ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-7281-6327-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICASSP49357.2023.10095937

FACTORIZED AED: FACTORIZED ATTENTION-BASED ENCODER-DECODER FOR TEXT-ONLY DOMAIN ADAPTIVE ASR

Xun Gong, Wei Wang, Hang Shao, Xie Chen, Yanmin Qian†

MoE Key Lab of Artificial Intelligence, AI Institute
X-LANCE Lab, Department of Computer Science and Engineering
Shanghai Jiao Tong University, Shanghai, China

ABSTRACT

End-to-end automatic speech recognition (ASR) systems have gained popularity given their simplified architecture and promising results. However, text-only domain adaptation remains a big challenge for E2E systems. Text-to-speech (TTS) based approaches fine-tune ASR models by synthesized speech with an auxiliary TTS model, thus increase deployment costs. Language model (LM) fusion based approaches can achieve good performance but are sensitive to interpolation parameters. In order to factorize out the language component in the AED model, we propose the factorized attention-based encoder-decoder (Factorized AED) model whose decoder takes as input the posterior probabilities of a jointly trained LM. Moreover, in the context of domain adaptation, the domain specific LM serves as a plug-and-play component for a well-trained factorized AED model. In-domain experiments on LibriSpeech and out-of-domain experiments adapting from LibriSpeech to a variety of domains in GigaSpeech are conducted to validate the effectiveness of our proposed methods. Results show 20% / 24% relative word error rate (WER) reduction for LibriSpeech test sets and 8∼34% relative WER reduction for 8 GigaSpeech target domains test sets compared to the AED baseline.

***Index Terms*—** text-only, domain adaptation, factorized AED, end-to-end speech recognition

1. INTRODUCTION

End-to-end automatic speech recognition (ASR) systems such as connectionist temporal classification (CTC) [1], attention-based encoder-decoder (AED) [2, 3, 4] and recurrent neural network transducer (RNN-T) [5] simplify the whole ASR procedure and bring a huge development in recognition accuracy [6]. However, the performance of end-to-end ASR degrades dramatically when there is a domain mismatch, which is a huge challenge compared with hybrid systems. Domain adaptation methods have been explored a lot in recent years [7, 8, 9, 10]. Conventional domain adaptation such as regularization methods [11], teacher-student learning [12], adversarial learning [13], domain vector [14], adapter [9] and mixture of experts [10] and so on usually require speech-text pairs for the target domain. However, it is known to all that collecting a large amount of speech-text matching data from the target domain is difficult, so text-only adaptive methods have been widely proposed and studied. For the hybrid DNN-HMM system, the acoustic model and language model (LM) are optimized independently, so text-only data can be easily used to adapt the LM part. In contrast, there are not many efficient methods recently in end-to-end systems.

One intuitive solution to the text-only adaptation is to utilize the text-to-speech (TTS) synthesized speech, and then fine-tune the ASR model with the synthesized speech. Although TTS-based approaches have shown promise in [15, 16, 17, 18, 19], they can be computationally expensive (require well-trained TTS models, new synthesized speech data and fine-tuned ASR models).

LM fusion is another possible adaptation way using text-only data, where a target LM is trained from target domain text. The simplest LM fusion is shallow fusion [20], which jointly computes the target LM score and the end-to-end model score during the beam search stage. To eliminate negative effects from source domain text, the density ratio [21] method is proposed, which subtracts the source LM score besides shallow fusion. However, the interpolation weight is task-dependent and needs to be tuned on dev data. Further more, structural LM fusion methods, such as deep fusion [22], cold fusion [23] and component fusion [24] learn the combination of E2E ASR system and an external LM. However, these methods have limited improvements and need multiple adaptation steps before usage. In recent years, researchers find that the decoders in AED and RNN-T systems play an important role in text modeling. Internal LM estimation [25] and hybrid autoregressive transducer [26] which try to make the decoder behave like an LM, however, is somehow against the design philosophy of transducer [27] and transformer [2].

We propose a factorized attention-based encoder-decoder (*Factorized AED*) model to solve the text-only domain adaptation problem without destroying the decoder design, inspired by factorized neural transducer [28], which factorizes the blank and vocabulary prediction part, and the vocabulary predictor works as a standalone LM in the transducer manner. The proposed method needs an internal LM besides the AED model. Then the LM posterior probability (LM information) is incorporated into the factorized decoder with different integration methods. LM information can be used as the text-part input of the factorized decoder, which is named as *Factorized Input*. Also, we can integrate LM information into the output projection layer of the decoder (*Factorized Output*), which is similar to component fusion but in the AED manner. Moreover, we integrate the LM information using contextual source-attention (*Factorized Attention*) inside each transformer layer to achieve an intensive adaptation. As a standalone LM is used in our model, different language model adaptation methods can be applied to do fast text-only domain adaptation, which is similar to the hybrid system. Our experiments show that the *Factorized Attention* achieves the best performance. The proposed method is validated on in-domain scenario (LibriSpeech) and out-of-domain adaptation scenario (Gi-

† corresponding author

gaSpeech) and results show that the proposed method consistently outperforms the baseline and shallow fusion method. Furthermore, the performance can be boosted by combining the proposed factorized AED and shallow fusion.

2. RELATED WORK

2.1. Joint CTC and Attention-based Encoder-Decoder ASR

Basically, E2E ASR models map speech $\boldsymbol{X} = [\boldsymbol{x}_1, \cdots, \boldsymbol{x}_T]^\intercal, \boldsymbol{x}_t \in \mathbb{R}^F$ to a token sequence $\boldsymbol{y} = [y_1, \cdots, y_L]^\intercal, y_l \in \mathcal{U}$, where F is the speech feature dimension and $\mathcal{U}$ denotes the vocabulary set.

After years of development, the attention-based encoder-decoder (AED) model has gradually formed a transformer-based structure [2, 4]. The transformer encoder is composed of basic transformer layers, and finally get output $\boldsymbol{H}_{\text{Enc}}$. Then for the decoder, one more multi-head context attention combines the encoder output $\boldsymbol{H}_{\text{Enc}}$ is combined with $\boldsymbol{y}_{<l}$ to estimate the whole posterior probability $p(\boldsymbol{y}|\boldsymbol{H}_{\text{Enc}})$:

$$\boldsymbol{H}_{\text{Enc}} = \text{Encoder}(\boldsymbol{X}), \quad \boldsymbol{H}_{\text{Dec}} = \text{Decoder}(\boldsymbol{y}_{<l}, \boldsymbol{H}_{\text{Enc}}), \tag{1}$$

$$p_{\text{AED}}(y_l|\boldsymbol{y}_{<l}, \boldsymbol{H}_{\text{Enc}}) = \text{Softmax} \cdot \text{Linear} \cdot \text{ReLU}(\boldsymbol{H}_{\text{Dec}}[l]), \tag{2}$$

$$p_{\text{AED}}(\boldsymbol{y}|\boldsymbol{H}_{\text{Enc}}) = p_{\text{AED}}(y_1) \prod_{l=2}^{L} p_{\text{AED}}(y_l|\boldsymbol{y}_{<l}). \tag{3}$$

Connectionist temporal classification (CTC) [1] acts as a good supplement to AED model by introducing a many-to-one function η from the frame-level alignment $\boldsymbol{Z} = [z_1, \cdots, z_T]^\intercal, z_t \in \mathcal{U} \cup \{\texttt{<B>}\}$ to the token sequence $\boldsymbol{y}$, by merging same labels and removing the token <B> in $\boldsymbol{Z}$. The sequence probability is represented as:

$$p_{\text{CTC}}(\boldsymbol{y}|\boldsymbol{H}_{\text{Enc}}) = \sum_{\boldsymbol{Z} \in \Psi^{-1}(\boldsymbol{y})} \prod_{t} p_{\text{CTC}}(z_t|\boldsymbol{H}_{\text{Enc}}), \tag{4}$$

where η is a many-to-one function from $\boldsymbol{Z}$ to $\boldsymbol{y}$.

The joint CTC/attention architecture [3] is widely used in modern architectures with a multi-task learning loss function:

$$\mathcal{L}_{\text{JCA}} = -\alpha \log p_{\text{CTC}}(\boldsymbol{y}|\boldsymbol{X}) - (1-\alpha) \log p_{\text{AED}}(\boldsymbol{y}|\boldsymbol{X}), \tag{5}$$

where $\alpha \in [0, 1]$ is a hyper-parameter.

2.2. Language Model Fusion Methods

The shallow fusion (SF) [20] method is developed based on beam search decoding of the E2E model, which uses a score combination with additional target LM score, and the density ratio (DR) [21] method subtract the source LM score:

$$s(y_l) = s_{\text{JCA}}(y_l) + \gamma s_{\text{tLM}}(y_l) - \beta s_{\text{sLM}}(y_l), \tag{6}$$

where $\beta, \gamma > 0$ are hyper-parameters, sLM and tLM denotes source/target LM.

Deep fusion (DF) [22] take one step forward by fusing external LM with pre-trained ASR output, according to the following procedure:

$$p(y_l) = \text{Softmax} \cdot \text{Linear}([\boldsymbol{H}_{Dec}[l]; \phi]), \tag{7}$$

$$\phi_{DF}(\boldsymbol{h}) = \sigma(\text{Linear}(\boldsymbol{h})) * \boldsymbol{h}, \tag{8}$$

where ϕ_{DF} is a coarse gating function and $\boldsymbol{h} = \boldsymbol{H}_{LM}[l]$. By fixing ASR and LM parameters, DF achieves a lower cost and expedites the convergence process.

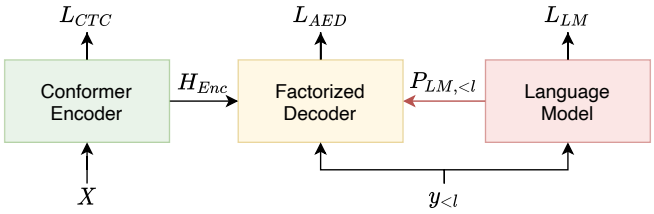

Fig. 1. The proposed factorized attention-based encoder-decoder (Factorized AED) architecture contains factorized decoder and a source language model.

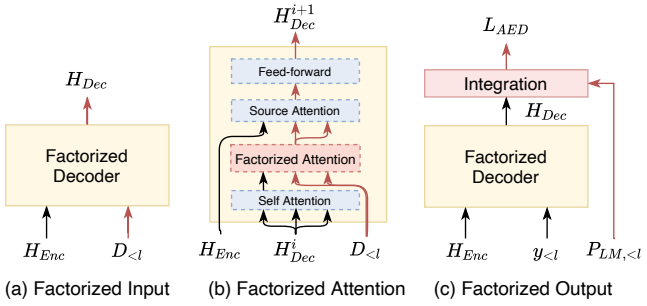

Fig. 2. Three different factorization architectures.

Cold fusion (CF) [23] and component fusion [24] on the other side, extend DF using another ϕ function as below:

$$\phi_{CF} = \sigma(\text{Linear}(\boldsymbol{H}_{Dec}[l]; \boldsymbol{H}_{LM}[l])) \odot \boldsymbol{H}_{LM}[l]. \tag{9}$$

The difference with DF is that CF uses LM logits and trains the ASR model from scratch with a pre-trained LM.

3. FACTORIZED ATTENTION-BASED ENCODER-DECODER SPEECH RECOGNITION

3.1. Factorization Architectures

The proposed factorized attention-based encoder-decoder (factorized AED) is shown in Figure 1. At first, a transformer language model is added to the network. The LM contains a vanilla transformer encoder and an output linear layer to match the token size.

$$p_{\text{LM}}(y_l|\boldsymbol{y}_{<l}) = \text{LM}(\boldsymbol{y}_{<l}) = \text{Softmax} \cdot \text{Transformer}(\boldsymbol{y}_{<l}). \tag{10}$$

The language information can be used in later factorizing integrations. Typical information features can be implemented as the hidden states $\boldsymbol{H}_{\text{LM}}$ or the posterior probability p_{LM}. To make the LM a replaceable component, we use the posterior probabilities instead of the hidden states of LM to integrate into the AED model as $\boldsymbol{P}_{\text{LM},<l}$, where $\geq l$ positions are masked. An auxiliary linear layer maps $\mathcal{U}$ to the feature dimension as below: $\boldsymbol{D}_{<l} = \text{Linear}(\boldsymbol{P}_{\text{LM},<l})$.

To factorize out the language part from the AED model, we propose three different approaches to integrate the language context with the transformer decoder shown in Figure 2.

Factorized Input: Shown in Figure 2(a), the token input of the decoder $\boldsymbol{y}_{<l}$ can be replaced with $\boldsymbol{D}_{<l}$:

$$\boldsymbol{H}_{\text{Dec}} = \text{Transformer Decoder}(\boldsymbol{D}_{<l}, \boldsymbol{H}_{\text{Enc}}), \tag{11}$$

where no more positional encoding or embedding layer is needed.

Factorized Attention: Another integration method Shown in Figure 2(b) is to behave as a context-attention module:

$$\boldsymbol{H}_{\text{Dec}\,i}' = \text{Factorized-Attention}(\boldsymbol{H}_{\text{Dec}\,i}, \boldsymbol{D}_{<l}, \boldsymbol{D}_{<l}), \tag{12}$$

where $\boldsymbol{H}_{\text{Dec}\,i}$ is the i-th transformer layer's hidden states, $\boldsymbol{D}_{<l}$ is used as key and value and $\boldsymbol{H}_{\text{Dec}\,i}$ is used as query. As mentioned before, there are different positions to inject the factorized attention module inside the three modules of the decoder layer: before self-attention (pos 0), before source-attention (pos 1), or before feed-forward layer (pos 2) to explore better performance.

Factorized Output: Shown in Figure 2(c), Also the factorization can be done after the transformer module of the AED decoder, and there are two positions to integrate the information. If the LM information is injected after the linear layer, then this method is similar to shallow fusion, but the joint decoding score is optimized during ASR training: $p'_{AED} = p_{AED} + \beta' p_{LM}$, where β' is a train-able parameter. If the LM information is injected before the linear layer, then there is an auxiliary transformation matrix from vocabulary size to the attention dimension, which is similar to cold/component fusion: $p'_{AED} = \text{Softmax} \cdot \text{Linear}([\boldsymbol{H}_{\text{Dec}}; \boldsymbol{D}_{LM,<l}])$.

3.2. Training and Adaptation Strategies

As the LM part is factorized out by the above methods, the LM part is jointly trained with the ASR model from scratch using Equation 5:

$$\mathcal{L}_{\text{Factorized}} = \mathcal{L}_{\text{JCA}} - \eta \log p_{\text{LM}}(\boldsymbol{y}), \tag{13}$$

where η is a hyper-parameter, and the LM loss is the standard token-level cross-entropy loss.

Although we can use a pre-trained source LM similar to deep / cold fusion, we find that joint training is more stable. The main reason is that the source LM has a little mismatch if it is pre-trained by a larger source text corpus. Another reason is that joint training makes the LM decouples from the ASR part synchronously.

The factorized architecture can be applied to improve the in-domain ASR performance by replacing the source LM with another source LM, which is trained in a larger source text corpus to improve the recognition accuracy. During text-only out-of-domain adaptation, the source LM can be replaced by any other language model type, such as n-gram, or recurrent neural network LM. Meanwhile, we could apply popular language model adaptation techniques on the target domain transcriptions. The important is that the new LM is fully plug-and-play, which means the adaptation step is to only train a new LM with large in-domain text or target domain text without fine-tuning the AED module.

4. EXPERIMENTS

4.1. Experimental Setups

Our experiments are conducted on in-domain and out-of-domain scenarios. The source domain data is LibriSpeech [29]. Different target domains in GigaSpeech [30] are adapted. LibriSpeech has around 960 hours of audiobook speech train set with test-clean / other used for testing. For experiments on the in-domain scenario, an auxiliary text corpus is used to train a new in-domain LM which contains ∼803 million words. GigaSpeech is a recently published multi-domain ASR corpus comprised of 10,000 hours of transcribed speech. In this work, we use the youtube and podcast partition of the GigaSpeech train-XL subset. 4 out of 5 different domains in youtube (education, news, people, science) and 4 out of 5 different domains in podcast (arts, health, people, science) are selected and shown in Table 1 [1]. Around 5 hours of dev set and 10 hours of test set are split and evaluated for the following experiments.

For acoustic feature extraction, 80-dimensional mel filterbank features are extracted with global level cepstral mean and variance normalization, frame length equal to 25ms and frame shift equal to 10ms. When it comes to data augmentation, standard SpecAugment [31] is used. 5,000 sentence pieces [32] are trained using LibriSpeech 960 hours paired text. The attention-based encoder-decoder (AED) baseline follows the basic settings of ESPnet recipe [33]. The subsampling layer is a two-layer convolution neural network, whose down-sampling rate is 4. The encoder has 12 conformer layers, in which the inner size of the feed-forward layer is 2,048, and the attention dimension is 512 with 8 heads. The decoder has 6 transformer layers with the same attention setup. The hyper-parameters are set as $\alpha = 0.3, \eta = 0.4$, and $\beta = 0.6$ for shallow fusion method for most sets with beam size equal to 20.

During adaptation, the target language models are trained from target domain text shown in Table 1 and the large LibriSpeech language model is trained from the auxiliary text corpus which has ∼803 million words for in-domain scenarios, respectively. All language models are based on standard transformer encoder with 18 layers, and each layer has 8 heads with 512 self-attention dimension.

4.2. In-domain Situation

In this section, we explore and validate the best factorized AED architecture on the LibriSpeech evaluation test sets for in-domain case by using external LM (ex-LM=✓). At the very beginning, we try to use log posterior probability $\log p_{\text{LM}}$ as the LM information. However, such log probability hurts the transformer decoder because the rare tokens get low probability, which has a large abs value. Then the transformer decoder tries to learn the distribution of rare tokens and those tokens' probability is unstable in log space for different LMs where the LM part is not replaceable.

Shown in Table 2, firstly, we compare the recognition accuracy when the factorized AED model with different factorized methods mentioned in Section 3. The replaced external LM perplexity (PPL) is 31.1 on averaging test-clean / other sets, while the internal LMs' PPL is shown in the Table 2.

Results show that factorized input has a better performance compared with two implementations of factorized output by 0.1∼0.2%. This is because the transformer decoder can learn more from the LM information and provide stronger modeling capability compared with only linear in the factorized output method. Also, we find that factorized output (after) performs better than factorized output (before) by 0.1%, which is maybe because the projection might hurt the effectiveness of LM information. And both factorized input / output perform worse than the shallow fusion method because the fusion style is fixed by η but shallow fusion tunes a better β. Finally, when it comes to the factorized attention (pos 1) in Table 2, we meet the best performance 2.0 / 4.3. And results shows that the factorized attention method (pos 2) does not perform well where the factorized attention hurts the acoustic information $\boldsymbol{H}_{\text{Enc}}$ in source-attention. Thus we choose factorized-attention (pos 1) as our proposed architecture.

Compared with the AED baseline, although the unadapted factorized AED model lags a little by 0.2 on the test-other set, the adapted *factorized AED* (Fact. Attn. (pos 1)) (ex-LM=✓) outperforms shallow fusion (+SF) or density ratio (+DR) with 0.1∼0.2 absolute (abs.) WER reduction (WERR). Shown in the last line

[1] To be notified, even if two target domains have the same name (i.e. people), the detail text conditions are different.

Dataset	Y-education	Y-news	Y-people	Y-science	P-arts	P-health	P-people	P-science
words (M)	17.33	3.74	4.24	3.41	0.69	1.44	1.38	4.08
duration (h)	1665.3	358.8	383.8	313.9	64.3	123.8	122.8	381.0

Table 1. Statistics of GigaSpeech target domains (alphabet order) train sets on their word counts (M, million words) and durations (h, hours). Y denotes youtube audios and P denotes podcast audios.

Adaptation data	Y-education	Y-news	Y-people	Y-science	P-arts	P-health	P-people	P-science
Source LM (PPL)	106.8	136.0	135.3	160.0	119.6	112.6	111.2	151.0
Target LM (PPL)	58.7	40.1	46.5	37.1	109.7	40.6	59.3	20.5
PPL reduction	45.0%	70.5%	65.6%	76.8%	8.3%	63.9%	46.7%	86.4%
AED	16.1 / 9.9	18.1 / 16.8	23.3 / 17.7	17.1 / 19.0	11.4 / 14.9	16.3 / 16.1	16.2 / 16.4	17.6 / 16.3
+Shallow Fusion	14.0 / 9.6	16.0 / 14.8	21.6 / 16.2	15.3 / 16.4	11.0 / 12.7	14.4 / 14.2	15.1 / 15.2	14.1 / 13.6
+Density Ratio	13.4 / 9.1	15.3 / 14.5	20.8 / 15.3	14.3 / 15.2	10.8 / 12.5	14.0 / 13.9	14.6 / 14.8	13.4 / 13.0
Factorized AED	16.0 / 9.8	18.3 / 17.0	23.5 / 17.8	17.5 / 19.4	11.5 / 15.1	16.4 / 16.2	16.3 / 16.5	17.9 / 16.5
+text adaptation	13.3 / 8.9	15.0 / 13.6	20.4 / 15.3	14.0 / 15.1	10.8 / 13.9	13.4 / 13.2	14.5 / 14.6	12.6 / 11.4
++Shallow Fusion	**12.5 / 8.4**	**14.3 / 13.0**	**19.2 / 14.7**	**13.3 / 14.4**	**10.6 / 10.9**	**12.8 / 12.7**	**14.1 / 14.3**	**11.8 / 10.7**

Table 3. LM description (perplexity) (PPL) and ASR Performance (WER) (%) comparison for domain adaptation with different methods. Factorized AED is referred to as the proposed factorized-attention (pos 1) model. Token-level perplexity (PPL) is averaged on dev and test sets. Source LM is the LM separated from the factorized AED model. Target LM is trained separately using text mentioned in Table 1.

Model		ex-LM	PPL	test-clean	test-other
AED		✗	-	2.5	5.4
		+SF	-	2.2	4.4
		+DR	61.7	2.1	4.3
Fact. Input		✗	62.4	2.5	5.7
		✓	62.4	2.3	4.7
Fact. Output	before	✗	64.2	2.6	6.0
		✓	64.2	2.4	5.0
	after	✗	63.8	2.6	5.8
		✓	63.8	2.3	4.9
Fact. Attn.	pos 0	✗	61.3	2.5	5.7
		✓	61.3	2.4	4.6
	pos 1	✗	59.8	2.5	5.6
		✓	59.8	2.1	4.3
	pos 2	✗	67.1	2.8	6.6
		✓	67.1	2.7	6.2
Fact. Attn.	pos 1	✓+SF	59.8	2.0	4.1

Table 2. Source (Internal) LM perplexity (PPL) and ASR Performance (WER) (%) Comparison of different factorized methods.

of Table 2, the factorized AED can be further boosted using shallow fusion (+SF) to achieve the best performance 2.0 / 4.1, which is 5∼10% relative (rel.) WERR compared to AED +SF / +DR.

4.3. Out-of-domain Adaptation

The target domain language models are trained using different scales performances are shown in the upper part of Table 3. With at least 50% rel. PPL reduction for most sets, the mismatch between source and target domain shows a huge potential to the text-only adaptation.

Shown in the 3rd block of Table 3, the factorized AED architecture after '+text adaptation' shows great improvement (i.e. 5%∼30% rel. WERR, averaged (ave.) 16% rel. WERR across all target domains) compared to the AED baseline. The proposed method demonstrates more potential compared with the shallow fusion model without fine-tuning the hyper-parameter β, and at the same time outperforms the baseline+SF method by ave. 6.1% rel. WERR and baseline+DR method by at most 10% rel. WERR. This further proves that the improvement of the language model part could be transferred to the factorized AED model. Meanwhile, the best performance achieves by 21% compared with the baseline when both shallow fusion and factorized AED architecture are preferred.

Experiments show that the factorized AED performance is sensitive to the target LM. After text-only adaptation, the improvement is limited if the target LM performs badly (P-arts, PPL reduction is 8%, rel. WERR are 5% / 7% while the average value is 16% compared to the baseline), and is boosted if target LM is advanced (P-science, PPL reduction is 86%, rel. WERR are 28% / 30%).

5. CONCLUSION

In our work, we propose a novel factorized AED architecture to integrate the external LM into the ASR model for fast text-only domain adaptation. The proposed factorized AED architecture uses the advanced conformer-based AED model and integrates the source LM into the transformer decoder using factorized attention. During text-only adaptation, the factorized AED model with plug-and-play target LMs shows its potential on various datasets. Text-only adaptation experiments show that the proposed method significantly outperforms the AED baseline by 20% / 24% relative WER reduction (rWERR) for in-domain LibriSpeech test sets and 8%∼33% for out-of-domain GigaSpeech test sets. Compared with density ratio method, we achieve ∼5% rWERR for in-domain sets and ∼8% rWERR for out-of-domain sets.

6. ACKNOWLEDGMENT

This work was supported by the National Natural Science Foundation of China (NSFC) under Grant No. 62122050, 62071288 and 62206171, in part by Shanghai Municipal Science and Technology Major Project under Grant 2021SHZDZX0102 and Alibaba Group through Alibaba Innovative Research Program.

7. REFERENCES

[1] Alex Graves, Santiago Fernández, et al., "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *Proc. ICML*, 2006, pp. 369–376.

[2] Ashish Vaswani, Noam Shazeer, et al., "Attention is all you need," *Advances in neural information processing systems*, vol. 30, Dec. 2017.

[3] Takaaki Hori, Shinji Watanabe, and John R Hershey, "Joint CTC/attention decoding for end-to-end speech recognition," in *Proc. ACL*, 2017, pp. 518–529.

[4] Anmol Gulati, James Qin, et al., "Conformer: Convolution-augmented transformer for speech recognition," in *Proc. Interspeech*, May 2020, pp. 5036–5040.

[5] Alex Graves, "Sequence transduction with recurrent neural networks," *Proc. ICML*, Nov. 2012.

[6] Jinyu Li, "Recent advances in end-to-end automatic speech recognition," *APSIPA Transactions on Signal and Information Processing*, vol. 11, no. 1, 2022.

[7] Peter Bell, Joachim Fainberg, et al., "Adaptation algorithms for neural network-based speech recognition: An overview," *IEEE Open Journal of Signal Processing*, vol. 2, pp. 33–66, 2021.

[8] Tian Tan, Yanmin Qian, et al., "Cluster adaptive training for deep neural network," in *Proc. ICASSP 2015*, Apr. 2015, pp. 4325–4329.

[9] Xun Gong, Yizhou Lu, et al., "Layer-wise fast adaptation for end-to-end multi-accent speech recognition," in *Proc. Interspeech*, 2021, pp. 1274–1278.

[10] Yizhou Lu, Mingkun Huang, et al., "Bi-encoder transformer network for mandarin-english code-switching speech recognition using mixture of experts," in *Proc. Interspeech*, Oct. 2020, pp. 4766–4770.

[11] Dong Yu, Kaisheng Yao, et al., "Kl-divergence regularized deep neural network adaptation for improved large vocabulary speech recognition," in *Proc. ICASSP*, 2013, pp. 7893–7897.

[12] Vimal Manohar, Pegah Ghahremani, et al., "A teacher-student learning approach for unsupervised domain adaptation of sequence-trained asr models," in *Proc. SLT*, 2018, pp. 250–257.

[13] Zhong Meng, Jinyu Li, et al., "Adversarial teacher-student learning for unsupervised domain adaptation," in *Proc. ICASSP*, 2018, pp. 5949–5953.

[14] Zhong Meng, Hu Hu, et al., "L-vector: Neural label embedding for domain adaptation," in *Proc. ICASSP*, 2020, pp. 7389–7393.

[15] Yan Deng, Rui Zhao, Zhong Meng, Xie Chen, Bing Liu, Jinyu Li, Yifan Gong, and Lei He, "Improving rnn-t for domain scaling using semi-supervised training with neural tts," in *Interspeech 2021*, August 2021.

[16] Andros Tjandra, Sakriani Sakti, and Satoshi Nakamura, "Listening while speaking: Speech chain by deep learning," in *2017 IEEE ASRU*, 2017, pp. 301–308.

[17] Fengpeng Yue, Yan Deng, Lei He, and Tom Ko, "Exploring machine speech chain for domain adaptation and few-shot speaker adaptation," *ArXiv*, vol. abs/2104.03815, 2021.

[18] Murali Karthick Baskar, Shinji Watanabe, et al., "Semi-supervised sequence-to-sequence asr using unpaired speech and text," *arXiv preprint arXiv:1905.01152*, 2019.

[19] Wei Wang, Zhikai Zhou, et al., "Towards data selection on tts data for children's speech recognition," in *Proc. ICASSP*, 2021, pp. 6888–6892.

[20] Anjuli Kannan, Yonghui Wu, et al., "An analysis of incorporating an external language model into a sequence-to-sequence model," in *Proc. ICASSP*, 2018, pp. 5824–5828.

[21] Erik McDermott, Hasim Sak, and Ehsan Variani, "A density ratio approach to language model fusion in end-to-end automatic speech recognition," in *Proc. ASRU*, 2019, pp. 434–441.

[22] Caglar Gulcehre, Orhan Firat, et al., "On using monolingual corpora in neural machine translation," *arXiv preprint arXiv:1503.03535*, 2015.

[23] Anuroop Sriram, Heewoo Jun, et al., "Cold fusion: Training seq2seq models together with language models," *arXiv preprint arXiv:1708.06426*, 2017.

[24] Changhao Shan, Chao Weng, et al., "Component fusion: Learning replaceable language model component for end-to-end speech recognition system," in *Proc. ICASSP*, Brighton, United Kingdom, May 2019, pp. 5361–5635.

[25] Zhong Meng, Sarangarajan Parthasarathy, et al., "Internal language model estimation for domain-adaptive end-to-end speech recognition," in *Proc. SLT*, 2021, pp. 243–250.

[26] Ehsan Variani, David Rybach, et al., "Hybrid autoregressive transducer (hat)," in *Proc. ICASSP*, 2020, pp. 6139–6143.

[27] Mohammadreza Ghodsi, Xiaofeng Liu, et al., "Rnn-transducer with stateless prediction network," in *Proc. ICASSP*, May 2020, pp. 7049–7053.

[28] Xie Chen, Zhong Meng, et al., "Factorized neural transducer for efficient language model adaptation," in *Proc. ICASSP*, 2022, pp. 8132–8136.

[29] Vassil Panayotov, Guoguo Chen, et al., "Librispeech: An ASR corpus based on public domain audio books," in *Proc. ICASSP*, 2015, pp. 5206–5210.

[30] Guoguo Chen, Shuzhou Chai, et al., "GigaSpeech: An Evolving, Multi-Domain ASR Corpus with 10,000 Hours of Transcribed Audio," in *Proc. Interspeech*, 2021, pp. 3670–3674.

[31] Daniel S. Park, William Chan, et al., "Specaugment: A simple data augmentation method for automatic speech recognition," *Proc. Interspeech*, pp. 2613–2617, Sept. 2019.

[32] Taku Kudo and John Richardson, "Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 2018, pp. 66–71.

[33] Shinji Watanabe, Takaaki Hori, et al., "Espnet: End-to-end speech processing toolkit," *arXiv preprint arXiv:1804.00015*, Mar. 2018.