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/ICASSP49357.2023.10095910

JOINT DISCRIMINATOR AND TRANSFER BASED FAST DOMAIN ADAPTATION FOR END-TO-END SPEECH RECOGNITION

Hang Shao[1], *Tian Tan*[2], *Wei Wang*[1], *Xun Gong*[1], *Yanmin Qian*[1†]

[1] MoE Key Lab of Artificial Intelligence, AI Institute
X-LANCE Lab, Department of Computer Science and Engineering,
Shanghai Jiao Tong University,Shanghai,China
[2]AISpeech Ltd, Suzhou, China

ABSTRACT

Adapting End-to-End (E2E) models to unseen domains is still a big challenge since training E2E models requires lots of paired audio and text training data. We propose a novel domain adaptation framework for the E2E model, which only uses the text of the target domain. Moreover, the proposed methods can keep the performance on the source domain intact while greatly improving the performance on the target domain. The proposed framework consists of two parts: the discriminator and the transfer which were optimized separately. Finally, optimized discriminator and transfer were combined and evaluated on two domain adaption tasks. In the experiments of adapting the English LIBRISPEECH to GIGASPEECH, we obtained an average relative 11.6% and 11.8% on word error rate (WER) reduction for the target domain dev and test sets, respectively, while almost without WER degradation on the source domain. For the in-house Chinese corpus aviation and TV, the character error rate (CER) of the source domain increased within 5%, while the CER on the target domain achieved around relative 85% and 42% improvement, respectively. In addition, our approach is also more effective in the mixed domain scenarios in the evaluation.

Index Terms— end-to-end speech recognition, domain adaptation, discriminator and transfer, log-likelihood ratio

1. INTRODUCTION

The E2E models including connectionist temporal classification (CTC) [1, 2], attention-based encoder-decoder (AED) models [3, 4] and recurrent neural network transducer (RNN-T) [5, 6] have gone mainstream and achieved state of the art performance for automatic speech recognition (ASR) [7, 8]. Traditional hybrid ASR models consist of separate acoustic model (AM), pronunciation models and language model (LM), which can be individually optimized using a variety of data sources, especially large amounts of textual data. However, E2E model training requires paired data, which is more difficult to obtain and expensive. Therefore, text-only domain adaptation for the E2E models is attracting more and more researchers. Some common methods include synthesizing the target-domain text into paired data using text-to-speech (TTS) [9] and fine-tuning the E2E model with the target-domain paired data. However, TTS model training suffers from high computational overhead.

LM fusion has been proven to be another effective way of domain adaptation with text-only data. Shallow Fusion (SF) [10, 11, 12, 13] is a simple yet effective method by doing a log-linear interpolation between the scores of the E2E model and a separately-trained LM during decoding. Such LM integration is further improved with the Density Ratio (DR) approach [14] that subtracts the score of the source domain LM from the interpolated score of shallow fusion. Following this idea, the hybrid autoregressive transducer (HAT) [15], internal LM estimation (ILME) [16] and internal LM training (ILMT) [17] are proposed to estimate the score of the internal LM component in a source domain E2E model. The internal LM score is used to replace the source domain LM score in the DR approach and shows promising improvement on the target domain.

Nevertheless, the improvement on the target domain brought by LM fusion methods often comes at the cost of severe deterioration on the source domain. Especially when a large difference exists between the source and target domain, the model may perform well on the target domain, but is completely unusable on the source domain. In addition, LM fusion based methods are also not suitable for the mixed domain scenarios where a speech contains multiple domain switches of both the source domain and target domain. Therefore, unlike previous work which looked at complete domain transfer at test time, [18] proposed a likelihood ratio (LLR) based domain adaptation method without causing degradation on general domains.

However, in LLR, an ngram was used to do domain adaption which is not good enough when used as LM. This is particularly evident in our experiments when transferred from LIBRISPEECH to GIGASPEECH. In addition, it is difficult for LLR to distinguish whether a word is a rare word or not when the perplexity (PPL) differences between the source domain and target domain are small.

Based on the above considerations, we proposed a new domain adaptation framework for E2E models, which consists of two parts: discriminator and transfer. The discriminator is used to identify the domain to which it belongs in real-time, and the transfer is used to better transfer to the corresponding domain. The discriminator and transfer are optimized separately. The final optimized discriminator and transfer are combined to do the domain adaption task. Our framework was evaluated on two test sets. On the English corpus, we obtained 11.6% and 11.8% relative WER reduction respectively in the target domain, while almost without sacrificing the source domain performance. On in-house Chinese corpus, a relative 85% and 42% improvement were obtained on the target domain while within 5% performance drop on the source domain. In the mixed domain scenario, our method obtained 10.0% and 28.0% relative CER reduction. Our contributions can be summarized as follow: (1) We propose a joint discriminator and transfer domain adaptive framework. (2) We optimize the discriminator and transfer separately, and the optimal joint system achieves better results. (3) Our approach also works well in a domain switching scenario where a speech contains a mixed situation of both source and target domains.

† corresponding author

2. METHODOLOGY

In this work a novel domain adaptation decoding framework is proposed to adapt a pre-trained AED model to a target domain with only Out-of-Domain (OOD) text data, meanwhile minimizing the degradation of the performance on domains already supported.

2.1. A Domain Adaptation Decoding Framework

The proposed framework consists of two parts: discriminator and transfer. The discriminator is used to identify the domain to which each token belongs in real-time, and the transfer is used to better transfer to the corresponding domain.

We take the speech X and the decoded prefix sequence Y_{u-1} (contains beamsize hypotheses) as input. For each hypothesis $\mathbf{y}_{u-1}$ in beam, the score of candidate tokens will be calculated. For each token in the token list (6979 for Chinese and 5000 for English), the discriminator is first used to discriminate whether it belongs to the source domain or the target domain. If this token belongs to the target domain, the transfer score is computed for it, otherwise the score of the source domain E2E model is computed for it. Finally, the $(beam, tokensize)$-dimensional Score and the already decoded prefix sequence Y_{u-1} are used in the BeamSearch algorithm to get a new decoded sequence Y_u in $(beam, u)$ dimensions. The following algorithm 1 flow describes the proposed decoding framework.

Algorithm 1 A Domain Adaptation Decoding Framework

Input: X, Audio input.
Input: Y_{u-1}, Prefix decoded sequence.
Output: Y_u, The new decoded sequence to be decoding
Require: TransferScore, Score after domain transfer, as mentioned in Eq. (6)
Require: SourceScore, Source model score, as mentioned in Eq. (7)

function JOINT-DISCRIMINANT-TRANSFER(X,Y_{u-1})
  for $b \leftarrow 1$ to beamsize **do**
    $\mathbf{y}_{u-1} \leftarrow Y_{b,u-1}$
    for $i \leftarrow 1$ to tokensize **do**
      Domain $\leftarrow$ Discriminator($\mathbf{y}_{u-1}, i$)
      if Domain = d_{tgt} **then**
        $Score_{b,i} \leftarrow$ TransferScore($\mathbf{y}_{u-1}, X, i$)
      else
        $Score_{b,i} \leftarrow$ SourceScore($\mathbf{y}_{u-1}, X, i$)
  $Y_u \leftarrow$ BeamSearch($Y_{u-1}, Score$)
  return Y_u

2.2. Discriminators on Domain Adaptation

In our framework, the main role of the discriminator is to discern whether the current token needs to be transferred or not. When a token that is the source domain is discriminated as the target domain, an error of offset decoding occurs. Since the discriminator affects the overall recognition performance, the performance of the discriminator must be as accurate as possible to better discriminate the domain switching dynamically and in real time. The following three discriminators have been proposed.

2.2.1. Discriminator based on Domain LM Scores

It was observed that the same utterances have obvious contrast of log likelihood calculated by different LMs from different domains. So a straightforward idea for the discriminator is using the LM score from the source domain and target domain. As shown in Eq. (1), when the source domain LM log-likelihood score $S_{src} < T_1$ and the target domain LM score $S_{tgt} > T_2$, this token is classified to the target domain and needs domain transfer.

$$Domain := \begin{cases} d_{tgt} & \text{if } S_{src} < T_1 \text{ and } S_{tgt} > T_2 \\ d_{src} & \text{otherwise,} \end{cases} \tag{1}$$

where S_{src} and S_{tgt} are obtained by computing the LM scores of token i on the source and target domains from the two inputs Y_{u-1} and token i, respectively, passed in by the discriminator in Alg. 1.

2.2.2. Discriminator based on Log-likelihood Ratio

Because the direct comparison of the score through the LM with the threshold is too broad and may not work well enough. Therefore, as shown in Eq. (4), the discriminator in the form of a relative LM scores difference is proposed. If the score difference exceeds a certain threshold, then it belongs to the target domain, otherwise it belongs to the source domain. In this work, three LMs were investigated including N-gram, neural network language model (NNLM) and ILME. Since N-gram can only see local information, we believe that using a sliding window to provide longer history would be helpful. A soft sliding window was applied, which is shown in Eq. (2). In order to make an accurate comparison with the threshold value, normalization was applied to the scores in Eq. (3). The NNLM can also see long historical information, so we compare the N-gram with the sliding window with the NNLM.

$$S_u = \beta S_{u-1} + \overline{S}_i \tag{2}$$

$$S'_u = \text{Norm}(S_u) = S_u / \frac{1-\beta^u}{1-\beta} \tag{3}$$

where $\overline{S}_i$ indicates the log-likelihood score of the i-th token in the token list and S_u denotes the momentum LM score of the decoded sequence Y_u.

Since ILME in [17] is able to evaluate the internal LM of the source domain more precisely, it may be a better choice for discriminator. Since the ILM in [17] is only a weak LM, the score of ELM may not be of the same magnitude in numerical scale. In order to make a more accurate comparison, a hyperparameter λ in Eq. 4 is used.

$$Domain := \begin{cases} d_{tgt} & \text{if } S_{tgt} - \lambda S_{src} > T \\ d_{src} & \text{otherwise,} \end{cases} \tag{4}$$

where S_{src} and S_{tgt} are S'_u on the source domain and target domain when the LM is N-gram with sliding window, respectively, otherwise S_{src} are S_{tgt} are LM scores on the source and target domain, respectively. T is a threshold value.

2.2.3. Discriminator based on Neural Domain Classifier

The other choice for domain discriminator is model based classifier. A neural networks based model was investigated in this work, detailed configuration is introduced in Sec 3.1. The training is similar to the LM training, for each token in the target domain text, we consider that its label belongs to the target domain, on the source domain as well. Since the amount of text data in the source domain is generally much larger than that in the target domain, we introduce the focal loss in [19] as the training loss function for our classification in Eq. (5).

$$L_{fl} = \begin{cases} -(1-\hat{p})^{\gamma} \log(\hat{p}) & \text{if } y = 1, \\ -\hat{p}^{\gamma} \log(1-\hat{p}) & \text{if } y = 0, \end{cases} \tag{5}$$

where γ is an adjustable factor to control the rate at which easy examples are down-weighted and $\hat{p}$ is the predicted probability of label $y = 1$.

In addition, we can also repeat the target domain text several times so that it is of the same order of magnitude as the source domain. For the inference of the neural domain classifier, The output of logits is passed through log_softmax to discriminate which domain the classifier predicts by argmax.

2.3. Transfer on Domain Adaptation

For the transfer, we currently use the LM fusion based approach. More precisely, as shown in Eq. (6), a linear interpolation is applied between the score of the E2E model and the score of the LM.

$$\begin{aligned}\text{TransferScore}(\mathbf{y}_{u-1}, X, i) &= \log P(y = i|X, \mathbf{y}_{u-1}; \theta^S_{AED}) \\ -\lambda_1 \log P(y = i|\mathbf{y}_{u-1}; \theta^S_{LM}) &+ \lambda_2 \log P(y = i|\mathbf{y}_{u-1}; \theta^T_{LM})\end{aligned} \tag{6}$$

where $P(y|X, \mathbf{y}_{u-1}; \theta^S_{AED})$ is the posterior score of the source domain E2E model, which is referred to as SourceScore Eq. (7), and $P(y|\mathbf{y}_{u-1}; \theta^S_{LM})$ and $P(y|\mathbf{y}_{u-1}; \theta^T_{LM})$ are the posteriors for the source LM and target LM, respectively.

$$\text{SourceScore}(\mathbf{y}_{u-1}, X, i) = \log P(y = i|X, \mathbf{y}_{u-1}; \theta^S_{AED}) \tag{7}$$

When $\lambda_1 = 0$, the transfer method corresponding to Eq. (6) is SF, which is a more moderate transfer approach that does not subtract the score of the source domain language model when doing the transfer, and thus the source domain performance drop is moderate. When $\lambda_1 \neq 0$ and $\lambda_2 \neq 0$, the transfer method corresponding to Eq. (6) is the DR-based transfer method, which is a complete transfer, i.e., the score of the source domain LM is subtracted, and the performance on the source domain degrades dramatically.

2.4. Comparison with LLR

In this section, a theoretical comparison is made between our proposed framework and the previous LLR method [18]. We illustrate that LLR is a special case of our framework. The LLR-based domain adaptation approach can be expressed by the following equation [18].

$$S(G) := \begin{cases} LLR(G) & \text{if } LLR(G) > T \\ 0 & \text{otherwise,} \end{cases} \tag{8}$$

where $S(G)$ is the boost score that does linear interpolation with the E2E model and $LLR(G) = \log P_T(G) - \log P_S(G)$, which $P_T(G)$ and $P_S(G)$ are the posterior of n-gram G on the target and source domain LM N-grams, respectively.

When the discriminator is $\lambda = 1$ in Eq. (4) and the transfer method is $\lambda_1 = \lambda_2$ in Eq. (6), under this condition it is equivalent to LLR. Thus LLR is a special case of our framework, and it is used as a baseline in our experiments.

3. EXPERIMENTS

3.1. Experiment Setup

3.1.1. Detailed Model Structure

Our baseline Conformer [20] model consists of 12 encoder layers and 6 decoder layers with 2048 hidden units. Each encoder layer is a Conformer block with 8 heads of 512 dimension self-attention [21] and each layer of decoder layer is a Transformer block with 8 heads of 512 dimension self-attention. When the transfer method is ILME, the weight λ_1 and λ_2 in Eq. (6) set to 0.3 and 0.6 respectively. The weight for CTC and attention is set to 0.3 and 0.7.We use an 80 dimensions log Mel-filterbank with 25ms window length computed every 10ms as inputs of audio encoder. SpecAugment [22] as a data augmentation policy is also used during model training. The Adam [23] optimizer is adopted with 0.0025 initial learning rate and 40,000 warmup steps. The structure of the neural network language model is the same as the decoder layer of the baseline conformer model. The neural domain classifier consists of 4 layers of lstm with 2,048 hidden units. The number of modeling units (BPE) [24] for the English corpus in the decoder is 5,000 and the number of modeling units (Char) for the Chinese corpus is 6,979. All models are trained with ESPnet [25] toolkit until convergence.

3.1.2. Dataset

We have experimented on both English and Chinese corpus. English corpus is taken on 960-hour LIBRISPEECH [26] as intra-domain and Youtube partition of GIGASPEECH [27] XL subset, which has five different domains such as science (sci), news, people (peo), entertainment (ent) and education (edu) were selected as target domains. In Table 1, there are the linguistic gap between the five domains of GIGASPEECH and LIBRISPEECH, but they are not significant.

Table 1: PPL of LIBRISPEECH LM on various domain test sets

Dataset	test-other	sci	news	peo	ent	edu
PPL	61	201	213	286	316	115

Chinese Corpus including in-house 3K-hours (aitrans) as general domain, aviation dialogue domain (aviage) has 103,307 transcripts and synthetic TV has 271,390 transcripts whose test set contains 13.8 hours of 5,000 utterances synthesized by TTS. The transcripts of TV are a variety of movies names populated into the generic domain to form a target domain containing rare words. The PPL of the source domain aitrans LM is 26 on its own domain and in the target domain TV and aviage are 796 and 894, respectively. there are the significant linguistic gap between aitrans and TV/aviage.

3.2. Experiment Results and Analysis

3.2.1. Optimizing Domain Transfer Methods

In Table. 3, the first row shows the decoding results of the source domain model without any discriminator and transfer under each domain. Since TV is a dedicated rare word corpus synthesized by TTS, nothing can go wrong acoustically, and its 7.89% CER is basically all errors caused by language mismatch. The second row is the case of the source domain model without a discriminator, we can see that without a discriminator, the target domain has better performance, however, the performance of the source domain drops dramatically. From row 3 to row 5, we optimize the transfer method when the relative CER of the source domain increases by less then 10%. The results showed that ILME has approximately 72% and 36% WER reduction over the two target domain test sets compared with the baseline N-gram. However, when the discriminator discriminates the domain incorrectly, the source domain CER increases will occur. Therefore, the discriminator also needs to be optimized separately in order to more accurately identify the domain type.

3.2.2. Optimizing Domain Discriminator

We selected the ILME method as our transfer method and optimized the discriminator in this section. The domain classification accuracy (ACC) of each token was used to rank each method. As in Table. 4, longer history does increase the classification accuracy. We

Table 2: Performance WER(%) comparison of different setups on English corpus. TF and DC stand for transfer and discriminator respectively.The left results in each column show WERs on LIBRISPEECH test clean / test other sets. The right results in each column show WERs on GIGASPEECH dev / test sets for the corresponding domain.

TF	DC	libri → science		libri → news		libri → people		libri → entertainment		libri → education	
		test c/o	dev/test	test c/o	dev/test	test c/o	dev/test	test c/o	dev/test	test c/o	dev/test
N/A	N/A	2.5/5.4	17.1/19.0	2.5/5.4	18.1/16.8	2.5/5.4	23.3/17.7	2.5/5.4	24.1/24.1	2.5/5.4	16.1/9.9
SF	N/A	2.9/6.5	15.3/16.4	2.9/6.4	16.0/14.8	3.0/6.3	21.6/16.2	3.0/6.2	23.3/22.9	2.9/6.5	14.0/9.6
N-gram	N-gram	2.5/5.5	17.1/19.1	2.5/5.4	18.1/16.7	2.5/5.5	23.4/17.8	2.5/5.4	24.1/24.2	2.5/5.5	16.0/9.8
ILME	N/A	3.1/6.7	13.8/14.8	3.3/6.8	14.9/13.2	3.1/6.3	19.9/15.1	3.1/6.3	22.0/21.7	2.9/6.5	12.9/8.5
ILME	ILME	2.5/5.5	14.2/15.1	2.6/5.6	15.8/14.3	2.5/5.4	20.9/16.2	2.5/5.5	22.8/22.7	2.5/5.4	13.6/8.9

Table 3: Performance CER(%) comparison of different configurations on Chinese corpus. The two numbers in the source domain aitrans column are the CER(%) of the source domain with the TV or aviage domain adaptation respectively.

Transfer	Discriminator	source	target	
		aitrans	TV	aviage
N/A	N/A	8.92/8.92	7.89	57.58
ILME	N/A	101.94/20.61	0.30	26.07
N-gram	N-gram	9.02/9.01	2.74	46.11
DR	N-gram	9.48/9.49	1.15	30.71
ILME	N-gram	9.72/9.67	0.75	29.44
ILME	NNLM	9.33/9.42	0.52	28.57
ILME	ILME	9.29/9.33	0.41	26.71

found that when β was adjusted to 0.9 in Eq. (2), the first 10 tokens could be seen, its accuracy is very close to the accuracy of NNLM. The accuracy of ILME is much higher than both N-gram and NNLM, even more it obtained a competitive performance as domain classifier. Since domain classifier needs extra training steps, ILME was chosen as our final discriminator. As shown in the last three rows of Table. 3, both the source and target domains could be further improved as the discriminator is optimized. Finally, our best system, i.e. the last row of Table. 3, gained 85% and 42% relative CER reduction on the target domain TV and aviage, respectively, compared to LLR, i.e. the third row of Table. 3.

Table 4: Domain classification accuracy ACC(%) of different Discriminators. The two numbers in the source domain aitrans column are the classification ACC(%) between the source domain and the TV/aviage target domain respectively.

Discriminator	source	target	
	aitrans	TV	aviage
Domain LM Scores	80.3/82.5	82.6	81.3
N-gram	82.0/85.3	85.4	84.5
+ sliding window	85.2/87.0	88.9	88.3
NNLM	85.6/87.6	89.1	88.7
ILME	89.1/89.2	93.1	91.3
Domain Classifier	89.5/89.0	93.5	91.1

3.2.3. Evaluation on the English corpus

The optimal transfer and discriminator were further evaluated on open-sourced English corpus. As shown in the Table. 2. The first row is the decoding result of the source domain model in each domain. The second row is shallow fusion, as a soft transfer approach, the improvement of the target domain is not as obvious as the fourth row of ILME, but the performance degradation of the source domain is smaller. The ILME without discriminator in the fourth row leads to a complete transfer to the target domain at the cost of performance degradation on the source domain. Our proposed approach, i.e. the last row, obtained significant improvement on the target domain without performance degradation on the source domain. Meanwhile, comparing it with the LLR method based on boost score, i.e. the third row, the newly proposed approach has the same performance in the source domain, while obtains average 11.6% and 11.8% relative CER reduction on dev and test, respectively, for all the five target domain in GIGASPEECH.

3.2.4. Evaluation on the Mixed Domain Scenarios

The test set of the mixed domain was generated by TTS, and the text was generated by random concatenating utterances in the source and target domain. The mixed domain is a situation where one speech has multiple domain switching. Table. 5 shows that comparing our approach (the last row) to the source domain unadapted model (the first row), we obtain relative 15.6% and 32.7% performance improvement on the TV and aviage domains, respectively. Moreover we still obtain relative 10.0% and 28.0% CER reduction over the basic LLR method (the second row). It shows that our proposed framework is also more effective in the mixed domain scenarios.

Table 5: Performance CER(%) comparison of different setups in the mixed domain scenarios.

Transfer	Discriminator	aitrans & TV mixed	aitrans & aviage mixed
N/A	N/A	5.76	12.58
N-gram	N-gram	5.38	11.74
ILME	N/A	7.19	10.01
ILME	ILME	**4.86**	**8.46**

4. CONCLUSIONS

In this paper, we proposed a new framework for text-only fast domain adaptation. The proposed framework consists of two parts: the discriminator and the transfer which were optimized separately. Finally, optimized discriminator and transfer were combined and evaluated on two domain adaption tasks. In the experiments of adapting the English LIBRISPEECH to GIGASPEECH, we obtained an average relative 11.6% and 11.8% WER reduction on the target-domain dev and test, respectively, while almost without any WER degradation on the source domain. For the in-house Chinese corpus aviation and TV, the CER on the source domain increased within 5%, while the CER on the target domain achieved around relative 85% and 42% improvement, respectively. In addition, our approach is also more effective in the mixed domain scenarios.

5. ACKNOWLEDGEMENTS

This work was supported in part by China NSFC projects under Grants 62122050 and 62071288, and in part by Shanghai Municipal Science and Technology Major Project under Grant 2021SHZDZX0102.

6. REFERENCES

[1] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *Proceedings of the 23rd international conference on Machine learning*, 2006, pp. 369–376.

[2] J. Li, G. Ye, A. Das, R. Zhao, and Y. Gong, "Advancing acoustic-to-word ctc model," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5794–5798.

[3] 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.

[4] W. Chan, N. Jaitly, Q. Le, and O. Vinyals, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," in *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2016, pp. 4960–4964.

[5] A. Graves and N. Jaitly, "Towards end-to-end speech recognition with recurrent neural networks," in *International conference on machine learning*. PMLR, 2014, pp. 1764–1772.

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

[7] C.-C. Chiu, T. N. Sainath, Y. Wu, R. Prabhavalkar, P. Nguyen, Z. Chen, A. Kannan, R. J. Weiss, K. Rao, E. Gonina *et al.*, "State-of-the-art speech recognition with sequence-to-sequence models," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 4774–4778.

[8] J. Li, R. Zhao, Z. Meng, Y. Liu, W. Wei, S. Parthasarathy, V. Mazalov, Z. Wang, L. He, S. Zhao *et al.*, "Developing rnn-t models surpassing high-performance hybrid models with customization capability," *arXiv preprint arXiv:2007.15188*, 2020.

[9] W. Wang, Z. Zhou, Y. Lu, H. Wang, C. Du, and Y. Qian, "Towards data selection on tts data for children's speech recognition," in *ICASSP 2021 IEEE*, 2021, pp. 6888–6892.

[10] 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 *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 1–5828.

[11] T. Mikolov, M. Karafiát, L. Burget, J. Cernockỳ, and S. Khudanpur, "Recurrent neural network based language model." in *Interspeech*, vol. 2, no. 3. Makuhari, 2010, pp. 1045–1048.

[12] A. Hannun, C. Case, J. Casper, B. Catanzaro, G. Diamos, E. Elsen, R. Prenger, S. Satheesh, S. Sengupta, A. Coates *et al.*, "Deep speech: Scaling up end-to-end speech recognition," *arXiv preprint arXiv:1412.5567*, 2014.

[13] 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 *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 1–5828.

[14] E. McDermott, H. Sak, and E. Variani, "A density ratio approach to language model fusion in end-to-end automatic speech recognition," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2019, pp. 434–441.

[15] E. Variani, D. Rybach, C. Allauzen, and M. Riley, "Hybrid autoregressive transducer (hat)," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6139–6143.

[16] 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," *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7338–7342, 2021.

[17] 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 *2021 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2021, pp. 243–250.

[18] C. Choudhury, A. Gandhe, X. Ding, and I. Bulyko, "A likelihood ratio based domain adaptation method for e2e models," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 6762–6766.

[19] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2980–2988.

[20] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu *et al.*, "Conformer: Convolution-augmented transformer for speech recognition," *arXiv preprint arXiv:2005.08100*, 2020.

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

[22] 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.

[23] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[24] T. Kudo, "Subword regularization: Improving neural network translation models with multiple subword candidates," *arXiv preprint arXiv:1804.10959*, 2018.

[25] 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.

[26] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.

[27] G. Chen, S. Chai, G. Wang, J. Du, W.-Q. Zhang, C. Weng, D. Su, D. Povey, J. Trmal, J. Zhang *et al.*, "Gigaspeech: An evolving, multi-domain asr corpus with 10,000 hours of transcribed audio," *arXiv preprint arXiv:2106.06909*, 2021.