



# WeNet 2.0: More Productive End-to-End Speech Recognition Toolkit

Binbin Zhang<sup>1,3</sup>, Di Wu<sup>1,3</sup>, Zhendong Peng<sup>1,3</sup>, Xingchen Song<sup>1,3</sup>, Zhuoyuan Yao<sup>2</sup>, Hang Lv<sup>2,3</sup>, Lei Xie<sup>2,\*</sup>, Chao Yang<sup>1,3</sup>, Fuping Pan<sup>1</sup>, Jianwei Niu<sup>1</sup>

<sup>1</sup>Horizon Robotics, Beijing, China

<sup>2</sup>Audio, Speech and Language Processing Group (ASLP@NPU), School of Computer Science, Northwestern Polytechnical University, Xi'an, China

<sup>3</sup>WeNet Open Source Community

binbin.zhang@horizon.ai, hanglv@nwpu-aslp.org, lxie@nwpu.edu.cn

## Abstract

Recently, we made available WeNet [1], a production-oriented end-to-end speech recognition toolkit, which introduces a unified two-pass (U2) framework and a built-in runtime to address the streaming and non-streaming decoding modes in a single model. To further improve ASR performance and facilitate various production requirements, in this paper, we present WeNet 2.0 with four important updates. (1) We propose U2++, a unified two-pass framework with bidirectional attention decoders, which includes the future contextual information by a right-to-left attention decoder to improve the representative ability of the shared encoder and the performance during the rescoring stage. (2) We introduce an n-gram based language model and a WFST-based decoder into WeNet 2.0, promoting the use of rich text data in production scenarios. (3) We design a unified contextual biasing framework, which leverages user-specific context (e.g., contact lists) to provide rapid adaptation ability for production and improves ASR accuracy in both with-LM and without-LM scenarios. (4) We design a unified IO to support large-scale data for effective model training. In summary, the brand-new WeNet 2.0 achieves up to 10% relative recognition performance improvement over the original WeNet on various corpora and makes available several important production-oriented features.

**Index Terms:** U2++, Language Model, Contextual Biasing, UIO, Toolkit

## 1. Introduction

End-to-end (E2E) methods, such as connectionist temporal classification (CTC) [2], recurrent neural network transducer (RNN-T) [3, 4, 5], and attention based encoder-decoder (AED) [6, 7, 8, 9], have drawn immense interest in the last few years in automatic speech recognition (ASR). Recent works [10, 11, 12, 13, 14] show the E2E systems not only extremely simplify the speech recognition pipeline, but also surpass the conventional hybrid ASR systems in accuracy.

Considering the advantages [15] of E2E models, deploying the emerging E2E ASR framework into real-world productions with stable and highly efficient characteristics becomes necessary. However, almost all of the well-known E2E speech recognition toolkits, such as ESPnet [16] and SpeechBrain [17], are research-oriented rather than production-oriented.

In [1], we present a production first and production ready E2E speech recognition toolkit, WeNet<sup>1</sup>, which focuses on ad-

ressing the production problems of transformer [18] and conformer [19] based E2E models. Specifically, WeNet adopts a joint CTC/AED structure as the basic model structure. And then, a unified two-pass (U2) framework is proposed to solve the streaming problem, where a dynamic chunk masking strategy is applied in the training procedure to unify the streaming and non-streaming modes in a unified neural model. A built-in runtime is provided in which developers can run x86 and android E2E system for speech recognition out of the box. WeNet dramatically reduces the workload of deploying an E2E model in real applications, so it is widely adopted by researchers and developers.

In this paper, we present **WeNet 2.0**, which introduces the recent developments and solutions updated in WeNet for production-oriented speech recognition. The key updates of WeNet 2.0 are as follows.

**(1) U2++:** we upgrade the original U2 framework to U2++, which simultaneously utilizes the left-to-right and right-to-left bidirectional contextual information of the labeled sequences to learn richer context information during the training stage, and combines the forward and backward predictions to achieve more accurate results during the decoding stage. The experiments show that U2++ achieves up to 10% relative reduction over the original U2 framework in error rate.

**(2) Production language model solution:** we support an optional n-gram LM, which is composed with the E2E modeling unit in a weighted finite state transducer (WFST) [20] based decoding graph, during the streaming CTC decoding stage. The n-gram LM can be trained rapidly on rich production-accumulated text data. Experiments show that the n-gram language model can provide up to 8% relative performance improvement.

**(3) Contextual biasing:** we design a unified contextual biasing framework which provides the opportunity to leverage the user-specific contextual information with or without LM during the streaming decoding stage. Utilizing user-specific contextual information (e.g., contact lists, particular dialog state, conversation topic, location, etc) plays a great role in both improving the ASR accuracy and providing rapid adaptation ability. Experiments show that our contextual biasing solution can bring clear dramatic improvement for both with-LM and without-LM cases.

**(4) Unified IO (UIO):** we design a unified IO system to support the large-scale dataset for effective model training. The UIO system provides a unified interface for different storage media (i.e., local disk or cloud) and datasets for different scale (i.e., small or large datasets). For small datasets, UIO keeps the

\* Lei Xie is corresponding author.

<sup>1</sup><https://github.com/wenet-e2e/wenet>

sample-level random access ability. While, for large datasets, UIO aggregates the data samples to shards (refer to Figure 4 for more details) and provides the shard-level random access ability. Thanks to UIO, WeNet 2.0 can elastically support training with a few hours to millions of hours of data.

In summary, our brand-new WeNet 2.0 makes available several important production-oriented features, which achieves substantial ASR improvement over the original WeNet and makes itself a more productive E2E speech recognition toolkit.

## 2. WeNet 2.0

In this section, we will describe the production-oriented key updates: U2++ framework, production language model solution, contextual biasing, and unified IO in detail, respectively.

### 2.1. U2++

U2++, a unified two-pass joint CTC/AED framework with bidirectional attention decoders, provides an ideal solution to unify streaming and non-streaming modes. As shown in Figure 1, U2++ consists of four parts. 1) A *Shared Encoder* that models the information of acoustic features. The *Shared Encoder* consists of multiple Transformer [18] or Conformer [19] layers which only takes limited right contexts into account to keep a balanced latency. 2) A *CTC Decoder* that models the frame-level alignment information between acoustic features and token units. The *CTC Decoder* consists of a linear layer, which transforms the *Shared Encoder* output to the CTC activation. 3) A *Left-to-Right Attention Decoder (L2R)* that models the ordered token sequence from left to right to represent the past contextual information. 4) A *Right-to-Left Attention Decoder (R2L)* that models the reversed token sequence from right to left to represent the future contextual information. The *L2R* and *R2L Attention Decoders* consist of multiple Transformer decoder layers. During the decoding stage, the *CTC Decoder* runs in the streaming mode in the first pass and the *L2R* and *R2L Attention Decoders* do rescore in the non-streaming mode to improve the performance in the second pass.

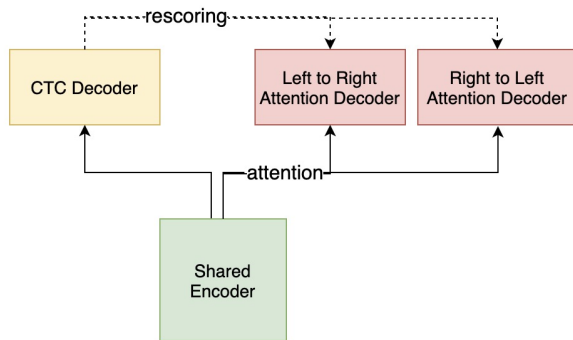


Figure 1: U2++ framework. The dotted line denotes the second pass rescoring during the decoding stage.

Compared with U2 [1], we add an extra right-to-left attention decoder to enhance the context modeling ability of our model, so that the context information comes not only from the past (left-to-right decoder) but also from the future (right-to-left decoder). This improves the representative ability of the shared encoder, the generalization ability of the whole system, and the performance during the rescoring stage.

The combined CTC and AED loss is used to train U2++:

$$\mathbf{L}_{\text{combined}}(\mathbf{x}, \mathbf{y}) = \lambda \mathbf{L}_{\text{CTC}}(\mathbf{x}, \mathbf{y}) + (1 - \lambda)(\mathbf{L}_{\text{AED}}(\mathbf{x}, \mathbf{y})), \quad (1)$$

where  $\mathbf{x}$  denotes the acoustic features, and  $\mathbf{y}$  denotes the corresponding labels,  $\lambda$  is a hyper parameter which balances the importance of the CTC loss  $\mathbf{L}_{\text{CTC}}(\mathbf{x}, \mathbf{y})$  and AED loss  $\mathbf{L}_{\text{AED}}(\mathbf{x}, \mathbf{y})$ .

To incorporate an extra R2L attention decoder into our model, we introduce a hyper parameter  $\alpha$  inside  $\mathbf{L}_{\text{AED}}(\mathbf{x}, \mathbf{y})$  to adjust the contributions of two unidirectional decoders:

$$\mathbf{L}_{\text{AED}}(\mathbf{x}, \mathbf{y}) = (1 - \alpha)\mathbf{L}_{\text{L2R}}(\mathbf{x}, \mathbf{y}) + \alpha(\mathbf{L}_{\text{R2L}}(\mathbf{x}, \mathbf{y})). \quad (2)$$

Similar to U2, the dynamic chunk masking strategy is used to unify the streaming and non-streaming modes. During the training stage, we firstly sample a random chunk size  $C$  from a uniform distribution,  $C \sim U(1, \text{max batch length } T)$ . And then, the input is split into several chunks with the chosen chunk size. At last, the current chunk does bidirectional chunk-level attention to itself and previous/following chunks in L2R/R2L attention decoder respectively in training. During the decoding stage, the n-best results achieved from the first-pass CTC decoder are rescored by the L2R and R2L attention decoder with its corresponding acoustic information generated by the *Shared Encoder*. The final results are obtained by fusing the scores of the two attention decoders and the CTC decoder.

Empirically, a larger chunk size results in better results with higher latency. Thanks to the dynamic strategy, U2++ learns to predict with arbitrary chunk size so that balancing the accuracy and latency is simplified by tuning the chunk size in decoding.

### 2.2. Language Model

To promote the use of rich text data in production scenarios, we provide a unified LM integration framework in WeNet 2.0, which is shown in Figure 2.

As a unified LM and LM-free system, CTC is employed to generate the first-pass n-best results. When LM is not provided, *CTC prefix beam search* is applied to get n-best candidates. When LM is supplied, WeNet 2.0 compiles the n-gram LM ( $G$ ), lexicon ( $L$ ), and end-to-end modeling CTC topology ( $T$ ) into a WFST-based decoding graph ( $TLG$ )

$$TLG = T \circ \min(\text{det}(L \circ G)), \quad (3)$$

and then *CTC WFST beam search* is applied to get n-best candidates. Finally, the n-best candidates are rescored by the *Attention Rescoring* module to find the best candidate.

We reuse the algorithm and code in Kaldi[21] for decoding, which is denoted as *CTC WFST beam search* in the WeNet 2.0 framework. To speedup decoding, blank frame skipping [22] technique is adopted.

### 2.3. Contextual Biasing

Utilizing user-specific contextual information (e.g., contact lists, driver’s navigation) plays a vital role in speech production, which always improves the accuracy significantly and provides the rapid adaptation ability. Contextual biasing technique is investigated in [23, 24, 25] for both traditional and E2E systems.

In WeNet 2.0, for leveraging the user-specific contextual information in both LM and LM-free cases during the streaming decoding stage, inspired by [24], we construct a contextual WFST graph on the fly when a set of biasing phrases are known in advance. Firstly, the biasing phrases are split into biasing units according to the E2E modeling units in the LM-free case or the vocabulary words in the with-LM case. And then, the contextual WFST graph is built on the fly as follows: (1) Each biasing unit with a boosted score is placed on a corresponding arc sequentially to generate an acceptable chain. (2) For each intermediate state of the acceptable chain, a special failure arc

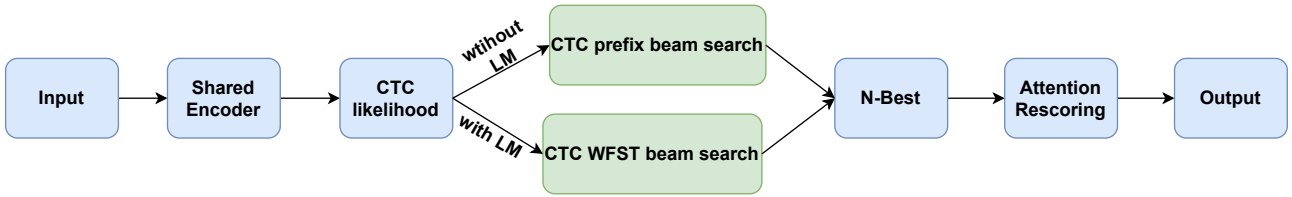


Figure 2: Language model solution in WeNet 2.0

with a negative accumulated boosted score is added. The failure arcs are used to remove the boosted scores when only partial biasing units are matched rather than the entire phrase. In Figure 3, a char (E2E modeling unit) level context graph in the LM-free case and a word-level context graph in the with-LM case are shown, respectively. Finally, during the streaming decoding stage, a boosted score is added immediately when the beam search result matches with a biasing unit through the contextual WFST graph, as shown in

$$\mathbf{y}^* = \arg \max_{\mathbf{y}} \log P(\mathbf{y}|\mathbf{x}) + \lambda \log P_C(\mathbf{y}), \quad (4)$$

where  $P_C(\mathbf{y})$  is the biasing score,  $\lambda$  is a tunable hyper-parameter to control how much the contextual LM influences the overall model score. Especially, when some biasing phrases share the same prefix, we do a greedy match to simplify the implementation.

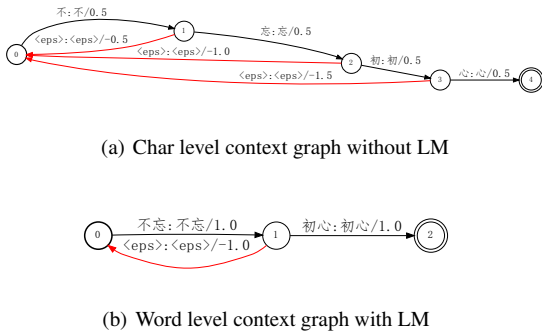


Figure 3: Context graph for Chinese phrase “不忘初心”. The failure arcs are highlighted in red lines.

## 2.4. UIO

Typically, the production-scale speech dataset is over tens of thousands of hours labeled speech, which always comprises massive small size files. The massive small files will cause the following problems:

- **Out of memory (OOM):** we have to keep the index information for all of the small files, which is very memory-consuming for large datasets.
- **Lower training speed:** reading data becomes the bottleneck of the training due to the intolerable time-consuming during random access to massive small files.

To address the aforementioned problems for large-scale production datasets and keep the high efficiency for small datasets at the same time, we design a unified IO system, which provides a unified interface for different storage (i.e., local disk or cloud) and different scale datasets (i.e., small or large datasets). The whole framework is shown in Figure 4.

For large datasets, inspired by TFRecord (Tensorflow) [26] and AIStore [27], we pack each set of samples (e.g., 1000 samples) with their associated metadata information into a corresponding bigger shard, which is done by GNU tar, an open-

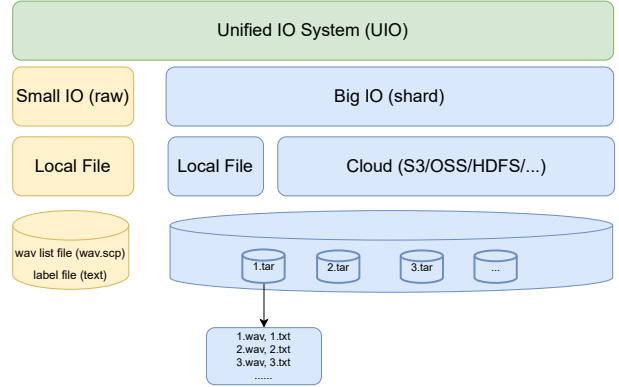


Figure 4: Unified IO in WeNet 2.0

source, open-format, ubiquitous and widely-supported archival tool. Tar files significantly save memory and overcome the OOM problem effectively. During the training stage, on-the-fly decompression is performed in the memory, and the data in the same compressed shard is read sequentially, which solves the time-consuming random data access problem and speeds up the training process. At the same time, different shards can be read randomly to ensure the global randomness of data. For small datasets, we can also load the training samples directly.

Especially, when using shards for big datasets, we support loading the shards from both the local disk and the distributed storage (e.g., S3, OSS, HDFS, etc). Similar to TFRecord, chain operations for processing data on-the-fly is designed, so that the unified IO is extensible and easy to debug<sup>2</sup>.

## 3. Experiments

In this section we describe our experimental setup, test sets, and analyze the experimental results. Most of the experimental setups are available in the WeNet recipes.

The experiments are carried on all of or some of the following corpora: AISHELL-1 [28], AISHELL-2 [29], LibriSpeech [30], GigaSpeech [31], and recent released Wenet-Speech [32]. The five corpora include different languages (English and Mandarin), recording environments (clean and noisy), and sizes (100 - 10000 hours).

### 3.1. U2++

To evaluate the effectiveness of our U2++ model, we conduct experiments on all of the 5 ASR corpora listed above. For most experiments, 80-dimensional log Mel-filter banks (FBANK) with a 25ms window and a 10ms shift are used as acoustic features. SpecAugment [33] is applied on-the-fly for data augmentation. Two convolution subsampling layers with kernel size 3\*3 and stride 2 are used in the front of the encoder. We use 12 conformer layers for the encoder. To keep comparable parameters for U2/U2++, we use 6 transformer decoder layers for

<sup>2</sup><https://wenet.org.cn/wenet/UIO.html>

Table 1: U2++/U2 comparison on various open-source ASR corpora. \* denotes the model is trained with dynamic mask, and it could be decoded with different chunk size (full/16 in the table). † denotes the model is trained with full attention.

Dataset	Language	Hours	Unit	Metric	Test Sets	U2	U2++
AISHELL-1 *	Mandarin	170	Char	CER	test(full)/test(16)	4.97/5.45	<b>4.63/5.05</b>
AISHELL-2 *	Mandarin	1000	Char	CER	test ios(full)/test ios(16)	6.08/6.46	<b>5.39/5.78</b>
LibriSpeech †	English	960	BPE	WER	test clean/test other	2.85/7.24	<b>2.66/6.53</b>
WenetSpeech †	Mandarin	10000	Char	MER	Test Net/Test Meeting	9.70/ <b>15.59</b>	<b>9.25/16.18</b>
GigaSpeech †	English	10000	BPE	WER	dev/test	11.30/11.20	<b>10.70/10.60</b>

U2, 3 left-to-right and 3 right-to-left decoder layers for U2++. Moreover, we obtain our final model by model averaging. In AISHELL-1 and AISHELL-2, the attention layer uses attention dimension 256, feed-forward 2048, and 4-head attention. In LibriSpeech, GigaSpeech and WenetSpeech, the attention layer uses attention dimension 512, feed-forward 2048, and 8-head attention. The kernel size of the convolution module is 8/8/31/31/15 for the five corpora, respectively. Accumulating gradient is used to stabilize training.

In Table 1, we report the character error rate (CER), word error rate (WER) or mixed error rate (MER) for each corpus. It shows that U2++ outperforms U2 on most of the corpora, and achieves up to 10% relative improvement over U2. From the results, we approve that U2++ shows superior performance in various types and sizes of ASR corpora consistently.

### 3.2. N-gram Language Model

The language model solution is evaluated on AISHELL-1, AISHELL-2, and LibriSpeech on the models listed in Section 3.1. For AISHELL-1 and AISHELL-2, a tri-gram trained on its own training corpus is used. For LibriSpeech, the official pre-trained four-gram large language model (fglarge) is used. To speed up the decoding, when the probability of the “blank” symbol of a frame is greater than 0.98, the frame will be skipped.

Table 2: N-gram LM evaluation

Dataset	Test Sets	without LM	with LM
AISHELL-1	test	4.63	<b>4.40</b>
AISHELL-2	test_ios	5.39	<b>5.35</b>
LibriSpeech	test_clean/test_other	2.66/6.53	<b>2.65/5.98</b>

Table 2 exhibits the comparison results on the three corpora in with-LM and without-LM scenarios. The with-LM solution shows 5% gain on AISHELL-1, comparable result on AISHELL-2, comparable result in test\_clean on LibriSpeech, and superior improvement (8.42%) in test\_other on LibriSpeech.

In conclusion, our production LM solution shows consistent gain on the three corpora, which illustrates n-gram language model works well when integrated to an E2E speech recognition system. Benefitting from this solution, the production system can enjoy the rich production-accumulated text data.

### 3.3. Contextual Biasing

We evaluate contextual biasing in a contact scenario, and we design two test sets selected from AISHELL-2 test sets.

- **test\_p**: positive test set with its relevant context. We select 107 utterances with person names, and all the names are attached as contextual biasing phrases in decoding.
- **test\_n**: negative test set without any contextual biasing phrases (i.e., names) in the positive testset. We randomly select 107 eligible utterances as the negative test set.

We carry on the test on the AISHELL-1 U2++ model with and without LM. We show that the strength of contextual biasing can be controlled by switching the boosted score. In our experiments, the boosted score is varied from 0 to 10, where 0 means context biasing is not applied. As shown in Table 3, our solution reduces the error rate on the positive test set and a bigger boosted score usually gives better improvement. Furthermore, when a proper boosted score is set, we can improve the performance on the positive testset and avoid the performance degradation on negative test set at the same time.

Table 3: Context Biasing evaluation

LM	test set	boosted score				
		0	3	5	7	10
NO	test_p	14.94	11.11	7.65	<b>6.17</b>	6.54
	test_n	7.45	7.45	7.45	7.57	8.06
YES	test_p	13.95	11.23	11.11	10.12	<b>9.14</b>
	test_n	7.20	7.20	7.20	7.20	7.33

### 3.4. UIO

We evaluate the performance of UIO on AISHELL-1 [28] test-set and WenetSpeech [32] testsets (i.e., test\_net/test\_meeting), including the accuracy, scalability and training speed. Specifically, we use 8 GPUs on one machine for AISHELL-1, and 24 GPUs on three machines for WenetSpeech.

As shown in Table 4, the UIO shows close accuracy in raw and shard modes on AISHELL-1, while the shard mode gets about 9.27% speedup in terms of the training speed. For WenetSpeech, since it is too slow to only use raw mode for training, the training speed is not shown. We adopt the result of ESPnet (marked with †) as our baseline, which has the same configuration as our model. The result of the model trained on WenetSpeech by shard mode is comparable to ESPnet, which further illustrates the effectiveness of UIO.

Table 4: Unified IO evaluation

Dataset	metric	raw	shard
AISHELL-1	speed(per epoch)	593 sec	<b>538 sec</b>
	WER	4.63	4.67
WenetSpeech	speed(per epoch)	NA	75 min
	WER	8.90/15.90(†)	9.70/15.59

## 4. Conclusion and Future Work

In this paper, we proposed our more productive E2E speech recognition toolkit, WeNet 2.0, which introduces several important production-oriented features and achieves substantial ASR performance improvement.

We are working on WeNet 3.0, which mainly focuses on unsupervised self-learning, on-device model exploration and optimization, and other characteristics for production-level ASR.

## 5. References

- [1] Zhuoyuan Yao, Di Wu, Xiong Wang, Binbin Zhang, Fan Yu, Chao Yang, Zhendong Peng, Xiaoyu Chen, Lei Xie, Xin Lei, “WeNet: Production oriented streaming and non-streaming end-to-end speech recognition toolkit,” in *ISCA Conference of the International Speech Communication Association (Interspeech)*, 2021, pp. 4054–4058.
- [2] Alex Graves, Santiago Fernández, Faustino Gomez, Jürgen Schmidhuber, “Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks,” in *ACM International Conference on Machine Learning (ICML)*, 2006, pp. 369–376.
- [3] Alex Graves, “Sequence transduction with recurrent neural networks,” in *ACM International Conference on Machine Learning Representation Learning Workshop (ICML)*, 2012.
- [4] Xiong Wang, Zhuoyuan Yao, Xian Shi, Lei Xie, “Cascade rnn-transducer: Syllable based streaming on-device mandarin speech recognition with a syllable-to-character converter,” in *IEEE Spoken Language Technology Workshop (SLT)*, 2021, pp. 15–21.
- [5] Senmao Wang, Pan Zhou, Wei Chen, Jia Jia, Lei Xie, “Exploring rnn-transducer for chinese speech recognition,” in *IEEE Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, 2019, pp. 1364–1369.
- [6] Jan Chorowski, Dzmitry Bahdanau, Kyunghyun Cho, Yoshua Bengio, “End-to-end continuous speech recognition using attention-based recurrent nn: First results,” in *Annual Conference on Neural Information Processing Systems (NIPS) Workshop on Deep Learning*, 2014.
- [7] William Chan, Navdeep Jaitly, Quoc Le, Oriol Vinyals, “Listen, attend and spell: A neural network for large vocabulary conversational speech recognition,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 4960–4964.
- [8] Haoneng Luo, Shiliang Zhang, Ming Lei, Lei Xie, “Simplified self-attention for transformer-based end-to-end speech recognition,” in *IEEE Spoken Language Technology Workshop (SLT)*, 2021, pp. 75–81.
- [9] Haoran Miao, Gaofeng Cheng, Changfeng Gao, Pengyuan Zhang, Yonghong Yan, “Transformer-based online ctc/attention end-to-end speech recognition architecture,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 6084–6088.
- [10] Rohit Prabhavalkar, Kanishka Rao, Tara N Sainath, Bo Li, Leif Johnson, Navdeep Jaitly, “A comparison of sequence-to-sequence models for speech recognition,” in *ISCA Conference of the International Speech Communication Association (Interspeech)*, 2017, pp. 939–943.
- [11] Tara N Sainath, Ruoming Pang, David Rybach, Yanzhang He, Rohit Prabhavalkar, Wei Li, Mirkó Visontai, Qiao Liang, Trevor Strohman, Yonghui Wu, Ian McGraw, Chung-Cheng Chiu, “Two-pass end-to-end speech recognition,” 2019, pp. 2773–2777.
- [12] Suyoun Kim, Takaaki Hori, Shinji 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.
- [13] Tara N Sainath, Yanzhang He, Bo Li, Arun Narayanan, Ruoming Pang, Antoine Bruguier, et al., “A streaming on-device end-to-end model surpassing server-side conventional model quality and latency,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 6059–6063.
- [14] Ke Hu, Tara N Sainath, Ruoming Pang, Rohit Prabhavalkar, “Deliberation model based two-pass end-to-end speech recognition,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 7799–7803.
- [15] Jinyu Li, “Recent advances in end-to-end automatic speech recognition,” in *arXiv preprint arXiv:2111.01690*, 2022.
- [16] Shinji Watanabe, Takaaki Hori, Shigeki Karita, Tomoki Hayashi, Jiro Nishitoba, Yuya Unno, Nelson Enrique Yalta Soplín, Jahn Heymann, Matthew Wiesner, Nanxin Chen, Adithya Renduchintala, Tsubasa Ochiai, “ESPnet: End-to-end speech processing toolkit,” in *ISCA Conference of the International Speech Communication Association (Interspeech)*, 2018, pp. 2207–2211.
- [17] Mirco Ravanelli, Titouan Parcollet, Peter Plantinga, Aku Rouhe, Samuele Cornell, Loren Lugosch, Cem Subakan, Nauman Dawalatabad, et al., “Speechbrain: A general-purpose speech toolkit,” in *arXiv preprint arXiv:2106.04624*, 2021.
- [18] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, Illia Polosukhin, “Attention is all you need,” in *Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2017, pp. 5998–6008.
- [19] Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, Ruoming Pang, “Conformer: Convolution-augmented transformer for speech recognition,” in *ISCA Conference of the International Speech Communication Association (Interspeech)*, 2020, pp. 5036–5040.
- [20] Mehryar Mohri, “Finite-state transducers in language and speech processing,” in *Computational Linguistics*, 1997, pp. 269–311.
- [21] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukáš Burget, Ondřej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlíček, Yanmin Qian, Petr Schwarz, Jan Silovsky, Georg Stemmer, Karel Veselý, “The Kaldi speech recognition toolkit,” in *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2011.
- [22] Zhehuai Chen, Wei Deng, Tao Xu, Kai Yu, “Phone synchronous decoding with ctc lattice,” in *ISCA Conference of the International Speech Communication Association (Interspeech)*, 2016, pp. 1923–1927.
- [23] Petar Aleksic, Mohammadreza Ghodsi, Assaf Michaely, Cyril Allauzen, Keith Hall, Brian Roark, David Rybach, Pedro Moreno, “Bringing contextual information to google speech recognition,” in *ISCA Conference of the International Speech Communication Association (Interspeech)*, 2015, pp. 468–472.
- [24] Yanzhang He, Tara N Sainath, Rohit Prabhavalkar, Ian McGraw, Raziel Alvarez, Ding Zhao, David Rybach, Anjali Kannan, et al., “Streaming end-to-end speech recognition for mobile devices,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 6381–6385.
- [25] Mahaveer Jain, Gil Keren, Jay Mahadeokar, Geoffrey Zweig, Florian Metzke, Yatharth Saraf, “Contextual RNN-T for open domain asr,” in *ISCA Conference of the International Speech Communication Association (Interspeech)*, 2020, pp. 11–15.
- [26] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, et al., “TensorFlow: A system for large-scale machine learning,” in *USENIX Symposium on Operating Systems Design and Implementation (OSDI)*, 2016, pp. 265–283.
- [27] Alex Aizman, Gavin Maltby, Thomas Breuel, “High performance i/o for large scale deep learning,” in *IEEE International Conference on Big Data (BigData)*, 2019, pp. 5965–5967.
- [28] Hui Bu, Jiayu Du, Xingyu Na, Bengu Wu, Hao Zheng, “Aishell-1: An open-source mandarin speech corpus and a speech recognition baseline,” in *IEEE Conference of the Oriental Chapter of the International Coordinating Committee on Speech Databases and Speech I/O Systems and Assessment (O-COCOSDA)*, 2017.
- [29] Jiayu Du, Xingyu Na, Xuechen Liu, Hui Bu, “Aishell-2: Transforming mandarin asr research into industrial scale,” in *arXiv preprint arXiv:1808.10583*, 2018.
- [30] Vassil Panayotov, Guoguo Chen, Daniel Povey, Sanjeev Khudanpur, “Librispeech: an asr corpus based on public domain audio books,” in *IEEE international conference on acoustics, speech and signal processing (ICASSP)*, 2015, pp. 5206–5210.
- [31] Guoguo Chen, Shuzhou Chai, Guan-Bo Wang, Jiayu Du, Wei-Qiang Zhang, Chao Weng, Dan Su, Daniel Povey, et al., “Gigaspeech: An evolving, multi-domain ASR corpus with 10,000 hours of transcribed audio,” in *ISCA Conference of the International Speech Communication Association (Interspeech)*, 2021, pp. 3670–3674.
- [32] Binbin Zhang, Hang Lv, Pengcheng Guo, Qijie Shao, Chao Yang, Lei Xie, Xin Xu, Hui Bu, Xiaoyu Chen, Chenchen Zeng, et al., “Wenetspeech: A 10000+ hours multi-domain mandarin corpus for speech recognition,” in *IEEE international conference on acoustics, speech and signal processing (ICASSP)*, 2022.
- [33] Daniel S Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D Cubuk, Quoc V Le, “SpecAugment: A simple data augmentation method for automatic speech recognition,” in *ISCA Conference of the International Speech Communication Association (Interspeech)*, 2019, pp. 2613–2617.