INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Automatic Speech Recognition Transformer with Global Contextual Information Decoder

Yukun Qian[*], *Xuyi Zhuang*[*], *Mingjiang Wang*[†]

Harbin Institute of Technology, Shenzhen, China
{20S052011, 19S052014}@stu.hit.edu.cn, mjwang@hit.edu.cn

Abstract

Most current automatic speech recognition (ASR) models use decoders that do not have access to global contextual information at the token level. Therefore, we propose a decoder structure with text-level global contextual information. We construct the global information encoder based on non-autoregressive recognition. To eliminate the non-autoregressive independence assumption, we add a self-attention layer with rotary position encoding. The obtained text-level global contextual information and the decoder are fused as cross-attention to construct a decoder with contextual information. Our model can achieve a character error rate of 3.92% on the AISHELL-1 validation set and 4.35% on the test set, reducing the error rate by 1.72%(dev)/2.13%(test) compared to the baseline model, achieving SOTA performance. Finally, we also use visualization techniques to explain the role of global information in the decoder.

Index Terms: speech recognition, transformer, attention, contextual information, non-autoregressive

1. Introduction

In recent years, significant progress has been made in research on end-to-end (E2E) models in automatic speech recognition (ASR) systems [1, 2, 3, 4, 5, 6, 7]. The decoding method of E2E ASR can be divided into two types: autoregressive and non-autoregressive [8, 9]. The autoregressive decoder is primarily based on the attention mechanism decoder [10]. It can only access information before its own time step at each step and cannot access future information. The CTC-based decoder dominates the non-autoregressive decoder [11, 2, 3], which makes a strong independence assumption. It decodes the token independently for each time step and outputs the maximum probability path result. Based on their decoding principles, autoregressive decoders face the problem of amplification of cumulative errors due to sequence antecedent errors, while non-autoregressive decoders face problems such as the lack of global contextual correlation due to independent decoding.

Several approaches have been proposed to address the limitations of these decoding mechanisms. For instance, CUSIDE proposes training an additional feature predictor to predict future block features and inject context information into the autoregressive Transformer [12]. Similarly, [13] proposes dynamically selecting contextual information with the most discriminatory degree for speech recognition based on the current frame. In [14], a dual decoding method based on tokens of different scales is proposed to fuse dense information of small-scale tokens with the sparse information of large-scale tokens to achieve internal contextual transfer. [15] proposed a self-attention mechanism based on the combination of similarity and content to influence the decoding of the current location using information from other locations with similar information. However, these studies are limited by the autoregressive decoding mechanism and usually provide only limited information on the future context of the receptive field. This contextual information lacks the ability to represent distant contexts. Therefore, it is ineffective in improving cumulative errors caused by antecedent errors, especially in the case of initial token recognition errors. Compared to token representations, the ability of audio to represent information is sparse, as it can still express its original information if the audio is downsampled by a factor of 2 or even 4, while it is difficult to recover the original information if the tokens are downsampled. Therefore, constructing contextual information at the text level is more beneficial to improve ASR performance. We propose leveraging the non-autoregressive one-step decoding feature to pre-identify audio representations and obtain potential recognition results to address this. These results are then recorded using the 2D rotary position embedding (2DRoPE) [16] and re-encoded using the multi-headed attention (MHA) mechanism to reduce the effect of independence assumption and generate global contextual representations. The latent representations of the audio and text-level global contextual information are then fed into the decoder for integrated decoding. In summary, the main contributions of this paper are as follows:

1. A non-autoregressive-based contextual encoder is proposed to construct text-level global contextual information to help the autoregressive decoder reduce the error rate.
2. In the contextual encoder, we introduce a self-attention structure with 2DRoPE position encoding to help the model break the non-autoregressive independence assumption.
3. We show that the performance of the decoder with the introduction of text-level global contextual information achieves SOTA performance, and we also show how the text-level global contextual information helps the autoregressive decoder to perform decoding.

2. Method

To make the global context information available to the decoder, we construct an Acoustic Encoder (AE), a Contextual Encoder (CE), and a Decoder. Figure 1 shows the overall ASR framework structure. The Acoustic Encoder is the Transformer structure responsible for encoding the acoustic feature signal. The Contextual Encoder predicts the corresponding text non-autoregressively, using the output from the Acoustic Encoder.

[*]These authors contributed to the work equally and should be regarded as co-first authors.

[†]Corresponding authors.

10.21437/Interspeech.2023-1450

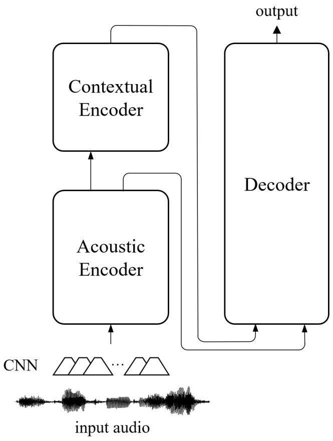


Figure 1: *General structure of the model.*

The global context encoding information is extracted from the last hidden layer of the Contextual Encoder. The Decoder then utilizes the Acoustic Encoder and Contextual Encoder outputs to make token predictions. In the following sections, we will describe each component in detail and explain how the Decoder leverages contextual information.

2.1. Feature Extraction and Acoustic Encoder

We opted to use FBank features over MFCC for our ASR system, as it provides a more detailed representation of the audio signal. However, the computation of the Transformer, the main structure of our Acoustic Encoder, is proportional to the square of the sequence length. Therefore, to improve computational efficiency, we downsampled the FBank features through a 2D convolutional network. Specifically, we employed a 4-layer VGG network to compress the FBank features to 25% of their original size. The resulting features are then fed into a standard 6-layer Transformer encoder to derive the potential audio representation, denoted as $O_{AE} = o_1, o_2, \ldots, o_T$.

2.2. Contextual Encoder

This section adopts the non-autoregressive structure introduced in a previous study [8]. To generate a position vector P, we set its length to a fixed value of L, which is determined by the maximum text length in the dataset. We then use the position vector V_{pos} as a query vector to perform cross-attention with the output O_{AE} of the Acoustic Encoder. The calculation formula is presented below:

$$V_{pos} = \text{POSEmbedding}(L) \quad (1)$$

$$H = \text{CrossAttentionBlock}(V_{pos}, O_{AE}, O_{AE}) \quad (2)$$

However, non-autoregressive prediction assumes that each word is independent of the others, which can limit the ability of the model to capture text-level global contextual information. To address this limitation, we introduce self-attention based on 2DRopE position coding, which combines relative and absolute positioning information [16]. To further enhance the ability of the model to extract global context, we add an additional self-attention module with $N-1$ layers. Figure 2 depicts the resulting Contextual Encoder structure.

The overall function of the Contextual Encoder is to predict the Acoustic Encoding range and the corresponding text for each decoding position using position vectors. The hidden layer vector in the contextual decoder contains text-level global contextual information, which is optimized to match the target text.

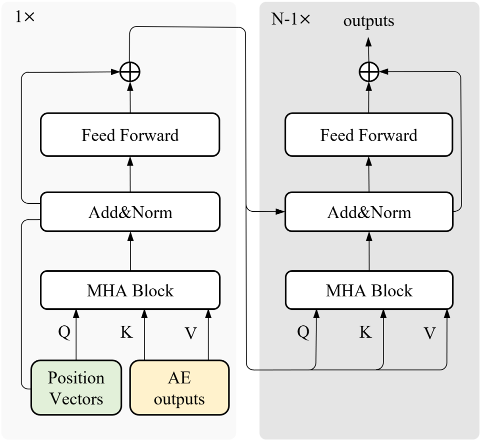


Figure 2: *Contextual Encoder structure.*

2.3. Decoder

The decoder design is rooted in the Transformer autoregressive decoder, which operates on a step-by-step prediction basis and lacks access to global contextual information. With the introduction of contextual information, possible future information can be sensed in advance through the attention mechanism, and the contextual information can be used to make corrections to existing predictions.

To harness the benefits of the global context information provided by the Contextual Encoder, we developed three distinct structures, illustrated in Figure 3.

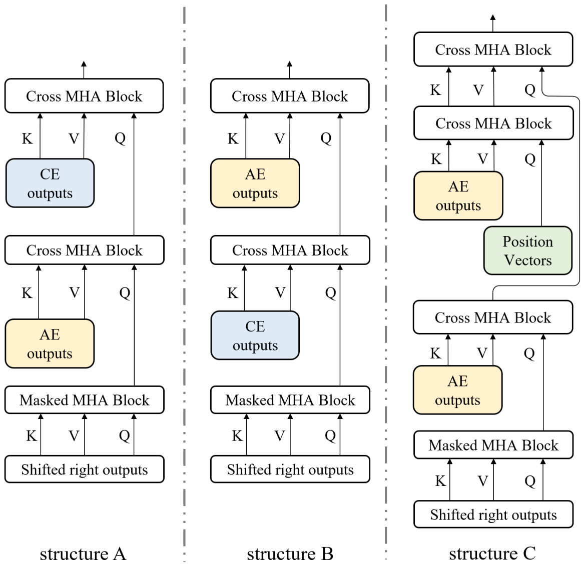


Figure 3: *Three structures that integrate global information.*

Structure A and B models utilize the outputs of the Contextual Encoder and incorporate them into the Decoder as additional information. The Contextual Encoder outputs play a similar role to Acoustic outputs in a standard Transformer decoder, enabling the decoder to access potential future contextual information and use it to refine its current predictions. The primary difference between the structure A model and structure B lies in the sequence of the Acoustic Encoder and Contextual Encoder outputs. In Structure A, the Acoustic Encoder outputs precede the Contextual Encoder outputs, while in Structure B, the order is reversed.

Since the Contextual Encoder outputs are fixed in structures A and B, we designed structure C, which integrates the Contextual Encoder into the Decoder. In structure C, we achieve dynamic Contextual Encoder output changes by feeding each hidden vector of the Contextual Encoder layer into the corresponding decoder layer. Since the layer structures of the Contextual Encoder and decoder are similar, we merged them to form structure C depicted in Figure 3. However, this structure designed this way leads to its inability to break the non-autoregressive independence assumption. We will describe a comparison of these three structures in the experimental section.

2.4. Joint Optimization

While training our speech recognition model, we employ the cross entropy loss function to compute the loss value for both the Decoder module and the Contextual Encoder module, optimizing them simultaneously. However, given the varying difficulty of these two tasks, we adopt a loss weight adaptation method based on [17], which adjusts the weight of different tasks according to their error rate. Specifically, we define the weight of each task as a function of γ_i and the character error rate (CER), which serves as a metric of the model's accuracy. This function is expressed in Equation 3.

$$\alpha_i = -(\text{CER}_i)^{\gamma_i} \log(1 - \text{CER}_i) \tag{3}$$

As evident from the formula, a higher weight is assigned to a task with a relatively higher error rate. It is important to note that at the start of training, the CER of a task may be equal to or greater than 1, which can cause an error in the weight calculation. Therefore, we treat the difficulty of the two tasks as equal at the beginning of training. The total loss calculation is expressed as Equation 4.

$$\ell = \begin{cases} \ell_{CE} + \ell_{Decoder} & \ell_{CE} \geq 1 \ or \ \ell_{Decoder} \geq 1 \\ \alpha_1 \times \ell_{CE} + \alpha_2 \times \ell_{Decoder} & others \end{cases} \tag{4}$$

Where $\ell_{Decoder}$ is the loss of the Decoder, ℓ_{CE} is the loss of the Contextual Encoder.

3. Experiment and result

3.1. Experimental Setups

Our model is trained on the open source dataset AISHELL-1 [18], which contains 150 hours of training speech, 18 hours of validation speech, and about 10 hours of test speech. All audio is in 16KHz WAV format. We use an 80-dimensional FBank feature, where the frame length is 25ms, and the frameshift is 10ms. Acoustic Encoder, Contextual Encoder, and Decoder layers are 6, 4, and 6, respectively. The hidden layer dimension is 512, and each layer has 6 attention heads. We use a dropout rate of 0.1 to avoid overfitting and employ SpecAugment [19] for data enhancement. We have a maximum of 20% for the time mask and 50% for the frequency mask. Each batch contains approximately 12.8 minutes of audio. We used 4 NVIDIA RTX4090 GPUs and the Pytorch framework [20] for training. We also use the Adam [21] optimizer and adopt the warm-up strategy [10], where the warm-up step is 12000.

3.2. Different Decoder Structure Comparison Experiment

We test the three decoders in Figure 3 on the AISHELL-1 dataset with all the same experimental parameters, using the greedy search for decoding while using the Transformer model as the baseline. The final test results are shown in the following table.

Table 1: *Decoding structure comparison*

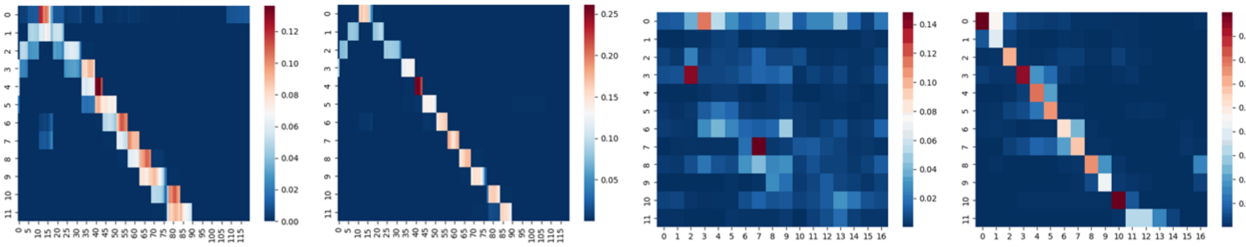
Model name	CER(%)	
	dev	test
Transformer	5.78	6.67
Structure A	**4.06**	**4.55**
Structure B	4.08	4.56
Structure C	4.49	5.03

Based on the data presented in Table 1, it is evident that all three structures exhibit improved accuracy rates compared to the baseline, with structures A and B performing best and showing the minimal difference. However, the error rate of structure C is relatively high. We posit that this is due to incorporating a self-attention mechanism with relative location encoding in structures A and B. This mechanism assists the Contextual Encoder in breaking the independence assumption in non-autoregression. To investigate this further, we conducted a comparison experiment by introducing a cross-attention similar to structure C in structure A, replacing the original self-attention, and assessing the error rate of the Contextual Encoder. The findings of this experiment are detailed in Table 2.

Table 2: *The effect of self-attention*

Model name	CER(%)	
	CE	Decoder
Structure A + Cross attention	6.28	4.97
Structure A + self attention	**5.93**	**4.55**

In this study, we compared the performance of two structures, one using self-attention and the other using cross-attention. We found that the Contextual Encoder error rate of the former is lower than that of the latter, with a reduction from 6.28% to 5.93%. This improvement leads to more accurate text-level global contextual information obtained by the decoder and ultimately results in better accuracy of the model. Interestingly, we observed that the accuracy rate of the Contextual Encoder is lower than that of the Decoder. This indicates that the Decoder cannot directly extract accurate text information from the Contextual Encoder. Instead, the Decoder relies more on the contextual information provided by the Contextual Encoder. Based on the above experiments, unless otherwise specified, all experiments in this section will use structure A.



(a) (b) (c) (d)

Figure 4: *Attention weight assignment for different cross MHA in structure A decoder. (a) Attention to the Acoustic Encoder output and decoder results in the penultimate layer. (b) Attention to the Acoustic Encoder output and decoder results in the last layer. (c) Attention to Contextual Encoder output and decoder results in the penultimate layer. (d) Attention to Contextual Encoder output and decoder results in the last layer.*

3.3. Results on AISHELL-1

We tested the proposed Transformer with Global Contextual Information Decoder (TGCID) model on the AISHELL-1 dataset, where the beam size was set to 5. We compared it with common automatic speech recognition models such as Conformer, WeNeT, etc. The test results are shown in the following table.

Table 3: *Comparison with recently published models*

Method	CER(%)	
	dev	test
(previous work)		
Transformer [22]	5.78	6.67
Transformer + beamsearch [22]	5.65	6.48
ESPNet(Conformer) [23]	4.50	4.90
Branchformer [24]	4.19	4.43
WeNet [25]	-	4.46
(our work)		
TGCID	4.06	4.55
TGCID + beamsearch	**3.92**	**4.35**

Our proposed TGCID model achieved a CER of 4.55% without beam search. With the addition of beam search, it achieved even better results, with a CER of 4.35% on the test set and 4.06% on the validation set, outperforming other comparison models. These results demonstrate that our model achieves state-of-the-art performance.

3.4. The Role of the Contextual Encoder

To assess the impact of our proposed Contextual Encoder on our model, we visualized the last two layers of the two cross-attention parts of the decoder, as presented in Figure 4. The horizontal axis represents the encoder output or contextual information, and the vertical axis denotes the decoder step. Figure 4a and Figure 4b illustrate the attention between the Decoder and the Acoustic Encoder output, while Figure 4c and Figure 4d exhibit the attention between the Decoder and the contextual information.

Figure 4a and 4b reveal that during decoding the first text, the decoder does not allocate its attention to the beginning of the encoder. Instead, it concentrates most of its attention on frames 10 to 15, which contradicts the expectation that the first word should be at the beginning of the audio. We speculate that this is because the first character is predicted with only the beginning symbol in front of it, making it challenging to locate the corresponding position accurately.

Figure 4c shows that after introducing the context information, the decoder initially allocates attention almost evenly to each upper context information position. This helps it to obtain more extensive context information to facilitate the determination of the predicted text. Meanwhile, Figure 4d depicts that the attention between the decoder and the contextual information becomes concentrated and varies almost monotonically. The initial attention focuses on the front, helping the decoder to correct previous character prediction errors caused by the attention allocation error at the beginning. Subsequent decoding steps concentrate on the vicinity of the corresponding position, allowing the decoder to predict the current text more accurately based on nearby information.

Thus, introducing context information enables the decoder to obtain global information but also helps it to correct incorrect information based on audio prediction, thereby enhancing the overall recognition rate. This result supports our hypotheses about the role of the Contextual Encoder.

4. Conclusions

This paper proposes a decoder that integrates global context information. We introduce a non-autoregressive module into the standard Transformer model. We use a self-attention mechanism with 2DRoPE position encoding to construct a Contextual Encoder, breaking the non-autoregressive independence assumption. Additionally, we explore how to fuse context information with the decoder, and through experiments, we determine that the structure of context information introduced after the encoder information yields the best results. Our model achieves an error rate of 4.35% on the AISHELL-1 dataset. Finally, we demonstrate that our contextual information can assist the decoder in recognizing and correcting text through visualization of the cross-attention in the decoder.

5. Acknowledgements

We would like to thank the National Natural Science Foundation of China under Grant No. 62276076, the National Natural Science Foundation of China under Grant No. 62176102, and the Natural Science Foundation of Guangdong Province under Grant No. 2020B1515120.

6. References

[1] A. Graves, A.-r. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2013, pp. 6645–6649.

[2] S. Kim, T. Hori, and S. Watanabe, "Joint ctc-attention based end-to-end speech recognition using multi-task learning," in *IEEE international conference on acoustics, speech and signal processing (ICASSP)*, 2017, pp. 4835–4839.

[3] S. Watanabe, T. Hori, S. Kim, J. R. Hershey, and T. Hayashi, "Hybrid ctc/attention architecture for end-to-end speech recognition," *Selected Topics in Signal Processing*, vol. 11, no. 8, pp. 1240–1253, 2017.

[4] M. Ghodsi, X. Liu, J. Apfel, R. Cabrera, and E. Weinstein, "Rnn-transducer with stateless prediction network," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 7049–7053.

[5] W. Chan, N. Jaitly, Q. V. Le, and O. Vinyals, "Listen, attend and spell," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 6528–6532.

[6] J. K. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, "Attention-based models for speech recognition," *Advances in neural information processing systems*, vol. 28, 2015.

[7] G. I. Winata, S. Cahyawijaya, Z. Lin, Z. Liu, and P. Fung, "Lightweight and efficient end-to-end speech recognition using low-rank transformer," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 6144–6148.

[8] Y. Qian, X. Zhuang, Z. Zhang, L. Zhou, X. Lin, and M. Wang, "Non-autoregressive speech recognition with error correction module," in *Pro. of APSIPA*, 2022, pp. 1104–1109.

[9] Y. Bai, J. Yi, J. Tao, Z. Tian, Z. Wen, and S. Zhang, "Listen attentively, and spell once: Whole sentence generation via a non-autoregressive architecture for low-latency speech recognition," *arXiv preprint arXiv:2005.04862*, 2020.

[10] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[11] A. Graves, "Sequence transduction with recurrent neural networks," *arXiv preprint arXiv:1211.3711*, 2012.

[12] K. An, H. Zheng, Z. Ou, H. Xiang, K. Ding, and G. Wan, "Cuside: Chunking, simulating future context and decoding for streaming asr," *arXiv preprint arXiv:2203.16758*, 2022.

[13] M. Han, L. Dong, Z. Liang, M. Cai, S. Zhou, Z. Ma, and B. Xu, "Improving end-to-end contextual speech recognition with fine-grained contextual knowledge selection," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 8532–8536.

[14] Z. Yang, W. Xi, R. Wang, R. Jiang, and J. Zhao, "Dual-decoder transformer for end-to-end mandarin chinese speech recognition with pinyin and character," *arXiv preprint arXiv:2201.10792*, 2022.

[15] K. Shim and W. Sung, "Similarity and content-based phonetic self attention for speech recognition," *arXiv preprint arXiv:2203.10252*, 2022.

[16] J. Su, Y. Lu, S. Pan, A. Murtadha, B. Wen, and Y. Liu, "Roformer: Enhanced transformer with rotary position embedding," *arXiv preprint arXiv:2104.09864*, 2021.

[17] M. Guo, A. Haque, D.-A. Huang, S. Yeung, and L. Fei-Fei, "Dynamic task prioritization for multitask learning," in *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

[18] 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 *2017 20th Conference of the Oriental Chapter of the International Coordinating Committee on Speech Databases and Speech I/O Systems and Assessment (O-COCOSDA)*, 2017, pp. 1–5.

[19] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "Specaugment: A simple data augmentation method for automatic speech recognition," *arXiv preprint arXiv:1904.08779*, 2019.

[20] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "Pytorch: An imperative style, high-performance deep learning library," *Advances in neural information processing systems*, vol. 32, 2019.

[21] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. of ICLR*, 2015.

[22] L. Dong, S. Xu, and B. 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.

[23] S. Watanabe, T. Hori, S. Karita, T. Hayashi, J. Nishitoba, Y. Unno, N. E. Y. Soplin, J. Heymann, M. Wiesner, N. Chen *et al.*, "Espnet: End-to-end speech processing toolkit," *arXiv preprint arXiv:1804.00015*, 2018.

[24] Y. Peng, S. Dalmia, I. Lane, and S. Watanabe, "Branchformer: Parallel mlp-attention architectures to capture local and global context for speech recognition and understanding," in *International Conference on Machine Learning*. PMLR, 2022, pp. 17 627–17 643.

[25] Z. Yao, D. Wu, X. Wang, B. Zhang, F. Yu, C. Yang, Z. Peng, X. Chen, L. Xie, and X. Lei, "Wenet: Production oriented streaming and non-streaming end-to-end speech recognition toolkit," *arXiv preprint arXiv:2102.01547*, 2021.