CONFORMER-BASED SPEECH RECOGNITION WITH LINEAR NYSTRÖM ATTENTION AND ROTARY POSITION EMBEDDING

Lahiru Samarakoon, Tsun-Yat Leung

Fano Labs, Hong Kong

ABSTRACT

Self-attention has become an important component for end-to-end (E2E) automatic speech recognition (ASR). Recently, Convolution-augmented Transformer (Conformer) with relative positional encoding (RPE) achieved state-of-the-art performance. However, the computational and memory complexity of self-attention grows quadratically with the input sequence length. Effect of this can be significant for the Conformer encoder when processing longer sequences. In this work, we propose to replace self-attention with a linear complexity Nyström attention which is a low-rank approximation of the attention scores based on the Nyström method. In addition, we propose to use Rotary Position Embedding (RoPE) with Nyström attention since RPE is of quadratic complexity. Moreover, we show that models can be made even lighter by removing self-attention sub-layers from top encoder layers without any drop in the performance. Furthermore, we demonstrate that Convolutional sub-layers in Conformer can effectively recover the information lost due to the Nyström approximation.

***Index Terms*—** End-to-end Automatic Speech Recognition, Conformer, Nyströmformer, Rotary Position Embedding

1. INTRODUCTION

The attention has become a vital component of many machine learning tasks including E2E ASR systems [1, 2, 3]. First, it was used to combine the recurrent neural networks (RNNs) based encoders and decoders of E2E ASR models [1]. Recently, transformer [4] models have successfully replaced the RNN architectures for E2E ASR [2, 3]. Transformer architecture utilizes self-attention and inter-attention to completely remove the RNNs from models. This is mainly because self-attention is capable of relating any two positions of an input sequence while RNNs suffer in processing long-term dependencies. However, effectiveness of self-attention comes with a cost due to quadratic computational and memory complexity with respect to the sequence length.

Considerable amount of work has been done to make transformer models faster and lighter [5]. Some general methods that are applicable to virtually all neural networks that can also be employed to transformers are quantization, pruning, parameter sharing and knowledge distillation. In addition, there are some specific methods that directly aim at reducing the complexity of self-attention. For instance, Longformer [6], Reformer [7], Linformer [8] and Performer [9] are some of the major mechanisms that aim at reducing the complexity of self-attention. However, application of these methods are not properly explored for ASR. In this paper, we are investigating the applicability of Nyström approximation based linear attention to ASR tasks as proposed in Nyströmformer [10].

In E2E ASR, the encoder processes longer sequences compared to the decoder. This is because the encoder processes speech frames while the decoder processes text tokens. Therefore, in this work we explore the linear Nyström attention for self-attention sub-layers of the encoder. In addition, it has been shown that Conformer [11] can outperform Transformer encoders for ASR. Therefore, we use Conformer encoder layers in this work. Self-attention is good at modeling long-range global context, but suffers in extracting fine grained local information. Convolution neural networks (CNNs), on the other hand, are good at exploiting local information. Consequently, Conformer proposed to include a depth-wise CNN sub-layer after the self-attention sub-layer. In addition, a Conformer layer uses two feed-forward sub-layers to sandwich the self-attention and CNN sub-layers in Macaron style. Furthermore, Conformer showed that relative positional information is essential for good ASR performance. Hence, it is important to find a good position encoding mechanism that complements the Nyström attention.

In this paper, we investigate the effect of linear Nyström attention [10] for Conformer-based ASR. Nyström attention uses the Nyström method to come up with a low-rank approximation of the attention matrix. In addition, we propose to use Rotary Position Embedding (RoPE) as proposed in Roformer [12] with Nyström attention sub-layers as RoPE does not require the calculation of full attention matrix. In [13], authors showed that in ASR encoders, the diagonality of attention matrices increases from the lower to upper self-attention sub-layers. They further showed that upper self-attention sub-layers can be removed without any performance drop. Therefore, we also propose to remove self-attention sub-layers from upper encoder layers to make encoder more efficient. We report our evaluations on the conversational part of the National Speech Corpus (NSC) [14].

2. RELATED WORK

In this section, we review the self-attention mechanism and briefly discuss its more efficient variants. Then, we give a summary of positional encodings that are commonly used with ASR transformers.

2.1. Self-Attention

Let $\mathbf{X} = \{\boldsymbol{x}_1, ..., \boldsymbol{x}_n\} \in \mathbb{R}^{n \times d}$ be the input sequence to an encoder layer which contains n vectors of dimension d. In the self-attention, three matrices are computed by linearly projecting the layer's input as given below:

$$\mathbf{Q} = \mathbf{X}\mathbf{W}_Q, \mathbf{K} = \mathbf{X}\mathbf{W}_K, \mathbf{V} = \mathbf{X}\mathbf{W}_V \tag{1}$$

where $\mathbf{Q}, \mathbf{K}, \mathbf{V} \in \mathbb{R}^{n \times d}$ and $\mathbf{W}_Q, \mathbf{W}_K, \mathbf{W}_V \in \mathbb{R}^{d \times d}$. These three matrices are known as queries ($\mathbf{Q}$), keys ($\mathbf{K}$), and values ($\mathbf{V}$) and the self-attention computes the weighted sum of the values based on the

compatibility score between $\mathbf{Q}$ and $\mathbf{K}$ as given below:

$$\text{Attention}(\mathbf{Q},\mathbf{K},\mathbf{V}) = \text{Score}(\mathbf{Q},\mathbf{K})\mathbf{V} \quad (2)$$

$$= \text{Softmax}(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d}})\mathbf{V} \quad (3)$$

Even though transformer models use multiple attention heads [4], for the sake of clarity, we only explain the formulation for a single-head. One of the biggest concerns with transformer models is the $O(N^2)$ computation and memory complexity of self-attention. This is due to the $\mathbf{Q}\mathbf{K}^T$ calculation in Equation 3, where compatibility scores are computed for every pair of positions. A considerable amount of research effort has been allocated to reduce this complexity and some of major methods will be discussed in the next subsection.

2.2. Efficient Self-Attention Mechanisms

In this subsection, we will briefly summarize some of the prominent works aimed at reducing the complexity of self-attention. These efficient self-attention mechanisms can be categorized into Recurrence, Sparse Attention, Locality-sensitive hashing, Kernel attention and low-rank factorization [5].

Transformer-XL [15], proposed to split the input sequence into non-overlapping segments and introduced a recurrence between segments to leverage information from the previous segments. In combination with their RPE, they showed significant improvements in language modelling.

Sparse attention is a way to approximate the full attention with one or more variants of sparse attention matrices. In Sparse Transformer [16], full attention matrix is factorized into two sparse attention patterns called strided and fixed. Sparse Transformer achieved $O(N\sqrt{N})$ complexity on the input sequence length. A similar work has been applied to E2E ASR by [17] where the full attention matrix is approximated by two patterns called restricted and dilated self-attention. Dilated self-attention was aimed at capturing the long-context information while restricted self-attention captured local information.

Since the Softmax function is dominated by largest values, Reformer [7] employed locality-sensitive-hashing (LSH) to cluster queries and keys into buckets based on the similarity. By restricting the attention calculation only within each bucket Reformer achieved linear complexity.

Performer [9] approximated the attention by means of a kernel and achieved linear complexity. In [18], a Gaussian Kernel-based self-attention with frame indexing was used for connectionist temporal classification (CTC) [19] based ASR and reported performance improvements.

In Linformer [8], authors observed that the self-attention matrix is low-rank and proposed to first project each key to a lower dimension before calculating the dot product with the queries. If the projected dimension is significantly smaller than the sequence length, Linformer can achieve linear complexity.

2.3. Positional Encodings

Self-attention is order-invariant. Therefore it is necessary to inject positional information into Transformer models. Naturally, it makes sense to incorporate positional information to the pair-wise comparison of the self-attention. The attention score ($\mathbf{A}_{i,j}$) between i^{th} query and j^{th} key can be computed as given below:

$$\mathbf{A}_{i,j} = [f_q(\boldsymbol{x}_i\mathbf{W}_Q, i)][f_k(\boldsymbol{x}_j\mathbf{W}_K, j)]^T \quad (4)$$

where $f_{q,k}(.)$ is a function that incorporates respective positional information to the input features. In general, positional information is added to the features and can be decomposed as given below:

$$\mathbf{A}_{i,j} = \boldsymbol{x}_i\mathbf{W}_Q\mathbf{W}_K^T\boldsymbol{x}_j^T + \boldsymbol{x}_i\mathbf{W}_Q\mathbf{W}_K^T\boldsymbol{p}_j^T + \boldsymbol{p}_i\mathbf{W}_Q\mathbf{W}_K^T\boldsymbol{x}_j^T + \boldsymbol{p}_i\mathbf{W}_Q\mathbf{W}_K^T\boldsymbol{p}_j^T \quad (5)$$

Where $\boldsymbol{p}_j, \boldsymbol{p}_j \in \mathbb{R}^{1\times d}$ are absolute positional encodings (APE) for positions i, j. It is shown that incorporating relative positional information is more beneficial [20, 21]. Moreover, incorporating relative positional encoding (RPE) as in [15] reported significant gains for Conformer based ASR [11], which is given below:

$$\mathbf{A}_{i,j}^{rel} = \boldsymbol{x}_i\mathbf{W}_Q\mathbf{W}_{K,E}^T\boldsymbol{x}_j^T + \boldsymbol{x}_i\mathbf{W}_Q\mathbf{W}_{K,R}^T\boldsymbol{R}_{i-j}^T + \mathbf{u}^T\mathbf{W}_{K,E}^T\boldsymbol{x}_j + \boldsymbol{v}^T\mathbf{W}_{K,R}^T\boldsymbol{R}_{i-j}^T \quad (6)$$

where $\boldsymbol{R}_{i-j}$ is the RPE and $\mathbf{u}, \mathbf{v} \in \mathbb{R}^d$ are trainable parameters. In addition, the key projection matrix ($\mathbf{W}_K$) in Equation 5 is separated into $\mathbf{W}_{K,E}$ and $\mathbf{W}_{K,R}$.

3. PROPOSED METHOD

In this section, we first present the Nyström method based approximation for attention. Second, Rotary Position Embedding is presented. Then, we describe the removal of self-attention sub-layers from upper encoder layers for improved efficiency.

3.1. Nyström Attention

The Nyströmformer [10] proposed a Nyström approximation based method to calculate the attention scores. First it selects two subsets $\widetilde{\mathbf{Q}}$ and $\widetilde{\mathbf{K}}$ from $\mathbf{Q}$ and $\mathbf{K}$, respectively. These subsets are called landmarks, and are calculated by first creating non-overlapping chunks and then averaging over each chunk. Then it is possible to approximate the Softmax($\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d}}$) as given below:

$$\text{Softmax}(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d}}) \approx S_{\mathbf{Q},\widetilde{\mathbf{K}}}S^+_{\widetilde{\mathbf{Q}},\widetilde{\mathbf{K}}}S_{\widetilde{\mathbf{Q}}\mathbf{K}} \quad (7)$$

where $S_{\mathbf{A},\mathbf{B}} = \text{Softmax}(\frac{\mathbf{A}\mathbf{B}^T}{\sqrt{d}})$ for any $\mathbf{A}$ and $\mathbf{B}$ and $S^+_{\widetilde{\mathbf{Q}},\widetilde{\mathbf{K}}}$ is the Moore-Penrose inverse of $S_{\widetilde{\mathbf{Q}},\widetilde{\mathbf{K}}}$. Since the number of landmarks are constant and significantly smaller than the sequence length, Nyströmformer gives linear complexity $O(n)$. Since Nyström attention is a low-rank approximation, it is not straightforward or meaningful to incorporate RPE as mentioned in Equation 6. This is because RPE needs constructing the full $n \times n$ attention matrix between positions. Therefore, we propose to use Rotary position embedding (RoPE) [12] with Nyström Attention.

3.2. Rotary Position Embedding (RoPE)

Rotary Position Embedding (RoPE) [12] can be considered as a method that unifies both absolute and relative position encoding approaches. In RoPE, first a rotation based on absolute positional information is applied to queries and keys separately. Then the multiplicative interaction between rotated queries and rotated keys in the attention calculation implicitly infers the relative positional information. In contrast to RPE, RoPE does not require the computation of the full attention matrix. Therefore, RoPE can be readily

used with efficient self-attention calculations like Nyströmformer. The RoPE can be applied to both query and key as given below:

$$f_q(\boldsymbol{x}_m \mathbf{W}_Q, m) = R^d_{\Theta,m} \boldsymbol{x}_m \mathbf{W}_Q$$
$$f_k(\boldsymbol{x}_n \mathbf{W}_K, n) = R^d_{\Theta,n} \boldsymbol{x}_n \mathbf{W}_K \quad (8)$$

where

$$R^d_{\Theta,m} = \begin{bmatrix} M_1 & & & \\ & M_2 & & \\ & & \ddots & \\ & & & M_{d/2} \end{bmatrix} \quad (9)$$

$$M_r = \begin{bmatrix} \cos m\theta_r & -\sin m\theta_r \\ \sin m\theta_r & \cos m\theta_r \end{bmatrix} \quad (10)$$

and $\Theta = \{\theta_r = 10000^{-2(r-1)/d}, r \in [1, 2, .., d/2]\}$.

3.3. Removal of Self-attention Sub-layers from Upper Encoder Layers

In [13], authors investigated the usefulness of self-attention for upper encoder layers. They argued that acoustic events that are essential for ASR occur within short time spans in a left-to-right order. In addition, the learned context increases from the lower to upper self-attention sub-layers. This prompted the removal of self-attention sub-layers from top encoder layers without any performance degradations. They further showed that the diagonality of attention matrices increases from the lower to upper encoder self-attention sub-layers. Based on the same arguments, we propose the removal of uppermost self-attention sub-layers to improve the efficiency while maintaining the same accuracy.

4. EXPERIMENTAL SETUP

We conduct experiments on the Singaporean English National Speech Corpus (NSC) [14]. We only perform the experiments on the conversational speech data of the NSC corpus which is around 1000 hours. Data is pre-processed and split into Train, Dev and Test sets as done in the Espnet recipe [22]. In all our models, the encoder consists of 12 blocks while the decoder contains 6 blocks. Self-attention sub-layers consist of 512 hidden units and 8 attention heads while feed-forward layers have 2048 hidden units. We use kernel size of 31 for Conformer convolution. SpecAugment is employed [23]. Models are trained for 200 epochs and the last 10 best checkpoints are averaged to get the final model. We use Adam optimizer with 30000 warm-up steps. 20000 sub-word units are used as output tokens. Hybrid CTC/Attention architecture [24] with a CTC weight of 0.3 during training and CTC weight of 0.2 during decoding. While decoding a beam size of 10 and shallow fusion with a transformer-based language model (LM) with a LM weight of 0.2 is used. For all the models trained in this paper, decoder configuration remains the same.

5. RESULTS

In Table 1, both RPE and RoPE improve the performance of the Transformer and Conformer models compared to APE, showing that the relative positional information is more beneficial. For all models, RPE gives slightly better results than the RoPE. In addition, applying RoPE to value representations, i.e. RoPE(V), does not give any improvement. This means that the multiplicative interaction between rotated queries and keys for implicit computation of the relative positional information is important. Furthermore, RPE performs slightly better than the RoPE for longer utterances as shown in Figure 1.

Table 1. Word error rate (WER) % for Transformer and Conformer models with various positional encodings. In RoPE(V), RoPE is also applied to the value representations.

Model	APE	RPE	RoPE	RoPE(V)	WER(%)	
					Dev	Test
Transformer	Y				21.2	22.3
		Y			20.2	21.2
			Y		20.3	21.5
				Y	21.8	22.8
Conformer	Y				21.4	23.7
		Y			19.5	20.9
			Y		20.1	21.3
				Y	20.7	21.3

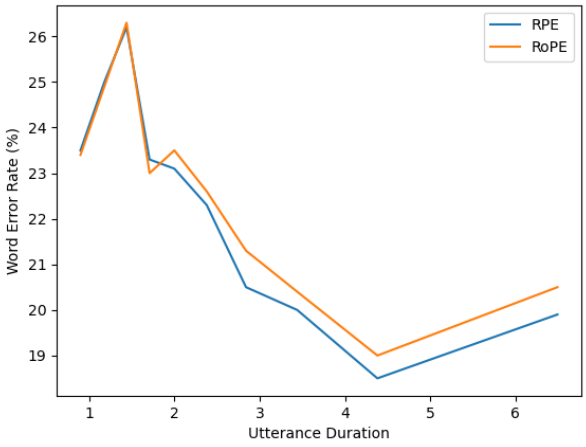

Fig. 1. WER % v.s. Utterance Duration with RPE and RoPE

Next, to investigate the performance of Nyström attention, we replace the self-attention sub-layers of the Conformer encoder with Nyström attention. As shown in Table 2, we train Nyströmformers with 24 landmarks with both APE and RoPE. 24 landmarks is a reasonable number as our encoders process sequences of around 100 frames on average after subsampling. Table 2 shows that RoPE also improves the performance of the Nyströmformer encoders significantly. The Nyströmformer also achieves a better performance than Conformer with RoPE on the test set, which suggests that Nyström attention with RoPE can be an effective way to replace self-attention

Table 2. WER % for Nyströmformer and Conformer models with various positional encodings.

Model	APE	RoPE	WER(%)	
			Dev	Test
Nyströmformer (24 landmarks)	Y		20.8	22.4
		Y	20.6	20.9
Conformer	Y		21.4	23.7
		Y	20.1	21.3

with RoPE. For the rest of this paper, we train all our Nyström attention sub-layers with RoPE.

Table 3. WER % for Conformer + RoPE model with various number of encoder and feed-forward(FF) layers.

Num. of Conformer Encoders	Num. of FF	WER(%)	
		Dev	Test
12	0	20.1	21.3
11	1	20.1	20.9
10	2	20.2	21.3
9	3	20.5	21.7

As mentioned in Section 3.3, based on the diagonality of the upper self-attention sub-layers, it is possible to remove them without a drop in performance. The results for that is shown in Table 3. Replacing one top self-attention sub-layers gives the best result on the test set. It supports the observation and results from [13].

Table 4. WER % for various number of Landmarks.

Num. of Nyström. Encoder	Num. of FF	Num. of Landmarks	WER(%)	
			Dev	Test
12	0	16	20.9	21.5
		24	20.6	20.9
		32	20.4	21.0
11	1	16	20.6	21.0
		24	20.5	21.2

Table 4 shows the WER for the Nyströmformer with different landmarks. It shows that a small number of landmarks, e.g. 16, is sufficient to ensure a good approximation on the self-attention matrix for ASR task.

Table 5. WER % for various number of Nyströmformer (24 landmarks) encoder, Conformer encoder and FF layers.

Num. of Preceding Nyström. Encoders	Num. of Conformer Encoders	Num. of FF	WER(%)	
			Dev	Test
0	11	1	20.1	20.9
11	0	1	20.5	21.2
1	10	1	19.9	20.9
2	9	1	20.3	21.0
3	8	1	19.9	21.1
4	7	1	20.3	20.9

In [13] showed that lower self-attention sub-layers have low diagonality, This suggests that lower self-attention sub-layers may be more robust to Nyströmformer landmark selection and approximation. We examine the effect of replacing lower Conformer encoder layers with Nyströmformer encoder layers, as shown in Table 5. Although using the approximation of the attention matrix is expected to lose local information, Nyströmformer does not significantly degrade the performance with a various number of preceding Nyströmformer encoder. One possible reason is that CNN sub-layer in the conformer blocks with skip connections helps recover the local information, which is examined in next.

The effect of CNN sub-layer in the Nyströmformer are shown in Table 6. Compared to the Transformer with RoPE, Nyströmformer without CNN sub-layer gives a significantly worse WER on the test set. It increases the absolute WER by 2.2% on test set. As the major difference between the Transformer and Nyströmformer without CNN is the attention, it means that the approximation of the attention matrix degrades the performance. However, when the Nyströmformer includes CNN sub-layers, performance improves significantly. CNN sub-layer alleviates the error from the approximation and landmark selection, possibly due to CNN with skip connections can effectively aggregate and recover local information before and after the approximation. This shows that CNN sub-layer is important when using Nyström attention.

Table 6. WER % for Transformer and Nystrom Attention Comparison.

Model	CNN Sub-layer	WER(%)	
		Dev	Test
Transformer (RoPE)	N.A.	20.3	21.5
Nyströmformer (RoPE, 16 landmarks)	No	22.1	23.7
	Yes	20.9	21.5

Although Nyström attention has linear complexity, we have not observed significant real time factor (RTF) improvements over the Conformer ASR models. We think this is mainly due processing of shorter utterances after subsampling. In order to verify this conjecture, we conducted speed improvement experiments for E2E diarization [25, 26, 27] using an audio segment of length 235 seconds subsampled by a factor of 10. We found that Nyströmformer can give speed improvements upto 2.57 times over a Conformer baseline. In our future work, we intend to extend our work to long-form ASR [28, 29] and E2E diarization.

6. CONCLUSIONS

In this paper, we proposed to use linear Nyström attention for end-to-end (E2E) automatic speech recognition (ASR) to replace the self-attention which has a quadratic computational and memory complexity with the input sequence length. In addition, we proposed to use Rotary Position Embedding (RoPE) with Nyström attention which reported significant performance improvement over absolute positional encoding (APE). Moreover, we showed that Conformer encoders can be made even lighter by removing self-attention sub-layers from top encoder layers without any drop in the performance. Furthermore, we demonstrated that Convolutional sub-layer alleviates the error from the Nyström approximation, possibly with skip connections can effectively aggregate and recover local information before and after the Nyström attention sub-layer. As future work, we will extend this work to long-form ASR and E2E diarization where models process longer speech sequences.

7. REFERENCES

[1] William Chan, Navdeep Jaitly, Quoc V Le, and Oriol Vinyals, "Listen, attend and spell," *arXiv preprint arXiv:1508.01211*, 2015.

[2] Linhao Dong, Shuang Xu, and Bo Xu, "Speech-transformer: a no-recurrence sequence-to-sequence model for speech recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5884–5888.

[3] Albert Zeyer, Parnia Bahar, Kazuki Irie, Ralf Schlüter, and Hermann Ney, "A comparison of transformer and lstm encoder decoder models for asr," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2019, pp. 8–15.

[4] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[5] Quentin Fournier, Gaétan Marceau Caron, and Daniel Aloise, "A practical survey on faster and lighter transformers," *arXiv preprint arXiv:2103.14636*, 2021.

[6] Iz Beltagy, Matthew E Peters, and Arman Cohan, "Longformer: The long-document transformer," *arXiv preprint arXiv:2004.05150*, 2020.

[7] Nikita Kitaev, Lukasz Kaiser, and Anselm Levskaya, "Reformer: The efficient transformer," *arXiv preprint arXiv:2001.04451*, 2020.

[8] Sinong Wang, Belinda Z Li, Madian Khabsa, Han Fang, and Hao Ma, "Linformer: Self-attention with linear complexity," *arXiv preprint arXiv:2006.04768*, 2020.

[9] Krzysztof Choromanski, Valerii Likhosherstov, David Dohan, Xingyou Song, Andreea Gane, Tamas Sarlos, Peter Hawkins, Jared Davis, Afroz Mohiuddin, Lukasz Kaiser, et al., "Rethinking attention with performers," *arXiv preprint arXiv:2009.14794*, 2020.

[10] Yunyang Xiong, Zhanpeng Zeng, Rudrasis Chakraborty, Mingxing Tan, Glenn Fung, Yin Li, and Vikas Singh, "Nystromformer: A nystrom-based algorithm for approximating self-attention," *arXiv preprint arXiv:2102.03902*, 2021.

[11] Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, et al., "Conformer: Convolution-augmented transformer for speech recognition," *arXiv preprint arXiv:2005.08100*, 2020.

[12] Jianlin Su, Yu Lu, Shengfeng Pan, Bo Wen, and Yunfeng Liu, "Roformer: Enhanced transformer with rotary position embedding," *arXiv preprint arXiv:2104.09864*, 2021.

[13] Shucong Zhang, Erfan Loweimi, Peter Bell, and Steve Renals, "On the usefulness of self-attention for automatic speech recognition with transformers," in *2021 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2021, pp. 89–96.

[14] Jia Xin Koh, Aqilah Mislan, Kevin Khoo, Brian Ang, Wilson Ang, and Charmaine Ng, "Building the singapore english national speech corpus," *Malay*, vol. 20, no. 25.0, pp. 19–3, 2019.

[15] Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc V Le, and Ruslan Salakhutdinov, "Transformer-xl: Attentive language models beyond a fixed-length context," *arXiv preprint arXiv:1901.02860*, 2019.

[16] Rewon Child, Scott Gray, Alec Radford, and Ilya Sutskever, "Generating long sequences with sparse transformers," *arXiv preprint arXiv:1904.10509*, 2019.

[17] Niko Moritz, Takaaki Hori, and Jonathan Le Roux, "Capturing multi-resolution context by dilated self-attention," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 5869–5873.

[18] Yosuke Kashiwagi, Emiru Tsunoo, and Shinji Watanabe, "Gaussian kernelized self-attention for long sequence data and its application to ctc-based speech recognition," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 6214–6218.

[19] Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *Proceedings of the 23rd international conference on Machine learning*, 2006, pp. 369–376.

[20] Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani, "Self-attention with relative position representations," *arXiv preprint arXiv:1803.02155*, 2018.

[21] Cheng-Zhi Anna Huang, Ashish Vaswani, Jakob Uszkoreit, Noam Shazeer, Ian Simon Curtis Hawthorne, Andrew M Dai, Matthew D Hoffman, Monica Dinculescu, and Douglas Eck, "Music transformer: Generating music with long-term structure generating music with long-term structure, december 2018," *arXiv preprint arXiv:1809.04281*.

[22] Shinji Watanabe, Takaaki Hori, Shigeki Karita, Tomoki Hayashi, Jiro Nishitoba, Yuya Unno, Nelson Enrique Yalta Soplin, Jahn Heymann, Matthew Wiesner, Nanxin Chen, Adithya Renduchintala, and Tsubasa Ochiai, "ESPnet: End-to-end speech processing toolkit," in *Proceedings of Interspeech*, 2018, pp. 2207–2211.

[23] Daniel S Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D Cubuk, and Quoc V Le, "Specaugment: A simple data augmentation method for automatic speech recognition," *arXiv preprint arXiv:1904.08779*, 2019.

[24] Shinji Watanabe, Takaaki Hori, Suyoun Kim, John R Hershey, and Tomoki Hayashi, "Hybrid ctc/attention architecture for end-to-end speech recognition," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 8, pp. 1240–1253, 2017.

[25] Yusuke Fujita, Naoyuki Kanda, Shota Horiguchi, Yawen Xue, Kenji Nagamatsu, and Shinji Watanabe, "End-to-end neural speaker diarization with self-attention," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2019, pp. 296–303.

[26] Shota Horiguchi, Yusuke Fujita, Shinji Watanabe, Yawen Xue, and Kenji Nagamatsu, "End-to-end speaker diarization for an unknown number of speakers with encoder-decoder based attractors," *Proc. Interspeech 2020*, 2020.

[27] Tsun-Yat Leung and Lahiru Samarakoon, "Robust end-to-end speaker diarization with conformer and additive margin penalty," *Proc. Interspeech 2021*, pp. 3575–3579, 2021.

[28] Chung-Cheng Chiu, Wei Han, Yu Zhang, Ruoming Pang, Sergey Kishchenko, Patrick Nguyen, Arun Narayanan, Hank Liao, Shuyuan Zhang, Anjuli Kannan, et al., "A comparison of end-to-end models for long-form speech recognition," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2019, pp. 889–896.

[29] Arun Narayanan, Rohit Prabhavalkar, Chung-Cheng Chiu, David Rybach, Tara N Sainath, and Trevor Strohman, "Recognizing long-form speech using streaming end-to-end models," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2019, pp. 920–927.