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

Self-and-Mixed Attention Decoder with Deep Acoustic Structure for Transformer-based LVCSR

Xinyuan Zhou[1,2]*,Grandee Lee*[2]*, Emre Yılmaz*[2]*, Yanhua Long*[1]*, Jiaen Liang*[3]*, Haizhou Li*[2]

[1]Shanghai Normal University, Shanghai, China
[2]National University of Singapore, Singapore
[3]Unisound AI Technology Co., Ltd., Beijing, China

{xinyuan_zhou,grandee.lee}@u.nus.edu, emrey@kth.se, yanhua@shnu.edu.cn,
liangjiaen@unisound.com, haizhou.li@nus.edu.sg

Abstract

Transformer has shown impressive performance in automatic speech recognition. It uses an encoder-decoder structure with self-attention to learn the relationship between high-level representation of source inputs and embedding of target outputs. In this paper, we propose a novel decoder structure that features a self-and-mixed attention decoder (SMAD) with a deep acoustic structure (DAS) to improve the acoustic representation of Transformer-based LVCSR. Specifically, we introduce a self-attention mechanism to learn a multi-layer deep acoustic structure for multiple levels of acoustic abstraction. We also design a mixed attention mechanism that learns the alignment between different levels of acoustic abstraction and its corresponding linguistic information simultaneously in a shared embedding space. The ASR experiments on Aishell-1 show that the proposed structure achieves CERs of 4.8% on the dev set and 5.1% on the test set, which are the best reported results on this task to the best of our knowledge.

Index Terms: speech recognition, attention, Transformer

1. Introduction

The sequence-to-sequence (S2S) approach [1] has achieved remarkable results in automatic speech recognition (ASR), in particular, large vocabulary continuous speech recognition (LVCSR) [2–8]. Unlike conventional hybrid ASR, S2S requires neither lexicons, prerequisite models, nor decision trees. It optimizes the acoustic and language model jointly and simultaneously, learning the mapping directly from speech to text.

The most commonly used structure in S2S approaches is the attention-based encoder-decoder model (AED) [9]. This model maps the input feature sequences to the output character sequences and has been widely used in ASR tasks [7, 10, 11]. Among them, the Listen, Attend and Spell (LAS) [12] structure has shown superior performance to a conventional hybrid system using large amounts of training data. LAS uses an encoder that is a pyramidal recurrent neural network (RNN) to convert low-level speech signals into higher-level acoustic representations, and then the relationship between these representations and targets is learned by an attention mechanism at the RNN-based decoder. However, due to the sequential nature of RNNs, the LAS model doesn't support parallelization of calculations, therefore, is prevented from big data training.

To remedy this problem, new encoder-decoder structures with self-attention networks have recently been proposed [13]. With self-attention, these structures can not only effectively capture global interaction between sequences [14], learning the direct dependence of long-distance sequences [15], but also support parallelized model training [13]. Now, these structures are widely used in a variety of machine learning tasks providing significant improvements. Vaswani et al. [13] first proposed a S2S self-attention based model called the Transformer, and it achieves state-of-the-art performance on WMT2014 English-to-French translation task with remarkable lower training cost. For the ASR task, Transformer also uses the AED structure. Unlike LAS, the Transformer uses the multi-head self-attention (MHA) sub-layer to learn the source-target relationship, and capture the mutual information within the sequences to extract the most effective high-level features. This enables Transformer-based ASR systems to achieve competitive performance over the conventional hybrid and other end-to-end approaches [11, 16–21].

Inspired by the Transformer and the layer-wise coordination [22], we propose a novel decoder structure that features a self-and-mixed attention decoder (SMAD) with a deep acoustic structure (DAS) to improve the acoustic representation of Transformer-based LVCSR. With reference to the standard Speech-Transformer in [11], several improvements have been made at the decoder.

In the Speech-Transformer decoder, the linguistic information is first extracted using a self-attention sub-layer, and then processed together with the encoder output in another source-target attention sub-layer. The same encoder output is repeatedly taken by every decoder layer to establish the acoustic-target relationship. In this paper, we propose a new attention block, called self-and-mixed attention (SMA), as an unified attention sub-layer in the decoder, that takes the concatenation of the encoded acoustic representation and the word label embedding as input. In this way, the acoustic and linguistic information is projected into the same subspace in the deep decoder network structure during the attention calculation.

Furthermore, our decoder learns the acoustic and linguistic information together in a layer-by-layer fashion with the SMA mechanism instead of repeatedly using the same acoustic representations in each decoder layer. This is motivated by two intuitions, 1) we hope to benefit from a deep decoder network structure that encodes multi-level of abstraction from both acous-

This work was done when Xinyuan Zhou was an intern at National University of Singapore. Yanhua Long is the corresponding author. The work is supported by the National Natural Science Foundation of China (No.61701306), and Human-Robot Interaction Phase 1 (Grant No. 192 25 00054), National Research Foundation (NRF) Singapore under the National Robotics Programme; AI Speech Lab (Award No. AISG-100E-2018-006), NRF Singapore under the AI Singapore Programme; Human Robot Collaborative AI for AME (Grant No. A18A2b0046), NRF Singapore.

http://dx.doi.org/10.21437/Interspeech.2020-2556

tic and linguistic representation, and 2) we hypothesize that a shared acoustic and linguistic embedding space will help the network to learn the association between acoustic and linguistic information, and improve their alignments.

We will introduce the Transformer-based ASR as the prior work in Section 2, and discuss the details of the new decoder in Section 3.

2. Transformer-based ASR

2.1. Encoder-Decoder with Attention

The Transformer model [13] uses an encoder-decoder structure similar to many other neural sequence transduction models. The encoder can be regarded as a feature extractor, which converts the input vector $\boldsymbol{x}$ into a high-level representation $\boldsymbol{h}$. Given $\boldsymbol{h}$, the decoder generates prediction sequence $\boldsymbol{y}$ one token at a time in an auto-regressive manner. In an ASR task [11, 23], tokens are usually modeling units, such as phones, characters or subword, etc.

The encoder has N layers, each of which contains two sub-layers: a multi-head self-attention and a position-wise fully connected feed-forward network (FFN). Similar to the encoder, the decoder is also composed of a stack of M identical layers. In addition to the two sub-layers, each layer of decoder also has a third sub-layer between the FFN and MHA to perform multi-head source-target attention over the output representation of the encoder stack.

2.2. Multi-Head Attention

Multi-head attention is the core module of the Transformer model. Unlike single-head attention, MHA can learn the relationship between queries, keys and values from different subspaces. It computes the "Scaled Dot-Product Attention" with the following form:

$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V, \quad (1)$$

where $Q \in \mathbb{R}^{t_q \times d_q}$ is the query, $K \in \mathbb{R}^{t_k \times d_k}$ is the key and $V \in \mathbb{R}^{t_v \times d_v}$ is the value. t_* are the length of input and d_* are the dimension of corresponding elements. To prevent pushing the softmax into extremely small gradient regions caused by large dimensions, the $\frac{1}{\sqrt{d_k}}$ is used to scale the dot products.

In order to calculate attention from multiple subspaces, multi-head attention is constructed as follow:

$$\text{MHA}(Q, K, V) = \text{Concat}(Head_1, \cdots, Head_H)W^O, \quad (2)$$

$$Head_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V), \quad (3)$$

where W_i^* is the projection matrix. $W_i^Q \in \mathbb{R}^{d_{\text{model}} \times d_Q}$, $W_i^K \in \mathbb{R}^{d_{\text{model}} \times d_K}$, $W_i^V \in \mathbb{R}^{d_{\text{model}} \times d_V}$ and $W^O \in \mathbb{R}^{d_{\text{model}} \times d_O}$, d_{model} is the dimension of the input vector to the encoder, H is the number of heads. For each Q, K, V in each attention, they are projected to d_* dimensions through three linear projection layers W_i^* respectively. After performing H attentions, the outputs are then concatenated and projected again to obtain the final values.

2.3. Positional Encoding

Unlike RNN, the MHA contains no recurrence and convolution, it cannot model the order of the input acoustic sequence. We follow the idea of "positional encoding" that is added to the input as described in [13].

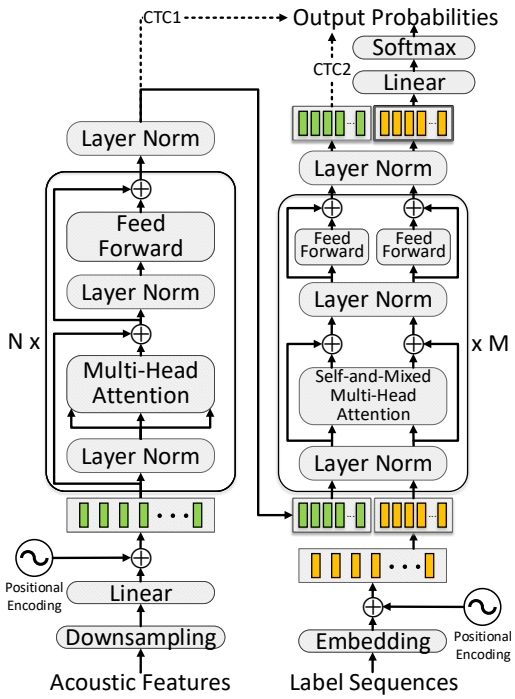

Figure 1: *The system architecture of the Transformer-based ASR with the proposed self-and-mixed attention decoder.*

3. Self-and-Mixed Attention Decoder

3.1. Architecture

Figure 1 shows an encoder-decoder architecture for ASR. We adopt the same encoder as in Transformer, and propose a Self-and-Mixed Attention Decoder (SMAD), that is an attention-based auto-regressive structure. The main difference between SMAD and the standard Transformer decoder [11] lies in the ways we integrate acoustic representations, $\boldsymbol{h}$.

Firstly, unlike the decoder in Transformer which takes the same $\boldsymbol{h}$ repeatedly to every decoder layer, SMAD employs a deep acoustic structure (DAS), which is a M-layer network to capture multiple level of acoustic abstraction. For simplicity, we use a single-head self-and-mixed attention in Figure 2 to illustrate the Self-and-Mixed MHA component in the decoder layer of Figure 1, where the self-attention handles the acoustic representations, and the mixed-attention handles the acoustic-target alignment. With M decoder layers stacking in a serial pipeline, the flow of acoustic information in Figure 2 (green) forms a deep acoustic structure.

Secondly, the decoder in the standard Transformer uses a self-attention module to learn the current target representation based on the previous tokens and learn the acoustic-targets dependencies using another separate source-target attention sub-layer. However, in SMAD, we merge these two attentions into one as illustrated in Figure 2. We concatenate the encoded acoustic representation and linguistic targets to form a joint embedding as the input to the decoder layer. After the self-and-mixed MHA, the concatenated representation with both acoustic and linguistic information is fed to the FFN and next self-and-mixed MHA. Since the information flow in the proposed decoder contains two modalities, we also employ modality-specific residual connections and position-wise feed forward networks to separate linguistic and acoustic information before

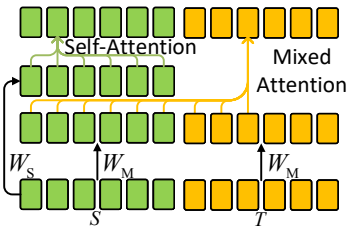

Figure 2: *A single-head self-and-mixed attention mechanism as a sub-layer of the decoder in Figure 1.*

obtaining the posterior probability at the output of the softmax layer preceded by a linear layer.

We perform the same downsampling as in [23] before the encoder using two 3×3 CNN layers with stride 2 to reduce the GPU memory occupation and and the length of the input sequence.

3.2. Self-and-Mixed Attention (SMA)

For simplicity, we take one head of the self-and-mixed MHA in Figure 1 as an example. The SMA consists of two independent attention mechanisms: a self-attention for acoustic-only representation, and a mixed attention to learn linguistic representation and the acoustic-target association. As shown in Figure 2, S refers to source, which is the acoustic representation (marked in green) and T refers to target, which is the linguistic information (marked in yellow).

Specifically, for self-attention in the SMA, Q, K, $V \in \mathbb{R}^{n \times d_{\text{model}}}$ are projected by $W_{\text{S}}^{Q}, W_{\text{S}}^{K}, W_{\text{S}}^{V}$ from S respectively, with n-length acoustic representation. The acoustic hidden representation in the current layer is generated using the accumulated acoustic information in the previous layer using the self-attention mechanism.

For the mixed attention, a linguistic token in the current layer is generated using the acoustic hidden representation and the preceding linguistic tokens in the previous layer using a mixed attention mechanism. The mixed attention is formulated as follows:

$$\text{MixedAtt}(Q,K,V) = \text{softmax}(\frac{QK^T}{\sqrt{d_{\text{model}}}} + Mask)V, \quad (4)$$

$$Q = TW_{\text{M}}^{Q}, \quad (5)$$

$$K = \text{Concat}(S,T)W_{\text{M}}^{K}, \quad (6)$$

$$V = \text{Concat}(S,T)W_{\text{M}}^{V}, \quad (7)$$

$$Mask(i,j) = \begin{cases} -\infty, & j > i+n \\ 0, & \text{otherwise} \end{cases}, \quad (8)$$

where $Q \in \mathbb{R}^{m \times d_{\text{model}}}$, K, $V \in \mathbb{R}^{(n+m) \times d_{\text{model}}}$, and $Mask \in \mathbb{R}^{m \times (n+m)}$ is the mask matrix, m is the number of tokens, i and j refer to the index of row and column of $Mask$. To project the acoustic and linguistic information into the same subspace, we concatenate S and T and apply the same projection matrix W_{M} for K and V. We use the acoustic representation $\boldsymbol{h}$ of entire sentence for the decoding of tokens, at the same time, we introduce a $Mask$ to ensure that the prediction of token sequence is causal, i.e., when predicting a token, we only use information of the tokens before it. When $Mask(i,j)$ equals to $-\infty$, the corresponding position in softmax output will approach zero, which prevents position i from attending to position j.

3.3. Multi-Ojective Learning

As often done in encoder-decoder structures, our model also uses the connectionist temporal classification (CTC) [23–25] to benefit from the monotonic alignment. The CTC loss function [24] is used to jointly train the extraction of acoustic representation by Multi-Objective Learning (MOL):

$$\mathcal{L}_{\text{MOL}} = \lambda \log P_{\text{ctc}}(C|X) + (1-\lambda)\log P_{\text{att}}^{*}(C|X), \quad (9)$$

$$P_{\text{att}}^{*}(C|X) = \prod_{l=1}^{L} P(c_l|c_1^{*}, \cdots, c_{l-1}^{*}, X), \quad (10)$$

$$P_{\text{ctc}}(C|X) = \sum_{Z} \prod_{t=1}^{T} P(z_t|z_{t-1}, C)P(z_t|X), \quad (11)$$

$$P(z_t|X) = \text{Softmax}(\text{LinB}(\mathbf{h}_t)), \quad (12)$$

where $\mathcal{L}_{\text{MOL}}$ is the multi-objective loss with a tuning parameter $\lambda \in [0,1]$, $P_{\text{att}}^{*}(C|X)$ is the Transformer loss modeled by Kullback-Leibler divergence [26] loss . $X = \{x_t \in \mathbb{R}^D | t = 1, \cdots, T\}$ is a T-length speech feature sequence, x_t is a D-dimensional speech feature vector at frame t, $C = \{c_t \in \mathcal{U} | l = 1, \cdots, L\}$ is an L-length letter sequence containing all the characters $\mathcal{U}$ in this task, and c_{l-1}^{*} is the ground truth of c_l^{*}'s previous token. $Z = \{z_t \in \mathcal{U} \cup \langle \text{b} \rangle | t = 1, \cdots, T\}$ is a framewise letter sequence with an additional blank symbol "$\langle b \rangle$", and $\mathbf{h}_t$ is the acoustic hidden representation vector, LinB$(\cdot)$is a linear layer to convert $\mathbf{h}_t$ to a $(|\mathcal{U}|+1)$ dimensional vector.

We explore two different locations where the CTC loss can be applied as shown in Figure 1. For CTC1,

$$\mathbf{h}_t = \text{Encoder}_t(X), \quad (13)$$

Encoder$_t(\cdot)$ accepts the full feature sequence X and output acoustic representation $\mathbf{h}_t$ at t. Similar to the previous techniques, this CTC loss is used to jointly train the encoder.

For CTC2, since the acoustic representation is also updated in the SMAD, $\mathbf{h}_t$ is produced as follows:

$$\mathbf{h}_t = \text{Decoder}_t^{A}(\text{Encoder}(X)), \quad (14)$$

where the Decoder$_t^{A}(\cdot)$ is the acoustic side of decoder stack output at t. In this way, the entire acoustic representation extraction process is jointly trained using the CTC loss.

4. Experiments

4.1. Experimental setup

We conduct experiments on 170 hours Aishell-1 [27] using the ESPnet [28] end-to-end speech processing toolkit. For all experiments, we extract 80-dimensional log Mel-filter bank plus pitch and its Δ, $\Delta\Delta$ as acoustic features and normalize them with global mean computed from the training set. The frame-length is 25 ms with a 10 ms shift.

The standard configuration of the state-of-the-art ESPnet Transformer recipe on Aishell-1 is used for both the baseline and proposed model. Each model contains 12-layer encoder and 6-layer decoder, where the $d_{\text{model}} = 256$ and the dimensionality of inner-layer in FFN $d_{\text{ff}} = 2,048$. In all attention sub-layers, 4 heads are used for MHA. The whole network is trained for 50 epochs and warmup [13] is used for the first 25,000 iterations. We use 4,230 Chinese characters which are extracted from the train set as modeling unit. A ten-hypotheses-width beam search is used with the the one-pass decoding for CTC as

Table 1: *Results comparsion on Aishell-1 in CER%*

System	Dev	Test
Kaldi (chain)	-	7.5
Kaldi (nnet3)	-	8.6
LAS [34]	-	10.6
ESPnet RNN [36]	6.8	8.0
RNN-T [35]	10.1	11.8
Transformer +SP+CTC (baseline) [36]	6.0	6.7
T-SMAD +SP+CTC1	5.9	6.4
T-SMAD +SP+CTC2	5.4	6.0

described in [24] and a two-layer RNN language model (LM) shallow fusion [29, 30], which was trained on the training transcriptions of Aishell-1 with 4,230 Chinese characters. We also evaluate the effect of speed perturbation (SP) [31], SpecAugment (SpecA) [32] and CTC joint training in our experiments.

4.2. Results and Discussion

Table 1 reports the results of the proposed Transformer-based ASR, referred to as T-SMAD, the conventional Kaldi hybrid [33] and other E2E ASR systems. Shallow fusion with 5-gram language model is used in both [34, 35]. In ESPnet RNN, Transformer [36] and T-SMAD, the RNN LM was also used for shallow fusion. We consider the Transformer with speed perturbation and CTC in ESPnet as our reference baseline.

According to Table 1, the T-SMAD system with the proposed CTC2 outperforms all other systems, including both the Transformer baseline and the Kaldi hybrid systems. A relative 20.0% CER reduction on the test set is obtained over the best hybrid system (chain). A relative 10% CER reduction on the dev set and 10.4% CER reduction on the test set is reported over the best E2E system (baseline). Moreover, it can be seen that CTC2 provides better ASR results than CTC1 due to the fact that the acoustic feature extraction of the entire network and the decoder are jointly trained in CTC2. In these experiments, the default parameter $\lambda = 0.3$, which is tuned for the baseline system in ESPnet, is used for both CTC1 and CTC2.

In addition to the default configuration of ESPnet, we further implement the SpecAugment in our system to investigate its impact on the ASR performance, all experiments are with RNN LM. As shown in Table 2, the baseline system with SpecAugment gives a relative CER reduction of 13.3% on dev set and 17.9% on test set. T-SMAD with SpecAugment continues to outperform the corresponding Transformer baseline with SpecAugment by relative 8.6% on dev and 9.4% on the test set. The best performing system, T-SMAD+SP+SpecA+CTC2 achieves a CER of 4.8% and 5.1% on the dev and test set, respectively. To the best of authors' knowledge, these are the best results reported on the Aishell-1 corpus. It can be concluded that the proposed SMAD achieves improved alignment due to the deep acoustic structure and the mixed attention, and yields consistent performance improvements over the standard Transformer architecture.

To examine the contribution of each component in SMAD, we perform several ASR experiments by removing the encoder, DAS, mixed attention and modelity-specific network one at a time. To focus on the SMAD mechanism, all the results are produced without additional LM and are reported in Table 3.

Firstly, we remove the encoder in T-SMAD. For a fair comparison, we increase the number of decoder layers to 18 in order to keep the number of model parameters the same. Directly concatenating the acoustic features with linguistic targets as the input to the decoder increases the CER from 6.1% to 8.2% on the test set, that suggests the encoder block is essential for effective acoustic representation. Secondly, we give the same encoder acoustic representation to each SMAD layer in the say way as the standard Transformer, without the deep acoustic structure ('T-SMAD w/o DAS'). This system gives a higher CER than T-SMAD, that confirms the effectiveness of the deep acoustic structure. Thirdly, we replace the mixed attention with the two standard attention mechanisms of Transformer to extract the linguistic features and learn the source-target alignment, respectively. We observe that the removal of mixed attention degrades the performance, that suggests that mapping acoustic-linguistic into the same subspace does help to learn a better alignment. Lastly, the contribution of a modality-specific network has been found to be less prominent than the previous components. It is worth noting that even without modality-specific network, T-SMAD still outperforms the standard Transformer without any increase in the number of model parameters.

Table 2: *Results (CER%) with SpecAugment on Aishell-1*

System	Dev	Test
Transformer +SP+SpecA	5.8	6.4
Transformer +SP+SpecA+CTC	5.2	5.5
T-SMAD +SP+SpecA	5.3	5.8
T-SMAD +SP+SpecA+CTC1	5.0	5.4
T-SMAD +SP+SpecA+CTC2	4.8	5.1

Table 3: *Contribution of each component in SMAD architecture to the ASR performance on Aishell-1. All experiments are with speed perturbation and SpecAugment.*

System	Dev	Test
T-SMAD	5.6	6.1
Transformer	6.7	7.4
T-SMAD w/o encoder	7.5	8.2
T-SMAD w/o DAS	6.6	7.3
T-SMAD w/o mixed attention	6.2	6.9
T-SMAD w/o modality-specific network	5.8	6.3

5. Conclusion

We propose a novel decoder structure for Transformer-based LVCSR that features a self-and-mixed attention decoder (SMAD) with a deep acoustic structure (DAS) to improve the acoustic representation. With SMAD mechanism, we have studied the interaction between acoustic and linguistic representation in the training and decoding of LVCSR system, that opens up a promising future direction for improving E2E ASR systems. We confirm that SMAD and DAS effectively improve the acoustic-linguistic representation in the decoder. The performance gain is attributed to the self-and-mixed attention mechanism that improves the acoustic-linguistic association and alignment in the Transformer decoder. The proposed technique has achieved the best results ever reported on both the dev and test sets of Aishell-1. Furthermore, we also investigate the impact of the components of the SMAD on the ASR performance and validate their effectiveness.

6. References

[1] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Advances in neural information processing systems*, 2014, pp. 3104–3112.

[2] 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*, 2015, pp. 577–585.

[3] D. Bahdanau, J. Chorowski, D. Serdyuk, P. Brakel, and Y. Bengio, "End-to-end attention-based large vocabulary speech recognition," in *Proc. ICASSP*. IEEE, 2016, pp. 4945–4949.

[4] J. Chorowski and N. Jaitly, "Towards better decoding and language model integration in sequence to sequence models," *arXiv preprint arXiv:1612.02695*, 2016.

[5] Y. Zhang, W. Chan, and N. Jaitly, "Very deep convolutional networks for end-to-end speech recognition," in *Proc. ICASSP*. IEEE, 2017, pp. 4845–4849.

[6] S. Kim, T. Hori, and S. Watanabe, "Joint ctc-attention based end-to-end speech recognition using multi-task learning," in *Proc. ICASSP*. IEEE, 2017, pp. 4835–4839.

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

[8] P. Huang, H. Lee, S. Wang, K. Chen, Y. Tsao, and H. Wang, "Exploring the encoder layers of discriminative autoencoders for lvcsr," in *Proc. INTERSPEECH*, 2019, pp. 1631–1635.

[9] 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*, 2015, pp. 577–585.

[10] J. Chorowski, D. Bahdanau, K. Cho, and Y. Bengio, "End-to-end continuous speech recognition using attention-based recurrent nn: First results," *arXiv preprint arXiv:1412.1602*, 2014.

[11] L. Dong, S. Xu, and B. Xu, "Speech-transformer: a no-recurrence sequence-to-sequence model for speech recognition," in *Proc. ICASSP*. IEEE, 2018, pp. 5884–5888.

[12] C. Chiu, T. N. Sainath, Y. Wu, R. Prabhavalkar, P. Nguyen, Z. Chen, A. Kannan, R. J. Weiss, K. Rao, E. Gonina *et al.*, "State-of-the-art speech recognition with sequence-to-sequence models," in *Proc. ICASSP*. IEEE, 2018, pp. 4774–4778.

[13] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[14] B. Yang, Z. Tu, D. F. Wong, F. Meng, L. S. Chao, and T. Zhang, "Modeling localness for self-attention networks," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2018, p. 4449–4458.

[15] S. Hochreiter, Y. Bengio, P. Frasconi, and J. Schmidhuber, "Gradient flow in recurrent nets: the difficulty of learning long-term dependencies," in *A Field Guide to Dynamical Recurrent Neural Networks*. IEEE Press, 2001, pp. 237–243.

[16] P. Zhou, R. Fan, W. Chen, and J. Jia, "Improving generalization of transformer for speech recognition with parallel schedule sampling and relative positional embedding," *arXiv preprint arXiv:1911.00203*, 2019.

[17] N.-Q. Pham, T.-S. Nguyen, J. Niehues, M. Muller, and A. Waibel, "Very deep self-attention networks for end-to-end speech recognition," in *Proc. INTERSPEECH*, 2019, pp. 66–70.

[18] J. Li, X. Wang, Y. Li *et al.*, "The speechtransformer for large-scale mandarin chinese speech recognition," in *Proc. ICASSP*. IEEE, 2019, pp. 7095–7099.

[19] S. Li, D. Raj, X. Lu, P. Shen, T. Kawahara, and H. Kawai, "Improving transformer-based speech recognition systems with compressed structure and speech attributes augmentation," in *Proc. INTERSPEECH*, 2019, pp. 4400–4404.

[20] S. Zhou, L. Dong, S. Xu, and B. Xu, "Syllable-based sequence-to-sequence speech recognition with the transformer in mandarin chinese," in *Proc. INTERSPEECH*, 2018, pp. 791–795.

[21] S. Zhou, L. Dong, S. Xu, and B. Xu, "A comparison of modeling units in sequence-to-sequence speech recognition with the transformer on mandarin chinese," in *International Conference on Neural Information Processing*, 2018, pp. 210–220.

[22] T. He, X. Tan, Y. Xia, D. He, T. Qin, Z. Chen, and T.-Y. Liu, "Layer-wise coordination between encoder and decoder for neural machine translation," in *Advances in Neural Information Processing Systems*, 2018, pp. 7944–7954.

[23] S. Karita, N. E. Y. Soplin, S. Watanabe, M. Delcroix, A. Ogawa, and T. Nakatani, "Improving Transformer-Based end-to-End speech recognition with connectionist temporal classification and language model integration," in *Proc. INTERSPEECH*, 2019, pp. 1408–1412.

[24] S. Watanabe, T. Hori, S. Kim, J. R. Hershey, and T. 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] A. Graves, S. Fernández, F. Gomez, and J. 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.

[26] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2818–2826.

[27] H. Bu, J. Du, X. Na, B. Wu, and H. Zheng, "Aishell-1: An open-source mandarin speech corpus and a speech recognition baseline," in *Oriental COCOSDA*, 2017.

[28] 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 *Proc. INTERSPEECH*, 2018, pp. 2207–2211.

[29] T. Hori, J. Cho, and S. Watanabe, "End-to-end speech recognition with word-based rnn language models," in *IEEE Spoken Language Technology Workshop (SLT)*, 2018, pp. 389–396.

[30] A. Kannan, Y. Wu, P. Nguyen, T. N. Sainath, Z. Chen, and R. Prabhavalkar, "An analysis of incorporating an external language model into a sequence-to-sequence model," in *Proc. ICASSP*. IEEE, 2018, pp. 5824–5828.

[31] T. Ko, V. Peddinti, D. Povey, and S. Khudanpur, "Audio augmentation for speech recognition," in *Proc. INTERSPEECH*, 2015, pp. 3586–3589.

[32] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "Specaugment: A simple augmentation method for automatic speech recognition," in *Proc. INTERSPEECH*, 2019, pp. 2613–2617.

[33] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The kaldi speech recognition toolkit," in *Proc. IEEE ASRU Workshop*, Dec. 2011.

[34] C. Shan, C. Weng, G. Wang, D. Su, M. Luo, D. Yu, and L. Xie, "Component fusion: Learning replaceable language model component for end-to-end speech recognition system," in *Proc. ICASSP*. IEEE, 2019, pp. 5361–5635.

[35] Z. Tian, J. Yi, J. Tao, Y. Bai, and Z. Wen, "Self-attention transducers for end-to-end speech recognition," in *Proc. INTERSPEECH*, 2019, pp. 4395–4399.

[36] S. Karita, N. Chen, T. Hayashi, T. Hori, H. Inaguma, Z. Jiang, M. Someki, N. E. Y. Soplin, R. Yamamoto, X. Wang, S. Watanabe, T. Yoshimura, and W. Zhang, "A comparative study on transformer vs rnn in speech applications," *Proc. IEEE ASRU Workshop*, pp. 449–456, 2019.