



Text-Only Domain Adaptation for End-to-End Speech Recognition through Down-Sampling Acoustic Representation

Jiayu Zhu^{1,3,‡}, Weinan Tong^{1,†}, Yaoxun Xu¹, Changhe Song^{1,2}, Zhiyong Wu^{1,2,5,*},
Zhao You^{3,*}, Dan Su³, Dong Yu⁴, Helen Meng⁵

¹Shenzhen International Graduate School, Tsinghua University, Shenzhen, China

²Peng Cheng Lab, Shenzhen, China

³Tencent AI Lab, Shenzhen, China ⁴Tencent AI Lab, Bellevue, WA, USA

⁵The Chinese University of Hong Kong, Hong Kong SAR, China

{zhu-jx21, twn21, xuyx22, sch19}@mails.tsinghua.edu.cn, zywu@sz.tsinghua.edu.cn,
{dennisyou, dansu, dyu}@tencent.com, hmmeng@se.cuhk.edu.hk

Abstract

Mapping two modalities, speech and text, into a shared representation space, is a research topic of using text-only data to improve end-to-end automatic speech recognition (ASR) performance in new domains. However, the length of speech representation and text representation is inconsistent. Although the previous method up-samples the text representation to align with acoustic modality, it may not match the expected actual duration. In this paper, we proposed novel representations match strategy through down-sampling acoustic representation to align with text modality. By introducing a continuous integrate-and-fire (CIF) module generating acoustic representations consistent with token length, our ASR model can learn unified representations from both modalities better, allowing for domain adaptation using text-only data of the target domain. Experiment results of new domain data demonstrate the effectiveness of the proposed method.

Index Terms: Speech Recognition, Text-Only, Continuous Integrate and Fire, Domain Adaptation

1. Introduction

Automatic speech recognition (ASR) is a technology that converts audio into text. In recent years, end-to-end (E2E) ASR has attracted much attention and made great progress. E2E ASR can convert audio to text using a single network model, simplifying the training and inference process. There are three main types of E2E ASR models: connectionist temporal classification (CTC) [1], recurrent neural network transducer (RNN-T) [2, 3], and attention-based encoder-decoder (AED) [4, 5]. Training with a large number of labeled data, the E2E ASR model has achieved excellent results. However, the performance still has a serious decline in new domains. While the E2E ASR model requires paired audio-text for training, it is expensive to acquire high-quality paired labeled data for new domains.

Even more, due to the training paradigm with paired audio-text data, it is difficult for E2E ASR to directly use text-only data for domain adaptation like the traditional hidden Markov model (HMM)-deep neural network (DNN) hybrid speech recognition model.

Considering that the acquisition of unpaired text is relatively more convenient and the amount of text data is large, many studies attempted to leverage text-only data to adapt the E2E ASR model in new domains. A common approach is to use an external language model. This external language model uses

a large amount of new domain unpaired text for training and fuses the E2E ASR model through the method of shallow fusion [6] or rescoring [7], to improve the recognition performance in the new domains. Another approach is to generate audio from large amounts of text in the new domains through a text-to-speech (TTS) synthesis model [8, 9], thus forming paired audio-text data that can be used for training E2E ASR models. However, this approach requires a reliable multi-speaker TTS model and the high computational cost of generating speech. Even worse, it still exists a mismatch between real and synthetic audio, which will affect the performance of speech recognition.

An alternative approach focuses on mapping the two modalities, speech and text, into shared representation spaces so that E2E ASR can be trained using paired audio-text or unpaired text [10, 11]. The main challenge of matching two modalities is that the length of speech representation and text representation is inconsistent, which makes it difficult for a model to learn a better-shared representation space. Although the previous approach aims to match the acoustic representation by replicating each text unit representation several times to increase the length of text representation or using a duration model to estimate phoneme and word alignments for each word in the transcript, it may not match the expected actual durations in practice [11, 12], which would affect the learning of shared representation space of two modalities. On the contrary, it will be more accurate to match the text length by down-sampling the acoustic representation [13]. We believe that modal matching with a more consistent length of representation will get better results.

In this paper, to explore the reasonable schemes of using text-only data for domain adaptation, we propose a new strategy to learn a shared representation space for the two modalities - speech and text, which can adapt E2E ASR to new domains more easily and effectively with text-only data. To solve the problem of the inconsistent length of acoustic representation and text representation, inspired by [13] and [14], we introduce a continuous integrate-and-fire (CIF) module to generate the acoustic representations consistent with token length. Furthermore, considering that syllable is more pronunciation-related than character, which can be more effectively matched with acoustic representation, and syllable is more robust than character which can reduce the impact of rare word or long-tail word, we explore using syllable instead of character to generate a shared representation similar to acoustic representation. Together with a transformer-based syllable encoder, our ASR model can learn unified representations from both modalities better. Experiment results for out-of-domain data show that the proposed text-only domain adaptation performs well.

[‡] Work done during internship at Tencent Inc.

[†] Equal contribution. * Corresponding authors.

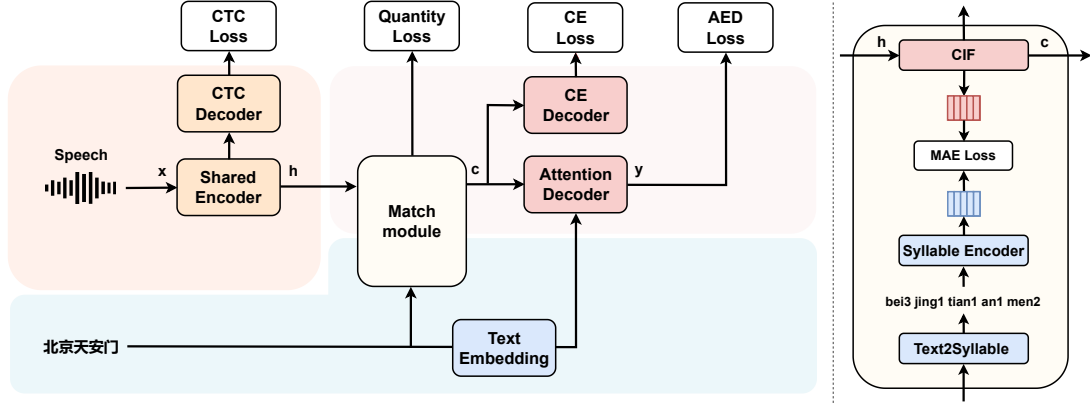


Figure 1: Proposed architecture of our E2E ASR model to learn unified representation from two modalities of speech and text. Left: the overall network; Right: the detailed architecture of the Match module

2. Methodology

In this section, we will review the architecture of our proposed method, which aims to apply a novel scheme to improve the ability to use text-only data for the E2E ASR model.

2.1. Architecture

Our proposed E2E ASR model is based on the one presented in WeNet [15], which uses both CTC and Attention-based Encoder-Decoder (AED) losses during training to speed convergence and increase the AED model’s robustness. As depicted in Figure 1, the proposed ASR model mainly contains five parts, Attention-based *Shared Encoder*, *CTC Decoder*, *CE Decoder*, *Attention Decoder* and *Match Module*. The *Shared Encoder* mainly consists of the Conformer [16] blocks. The *CTC Decoder* and *CE Decoder* consist of a linear layer and a log softmax layer. The *Attention Decoder* mainly consists of multiple transformer [17] blocks. The *Match Module* mainly consists of a CIF module and a transformer-based *Syllable Encoder*. Given that the module structure of *CTC Decoder* and *CE Decoder* is relatively simple, we mainly introduce other modules.

2.1.1. The Shared Encoder

The *Shared Encoder* consists of a convolution subsampling layer containing two convolutional layers with stride 2 for downsampling, a linear projection layer, and a positional encoding layer, followed by multiple Conformer encoder layers. The *Shared Encoder* transforms a T -length speech feature sequence $\mathbf{x} = (x_1, \dots, x_T)$ to a L -length intermediate representation $\mathbf{h} = (h_1, \dots, h_L)$, where $L \leq T$ owing to downsampling.

2.1.2. The Match Module

As depicted in Figure 1, the *Match Module* mainly consists of a CIF module and a transformer-based *Syllable Encoder*. The CIF module consists of a 1-dimensional convolution layer and a linear layer to achieve a soft monotonic alignment. The CIF encodes the *Shared Encoder*’s outputs $\mathbf{h} = (h_1, \dots, h_L)$ to predict the corresponding float weights $\mathbf{a} = (a_1, \dots, a_L)$ ranging from 0 to 1. We then carry out a weighted sum between \mathbf{h} and \mathbf{a} until the accumulated weight reaches a threshold which means reaching an acoustic boundary and generating a new integrated embedding. The threshold is recommended to be 1.0

in [13]. In this way, CIF outputs high-level acoustic sequence $\mathbf{c} = (c_1, \dots, c_i, \dots, c_I)$, which is consistent with the length of the text representation. On the other hand, a transformer-based *Syllable Encoder* takes syllable embedding as inputs and output text representations. Considering that the acoustic sequence \mathbf{c} is strictly aligned with the text sequence during training, we map the two modalities into a shared space with a mean absolute error (MAE) training objective.

2.1.3. The Attention Decoder

The *Attention Decoder* consists of a positional encoding layer, multiple transformer decoder layers, and a linear projection layer. Given \mathbf{c} and previously emitted character outputs $\mathbf{y}_{0:i-1} = (y_0, \dots, y_{i-1})$, the attention decoder predicts the next character y_i .

2.2. The Overall Training Pipeline

In this subsection, we describe the overall modality-matched training process using available paired speech-text data and the text-only training process using unpaired text-only data. Our method aims to utilize a large number of unpaired text data without modifying the model. To do so, we allow the model to be trained on either paired audio-text data or unpaired text data.

2.2.1. Modality Matched Training

To solve the actually expected duration mismatch of text units in previous approaches, we have changed a perspective to down-sample the acoustic representation instead of up-sampling text representation. As shown in Figure 1, speech is first transformed into a long acoustic representation by the *Shared Encoder*. Then, in the *Match Module*, a CIF module down-samples the long acoustic representation to a relatively short acoustic representation with the same length as the text representation generated by the *Syllable Encoder*. Although we use syllables as model unit to get a hidden representation, we still use the word “text representation” to refer to it, because the syllables come from text through a text2syllable process which mainly uses an open-source Chinese character to pinyin tool python-pinyin¹.

¹<https://pypi.org/project/python-pinyin/>

2.2.2. Text-only Training

The *Attention Decoder* in our E2E ASR model is a transformer-based decoder that mainly consists of a self-attention module, a cross-attention module, and a feed-forward module. The self-attention module allows the decoder to model the content text between token sequences [18]. Therefore, the *Attention Decoder* can interpret as an internal language in E2E ASR [19, 20, 21]. However, the cross-attention module makes the decoder dependent on the acoustic encoder output and thus can not be separately trained on text-only data, which also makes the internal language model hard to update for domain adaptation. And the approach of shared representation space of speech and text would be an effective way to use text-only data, in which the text representation plays a similar role to acoustic representation.

After the paired speech-text data training process, the *Syllable Encoder* can transform syllable embedding to a hidden representation, which shares a representation space with acoustic representation. Therefore, we can utilize several text-only data to train the E2E ASR model. The detailed process is as depicted in Figure 2, we first get a hidden representation from *Syllable Encoder* instead of CIF to replace the absent acoustic representation in the text-only process. And then we only train the *Attention Decoder* while other module parameters are fixed to not update. The *Attention Decoder* is based on transformer architecture and it does not require changing the original objective function. Considering that the internal language model has some parts related to acoustic modeling, it is not a real language model. In order to prevent the decoder from affecting the acoustic modeling part and suppress catastrophic forgetting during text-only training, we randomly use audio-text data of the source domain in training to ensure that the decoder can perform ASR tasks when conditioned on audio features [22, 23]. Therefore, in our novel modality-matched schemes, we can implement the text-only training process more simply and effectively.

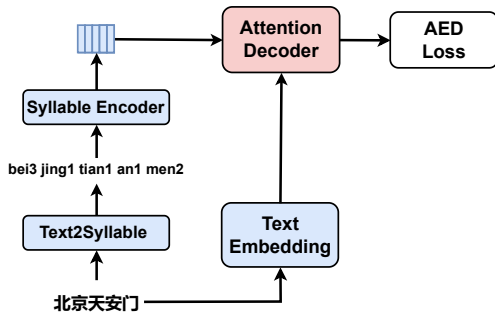


Figure 2: Overview of text-only training process.

2.3. Loss Function

We used five loss functions to train our model, namely the CTC, Quantity, CE, AED, and MAE losses, where the CTC, Quantity, and CE loss just like that in [13]. The types are jointly trained, as follows:

$$\mathcal{L} = \alpha \mathcal{L}_{CTC} + \beta \mathcal{L}_{QUA} + \gamma \mathcal{L}_{CE} + \lambda \mathcal{L}_{AED} + \delta \mathcal{L}_{MAE} \quad (1)$$

Where $\alpha, \beta, \gamma, \lambda$ and δ are tunable parameters. In the experiment, we set α and γ to 0.5, and other parameters to 1.

In text-only training, AED loss is the only one used:

$$\mathcal{L}_{text} = \mathcal{L}_{AED} \quad (2)$$

3. Experiments

3.1. Datasets

In this paper, we train our proposed E2E ASR on public Mandarin Aishell-1 [24] datasets. The Aishell-1 corpus consists of 178 hours of labeled speech collected from 400 speakers. The content of the datasets covers 5 domains including Finance, Science and Technology, Sports, Entertainment, and News. To compare the domain adaptation ability of ASR in the text domain while minimizing the influence of differences in the acoustic environment, we chose another public Mandarin dataset Aishell-2 [25] that has a similar acoustic environment for sound recording but the corresponding text contents cover different text domains. The Aishell-2 corpus consists of 1000 hours of labeled speech collected from 1991 speakers. The content of Aishell-2 correspondent-only domains of voice commands, digital sequence, places of interest, entertainment, finance, technology, sports, English spellings, and free speaking without specific topics. Furthermore, we also conducted further experiments on different domains on the WenetSpeech [26], which is a multi-domain Mandarin corpus consisting of high-quality labeled speech but a relatively more complex acoustic environment than Aishell-1. We use the Aishell-1 training set for training and the development set for early stopping.

3.2. Experimental Setup

For all experiments, we use the open-source WeNet toolkit [15] to build our proposed ASR model. And we used the default values in the WeNet for the main parameters which have been validated by the WeNet contributor. The input features are 80-dimensional log Mel-filterbank (FBank) computed on a 25ms window with a 10ms shift. We use SpecAugment [27] and speed perturb for data augmentation. We choose 4233 characters (including $\langle \text{blank} \rangle$, $\langle \text{unk} \rangle$, $\langle \text{sos/eos} \rangle$ labels) as model units for Aishell-1.

Following the WeNet recipe [15], we construct the base model using 12 Conformer blocks in the *Shared Encoder*, 6 transformer blocks in the *Attention Decoder* and 4 transformer blocks in the *Syllable Encoder*. We employ $h = 4$ parallel attention heads in both the Conformer block and transformer block. For every layer, we use $d_k = d_v = d_{model}/h = 64$, $d_{ffn} = 2048$.

We train the model with Adam Optimizer [17] for at most 240 epochs with 12 batches. And $learning\ rate = 0.002$, $warm\ up = 25000$, and gradient clipping at 5.0. Additionally, during training, we employ the gradient accumulation method, in which the gradients are modified every four batches. Moreover, we employ label smoothing of value $\epsilon_{ls} = 0.1$ and a dropout rate of $P_{drop} = 0.1$. We set the weight λ of the CTC branch during joint training to 0.3. We also train the n-gram language model with new domains of text-only data follow by the WeNet recipe. During joint decoding, we set the CTC-weight λ to 0.5. To avoid overfitting, we averaged the 30 best model parameters in the development dataset.

4. Results

The performance of the models is evaluated based on character error rates (CER). Our experimental results are mainly based on the attention-rescore two-step decoding method.

4.1. Main Result

Our method is evaluated on the Aishell-1 dataset. We compare the proposed ASR model with other models in the literature. As shown in Table 1, the proposed ASR model achieves comparable performance with a series of state-of-the-art approaches.

Table 1: *Main results on Aishell-1 (CER)*

Model	dev	test
EspNet Conformer [28]	4.5	4.9
WeNet Conformer [15]	-	4.6
Branchformer [29]	4.2	4.4
Blockformer [30]	-	4.4
CIF-based model [13]	4.5	4.9
+ CE Decoder	4.2	4.7
+CE Decoder and Match module (Proposed)	4.1	4.5

4.2. Domain Adaption

In order to prove the effectiveness of our text-only method in domain adaptation, we also compare the results on the Aishell-2 test and dev datasets, which have a similar acoustic environment with Aishell-1 but cover different text domains. We use the text data of the Aishell-2 training dataset for text-only training. As shown in Table 2, after text-only training, the performance of the ASR model in a new domain is significantly improved. The LM is trained with Aishell-2 text data from the training set. And text-only refers to the model after text-only training.

Table 2: *Comparison of the performance after text-only domain adaption on Aishell-2 (CER)*

Model	Aishell-2 dev	Aishell-2 test
Proposed Model	11.7	11.6
+ LM	11.5	11.4
Text-Only	11.2	11.0

In addition, to further illustrate the validity of our approach in more difficult text domains and more complex acoustic environments, we conduct further experiments on three domains of the WenetSpeech dataset. As shown in Table 3, the proposed text-only method can also improve recognition performance. However, with the increasing complexity of data, the recognition performance of the proposed model is poor, and the performance of text-only is not obvious. Our proposed model is trained on a relatively small and quiet dataset, and its ability for acoustic modeling is not strong enough, so the performance is poor in a dataset with a complex acoustic environment.

Table 3: *Comparison of the performance after text-only domain adaption on WeNetSpeech (CER)*

Model	audiobook	interview	drama
Proposed Model	15.0	37.9	56.6
Text-Only	14.2	37.9	56.4

4.3. Effects of the Epoch of Text-Only

We further study the out-of-domain performance changed with the training epoch. As shown in Figure 3, in the beginning, with the increase of training epochs, the ASR performance was

significantly improved. When the training epoch reaches nearly 40, the text-only performance will be achieved. This shows that the text-only method can achieve domain adaptation quickly.

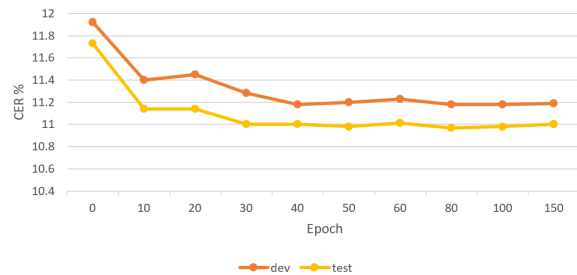


Figure 3: *Text-only performance changes with epoch.*

4.4. Analysis on Different Model Unit for Text-Only

Furthermore, we also study the performance of the different model units of character and syllable in our text-only. As shown in Table 4, using syllables as model unit achieves a better performance than using the character in the proposed model and is more effective in text-only domain adaptation. On the one hand, the syllable modeling units are more pronunciation-related than character, which can be more effectively matched with acoustic representation. On the other hand, the rare character or long-tail character may be difficult to fully model in a modality matching.

Table 4: *Comparison of the performance of different model unit in text-only domain adaption on Aishell-2 (CER)*

Model	Character dev / test	Syllable dev / test
Proposed Model	11.8 / 11.7	11.7 / 11.6
+ LM	11.6 / 11.5	11.5 / 11.4
Text-Only	12.0 / 11.9	11.2 / 11.0

5. Conclusions

In this paper, we proposed a novel representations match module through down-sampling acoustic representation to align with text modality. By introducing a continuous integrate-and-fire (CIF) module generating acoustic representations consistent with token length and using pronunciation-related model unit syllable matching acoustic representation effectively, our ASR model can learn unified representations from both modalities better, allowing for domain adaptation using text-only data of the target domain. Experimental comparisons for out-of-domain settings demonstrate that the proposed text-only domain adaptation achieves a good performance. In the future, we will further explore the performance of different model units with large-scale datasets and verify the performance of our method on the English datasets.

6. Acknowledgements

This work is supported by Shenzhen Science and Technology Program (WDZC20200818121348001, JCYJ20220818101014 030), the Major Key Project of PCL (PCL2021A06, PCL2022 D01) and AMiner.Shenzhen SciBrain fund.

7. References

- [1] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, “Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks,” in *ACM International Conference on Machine Learning (ICML)*, 2006, pp. 369–376.
- [2] A. Graves, “Sequence transduction with recurrent neural networks,” *arXiv preprint arXiv:1211.3711*, 2012.
- [3] S. Wang, P. Zhou, W. Chen, J. Jia, and L. Xie, “Exploring rnn-transducer for chinese speech recognition,” in *Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2019, pp. 1364–1369.
- [4] J. Chorowski, D. Bahdanau, K. Cho, and Y. Bengio, “End-to-end continuous speech recognition using attention-based recurrent nn: First results,” in *Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2014.
- [5] W. Chan, N. Jaitly, Q. Le, and O. 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.
- [6] 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 *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 1–5828.
- [7] T. N. Sainath, Y. He, A. Narayanan, R. Botros, R. Pang, D. Rybach, C. Allauzen, E. Variiani, J. Qin, Q.-N. Le-The, S.-Y. Chang, B. Li, A. Gulati, J. Yu, C.-C. Chiu, D. Caseiro, W. Li, Q. Liang, and P. Rondon, “An Efficient Streaming Non-Recurrent On-Device End-to-End Model with Improvements to Rare-Word Modeling,” in *ISCA Conference of the International Speech Communication Association (Interspeech)*, 2021, pp. 1777–1781.
- [8] Z. Chen, Y. Zhang, A. Rosenberg, B. Ramabhadran, G. Wang, and P. Moreno, “Injecting text in self-supervised speech pretraining,” in *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2021, pp. 251–258.
- [9] Z. Chen, Y. Zhang, A. Rosenberg, B. Ramabhadran, P. Moreno, and G. Wang, “Tts4pretrain 2.0: Advancing the use of text and speech in asr pretraining with consistency and contrastive losses,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 7677–7681.
- [10] Z. Chen, Y. Zhang, A. Rosenberg, B. Ramabhadran, P. J. Moreno, A. Bapna, and H. Zen, “Maestro: Matched speech text representations through modality matching,” in *ISCA Conference of the International Speech Communication Association (Interspeech)*, 2022, pp. 4093–4097.
- [11] T. N. Sainath, R. Prabhavalkar, A. Bapna, Y. Zhang, Z. Huo, Z. Chen, B. Li, W. Wang, and T. Strohm, “Joist: A joint speech and text streaming model for asr,” in *IEEE Spoken Language Technology Workshop (SLT)*, 2023, pp. 52–59.
- [12] S. Thomas, H.-K. J. Kuo, B. Kingsbury, and G. Saon, “Towards reducing the need for speech training data to build spoken language understanding systems,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 7932–7936.
- [13] L. Dong and B. Xu, “Cif: Continuous integrate-and-fire for end-to-end speech recognition,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 6079–6083.
- [14] Z. Gao, S. Zhang, I. McLoughlin, and Z. Yan, “Paraformer: Fast and accurate parallel transformer for non-autoregressive end-to-end speech recognition,” in *ISCA Conference of the International Speech Communication Association (Interspeech)*, 2022, pp. 2063–2067.
- [15] 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,” in *ISCA Conference of the International Speech Communication Association (Interspeech)*, 2021, pp. 4054–4058.
- [16] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, and R. Pang, “Conformer: Convolution-augmented transformer for speech recognition,” in *ISCA Conference of the International Speech Communication Association (Interspeech)*, 2020, pp. 5036–5040.
- [17] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” in *Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2017, pp. 5998–6008.
- [18] K. Deng, S. Cao, Y. Zhang, and L. Ma, “Improving hybrid ctc/attention end-to-end speech recognition with pretrained acoustic and language model,” in *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2021, pp. 76–82.
- [19] Z. Meng, S. Parthasarathy, E. Sun, Y. Gaur, N. Kanda, L. Lu, X. Chen, R. Zhao, J. Li, and Y. Gong, “Internal language model estimation for domain-adaptive end-to-end speech recognition,” in *IEEE Spoken Language Technology Workshop (SLT)*, 2021, pp. 243–250.
- [20] Z. Meng, N. Kanda, Y. Gaur, S. Parthasarathy, E. Sun, L. Lu, X. Chen, J. Li, and Y. Gong, “Internal language model training for domain-adaptive end-to-end speech recognition,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 7338–7342.
- [21] M. ZeinElddeen, A. Glushko, W. Michel, A. Zeyer, R. Schlüter, and H. Ney, “Investigating Methods to Improve Language Model Integration for Attention-Based Encoder-Decoder ASR Models,” in *ISCA Conference of the International Speech Communication Association (Interspeech)*, 2021, pp. 2856–2860.
- [22] P. Wang, T. N. Sainath, and R. J. Weiss, “Multitask training with text data for end-to-end speech recognition,” in *ISCA Conference of the International Speech Communication Association (Interspeech)*, 2021, pp. 2566–2570.
- [23] S. Kim, K. Li, L. Kabela, R. Huang, J. Zhu, O. Kalinli, and D. Le, “Joint audio/text training for transformer rescorer of streaming speech recognition,” in *Findings of the Association for Computational Linguistics: EMNLP*, 2022, pp. 5717–5722.
- [24] 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 *IEEE 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.
- [25] J. Du, X. Na, X. Liu, and H. Bu, “Aishell-2: Transforming mandarin asr research into industrial scale,” *arXiv preprint arXiv:1808.10583*, 2018.
- [26] B. Zhang, H. Lv, P. Guo, Q. Shao, C. Yang, L. Xie, X. Xu, H. Bu, X. Chen, C. 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, pp. 6182–6186.
- [27] 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,” in *ISCA Conference of the International Speech Communication Association (Interspeech)*, 2019, pp. 2613–2617.
- [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 *ISCA Conference of the International Speech Communication Association (Interspeech)*, 2018, pp. 2207–2211.
- [29] 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 *ACM International Conference on Machine Learning (ICML)*, 2022, pp. 17 627–17 643.
- [30] X. Ren, H. Zhu, L. Wei, M. Wu, and J. Hao, “Improving mandarin speech recognition with block-augmented transformer,” *arXiv preprint arXiv:2207.11697*, 2022.