self-Attention Representation

Self-attention is a special case of multi-head attention, where the Q and V are from a same sentence $x = (x_1, x_2, ..., x_n)$. We look into the Eq. (2) to see how self-attention works. Given x of n elements, we want to get a latent representation z with the same length n:

$$z_i = \sum_{j=1}^{n} \alpha_{ij}(x_j W^V) \tag{4}$$

where α_{ij} is the weight computed from the score function described in Eq. (2):

$$\alpha_{ij} = \frac{exp(e_{ij})}{\sum_{k=1}^{n} exp(e_{ik})} \tag{5}$$

$$e_{ij} = \frac{(x_i W^Q)(x_j W^K)^T}{\sqrt{d_k}} \tag{6}$$

Therefore, each latent z_i can assemble a global dependencies on the whole sequence x. Since there are no recurrence or convolution in self-attention, it's hard to model sequential information. To overcome this problem, transformer injects position information in the network inputs, where they use sine and cosine functions to encode positions [12].

2.3. Relative-Position-Aware Self-Attention

When modeling global dependencies, the self-attention ignores the distances between symbols or frames. However in speech synthesis, local contexts usually play a critical role in learning how people speak. Inspired by [15], we propose to add edges between elements x_i and x_j to enhance the local relations. Eq. (6) shows that the part $x_j W^K$ mainly decides the contribution of x_j for generating e_i given x_i. So we modify the Eq. (6) to additionally model the localness:

$$e_{ij} = \frac{(x_i W^Q)(x_j W^K + a_{ij}^K)^T}{\sqrt{d_k}} \tag{7}$$

where a_{ij}^K is the edge representation for matrix K, which strengthens the relative contribution of x_j for e_i. We can also inject another edge a_{ij}^V for V in Eq. (4). But we found it didn't bring improvements in our system. So we only use relative representations a^K.

In order to enhance the neighbouring relations, and to generate different sequence lengths not seen in training, maximum relative position is clipped to m in both directions. Then we get $2m+1$ unique relative representations. Each edge representation can be written as:

$$a_{ij}^K = \omega_{clip(j-i,m)}^K \tag{8}$$

$$clip(x, m) = max(-m, min(m, x)) \tag{9}$$

Fig. 2 shows an example of relative edges representation. The relative position representations are simply learned by $\omega^K = (\omega_{-m}^K, ..., \omega_m^K)$. Instead of fixed distances, learnable Gaussian bias can also be applied here to model localness, which may be softer and more suitable.

2.4. Hybrid Architecture

Though self-attention show its ability in modeling long-range dependencies [12], only using self-attention is still proved not as good as recurrent networks in modeling hierarchical structures [16, 17]. So we propose to use the multi-tower hybrid system to make use of both the advantages of the recurrent units and the self-attention mechanism.

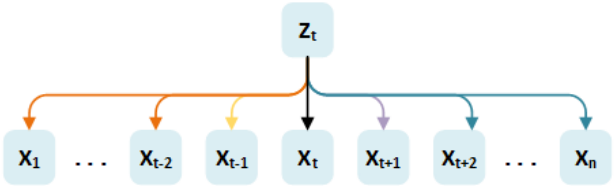

Fig. 2. Example relative edges representation with $m = 2$. When generating z_t, there are n additional edges a_{tj}^K, where different colors represent different vectors. For $j \leq t-2$ or $j \geq t+2$, the representation is clipped to the same ω_{-2}^K and ω_2^K, respectively.

3. EXPERIMENTS

3.1. Basic Setup

Our experiments are conducted on the public available English corpus from Blizzard Challenge 2011, which contains about 13 hours of speech of a single female speaker. For text, we simply feed character sequences as input. And 80-band mel-scale spectrogram is extracted from audios as the modeling target. For each system in our experiments, we trained about 500K steps (about two days) with a single Nvidia Tesla P40 GPU. With proper caching, the inference time is a little slower than Tacotron2 system.

To reconstruct audios, we train a WaveNet vocoder conditioned on the ground-truth mel spectrogram [19, 21, 22], which contains 30 dilated layers as described in [21]. For subjective evaluation, we conduct mean opinion score (MOS) tests to evaluate the naturalness of each system. 30 out of 100 synthesized utterances in test set are randomly chosen as test cases. And 35 listeners take part in the MOS test. All the systems share the same WaveNet vocoder for a fair comparison[1].

3.2. Model Architecture

For the hybrid encoder part, we concatenate RNNs and multi-head self-attention to catch salient behaviors of both models. We will also assess each individual model. In RNN encoder, we follow the encoder in Tacotron [8], which contains a convolutional bank and bidirectional recurrent units. In self-attention tower, we stack 6 self-attention blocks. Each attention block includes an 8-head self-attention and a position-wise feed-forward layer [12]. Residual connection and layer-norm are applied to these two layers. The projected 256-dimensional outputs from each of the two towers are finally concatenate as decoder attention input.

For the decoder part of the proposed system, there are also 6 sub-blocks. In order to communicate with the encoder, a 8-head attention is added in each block after the self-attention layer. Since attention layer considers both leftward and rightward contexts, which conflicts with the auto-regressive prop-

[1]Samples can be found at `https://syang1993.github.io/relation-aware/index.html`

Table 1. System components

Systems	Modules				
	CBHG Encoder	CNN Encoder [13]	Self Encoder	Position Encoding	Relation Aware
SELF-P			✓	✓	
SELF-R			✓		✓
CNN-P		✓		✓	
CNN-R		✓			✓
CBHG-P	✓			✓	
CBHG-R	✓				✓
Hybrid-P	✓		✓	✓	
Hybrid-R	✓		✓		✓

erty of decoding process, we add negative infinite bias before softmax to mask out illegal connections. Relative edge representation described in the previous section is added to all self-attention layers in both the encoder and the decoder. In that case, no position encoding or position embedding are used.

3.3. Subjective Evaluation

To evaluate the proposed methods, we built and compared 8 different systems. The details of each system are given in Table 1. CBHG encoder means CBHG module described in Tacotron. Self encoder represents a solely self-attention based encoder, and CNN encoder means that convolutional pre-net is added before self-attention blocks [13]. For all self-attention layers in both encoder and decoder, position encoding indicates absolute position representations [12], and relative-position-aware is the relative edge representations. We tested different clipping value m, and here we set $m = 10$. Fig. 3 shows the MOS scores of all systems.

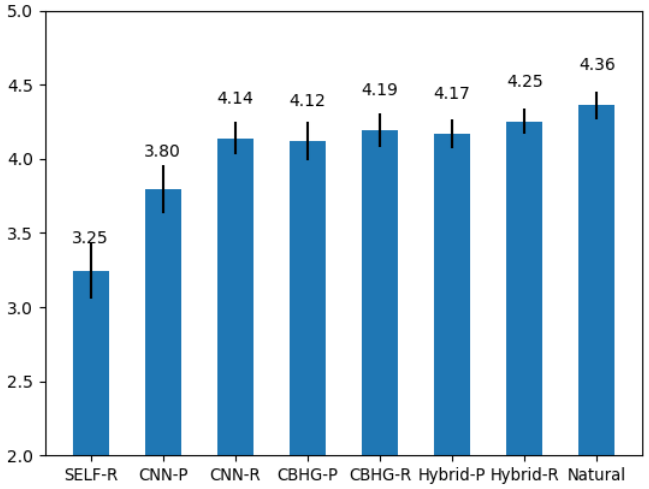

Fig. 3. MOS test results of various systems with 95% confidence intervals.

Firstly, we built a solely attention-based system SELF-P without any recurrent or convolutional unit. The sequential information was only injected from the position encoding. It is found that SELF-P cannot generate intelligible speech in inference, even after 1M training steps or using multi-GPU training like [13]. So we didn't put SELF-P in MOS evaluation. And there is no pattern in all the attention matrix, which shows that simply using absolute position encoding cannot learn good attention contributions of neighboring signals.

So we further proposed to conduct relative-position-aware self-attention to enhance the local contexts, as system SELF-R. We found that relative-position-aware mechanism can significantly improve the model performance. Unlike generating nonsense speech in SELF-P, the SELF-R model can generate normal and intelligible speech with the help of using relative edges, although it still suffers from the skip and mispronouncing problem. And like [14], the attention matrix of the lower three self-attention layers are nearly diagonal lines.

The performance of different individual towers were also evaluated as CNN-P and CBHG-P. CNN-P indicates the self-attention tower with CNN pre-nets and position encoding like [13]. And CBHG-P is the RNN tower equals to CBHG module used in Tacotron [8]. The results indicate that the only CBHG-based system outperforms the CNN + self-attention based system. More mispronouncing cases are observed in CNN-P than CBHG-P. Similar to Deep Voice 3 [10], replacing input sequences from English characters to phonemes sequences may alleviate this problem. Further with relative-position-aware representations, CNN-R significantly improves the performance of CNN-P, especially in mispronouncing and skip pronouncing. Noted that though CBHG-P only contains self-attention in decoder, CBHG-R also outperforms CBHG-P with enhanced relative representation.

The MOS results show that the proposed Hybrid-P outperforms all other individual systems without relation representations. We then get further improvements in system Hybrid-R with the edge representations. This result shows the effectiveness of combining both the advantages of recurrent architecture and self-attention in speech synthesis. And relative-position-aware edge connections can further strengthen the local relations over the basic position encoding.

4. CONCLUSION

We have proposed a hybrid self-attention structure which combines self-attention with recurrent networks for speech synthesis. We further enhance the proposed structure with relative-position-aware biases. From the MOS test results, we can conclude that the proposed hybrid structure outperforms any single CBHG or CNN self-attention structure. And by adding relative-position-aware representation into self-attention, all systems with absolute position encoding can be further improved in terms of the naturalness of synthesized speech. As the MOS test shows, the proposed relative-position-aware representation can enhance the hybrid self-attention speech synthesis system with the best MOS score which is only 0.11 below the natural recording.

5. REFERENCES

[1] Alan W Black, Heiga Zen, and Keiichi Tokuda, "Statistical parametric speech synthesis," in *Proc. ICASSP*. IEEE, 2007, pp. 1229–1232.

[2] Heiga Ze, Andrew Senior, and Mike Schuster, "Statistical parametric speech synthesis using deep neural networks," in *Proc. ICASSP*. IEEE, 2013, pp. 7962–7966.

[3] Zhen-Hua Ling, Shi-Yin Kang, Heiga Zen, Andrew Senior, Mike Schuster, Xiao-Jun Qian, Helen M Meng, and Li Deng, "Deep learning for acoustic modeling in parametric speech generation: A systematic review of existing techniques and future trends," *IEEE Signal Processing Magazine*, vol. 32, no. 3, pp. 35–52, 2015.

[4] Ilya Sutskever, Oriol Vinyals, and Quoc V Le, "Sequence to sequence learning with neural networks," in *Proc. NIPS*, 2014, pp. 3104–3112.

[5] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

[6] Wenfu Wang, Shuang Xu, and Bo Xu, "First Step Towards End-to-End Parametric TTS Synthesis: Generating Spectral Parameters with Neural Attention," in *Proc. INTERSPEECH*, 2016, pp. 2243–2247.

[7] Jose Sotelo, Soroush Mehri, Kundan Kumar, Joao Felipe Santos, Kyle Kastner, Aaron Courville, and Yoshua Bengio, "Char2wav: End-to-end speech synthesis," in *Proc. ICLR workshop*, 2017.

[8] Yuxuan Wang, RJ Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J Weiss, Navdeep Jaitly, Zongheng Yang, et al., "Tacotron: Towards end-to-end speech synthesis," in *Proc. INTERSPEECH*, 2017, pp. 4006–4010.

[9] Hideyuki Tachibana, Katsuya Uenoyama, and Shunsuke Aihara, "Efficiently Trainable Text-to-Speech System Based on Deep Convolutional Networks with Guided Attention," in *Proc. ICASSP*. IEEE, 2018, pp. 4784–4788.

[10] Wei Ping, Kainan Peng, Andrew Gibiansky, Sercan O Arik, Ajay Kannan, Sharan Narang, Jonathan Raiman, and John Miller, "Deep voice 3: 2000-speaker neural text-to-speech," *arXiv preprint arXiv:1710.07654*, 2017.

[11] Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N Dauphin, "Convolutional sequence to sequence learning," *arXiv preprint arXiv:1705.03122*, 2017.

[12] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," in *Proc. NIPS*, 2017, pp. 5998–6008.

[13] Naihan Li, Shujie Liu, Yanqing Liu, Sheng Zhao, Ming Liu, and Ming Zhou, "Close to Human Quality TTS with Transformer," *arXiv preprint arXiv:1809.08895*, 2018.

[14] Baosong Yang, Zhaopeng Tu, Derek F. Wong, Fandong Meng, et al., "Modeling Localness for Self-Attention Networks," in *Proc. EMNLP*, 2018.

[15] Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani, "Self-Attention with Relative Position Representations," *arXiv preprint arXiv:1803.02155*, 2018.

[16] Ke Tran, Arianna Bisazza, and Christof Monz, "The Importance of Being Recurrent for Modeling Hierarchical Structure," *arXiv preprint arXiv:1803.03585*, 2018.

[17] Mostafa Dehghani, Stephan Gouws, Oriol Vinyals, Jakob Uszkoreit, and Lukasz Kaiser, "Universal Transformers," *arXiv preprint arXiv:1807.03819*, 2018.

[18] Mia Xu Chen, Orhan Firat, Ankur Bapna, Melvin Johnson, Wolfgang Macherey, George Foster, Llion Jones, Niki Parmar, et al., "The Best of Both Worlds: Combining Recent Advances in Neural Machine Translation," *arXiv preprint arXiv:1804.09849*, 2018.

[19] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu, "Wavenet: A generative model for raw audio," *arXiv preprint arXiv:1609.03499*, 2016.

[20] Minh-Thang Luong, Hieu Pham, and Christopher D Manning, "Effective approaches to attention-based neural machine translation," *arXiv preprint arXiv:1508.04025*, 2015.

[21] Akira Tamamori, Tomoki Hayashi, Kazuhiro Kobayashi, Kazuya Takeda, and Tomoki Toda, "Speaker-dependent WaveNet vocoder," in *Proc. INTERSPEECH*, 2017, vol. 2017, pp. 1118–1122.

[22] Jonathan Shen, Ruoming Pang, Ron J Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, RJ Skerry-Ryan, et al., "Natural TTS synthesis by conditioning wavenet on mel spectrogram predictions," *arXiv preprint arXiv:1712.05884*, 2017.