INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

A More Accurate Internal Language Model Score Estimation for the Hybrid Autoregressive Transducer

Kyungmin Lee, Haeri Kim*, Sichen Jin, Jinhwan Park, Youngho Han*

Samsung Research, Seoul, South Korea

{k.m.lee, haeri.kim, sc.ehkim.jin, jh0354.park, yho.han}@samsung.com

Abstract

We present a novel constrained learning method for hybrid autoregressive transducer (HAT) models that results in more validated language model (LM) adaptation. LM adaptation in HAT is justified only when the transducer logits and the sum of speech and text logits in the label estimation sub-networks are approximately the same. The mean squared error (MSE) between the two logits was added to the HAT loss to encourage the HAT models to satisfy the required condition. The proposed method exhibited significantly lower and more stable internal language model perplexities than those of HAT. Consequently, it attained lower word error rates (WERs) compared to HAT in various model architecture settings and in both cases with and without LM adaptation. In the television content task, the proposed method achieved a relative reduction in WERs of up to 28.60% compared to HAT. In most cases, the accuracy of pre-trained HAT models also improved upon training with the additional MSE loss.

Index Terms: contextual speech recognition, language model adaptation

1. Introduction

Neural speech recognition (NSR) systems possess an all-neural architecture in which the mapping between speech signals and transcriptions is directly learned [1, 2, 3]. NSR systems have attracted considerable research attention because they can learn alignments between variable input and output sequences without hand-crafted data such as pronunciation dictionaries, whereas conventional automatic speech recognition (ASR) systems based on hidden Markov models require this mechanism [4, 5]. Moreover, the model of NSR systems can be lightweight, making them suitable for use on mobile devices while still maintaining the accuracy of conventional server-based ASR systems [1, 2]. In addition to these structural advantages, NSR systems exhibit high accurate recognition rates because of their excellent sequence representation capability, which exceeds that of conventional ASR systems [3].

In ASR systems, contextual biasing is required to accurately recognize unseen domain speech inputs, such as content titles or voice-command-related utterances. On-demand language model (LM) adaptation is the one of the most prevalently employed mechanisms. According to Bayes' theorem, accurately computing linguistic prior probabilities is critical. [6] introduced the density ratio method to estimate prior probability based on an assumption, which has not been validated, that output probabilities of NSR systems can be factorized, similar to conventional ASR systems. As factorization-suitable structures, transducer-based NSR systems, which can compute prior probability with a separate part of the network, have been developed [7, 8]. The LM factorization models [7] consist of two kinds of prediction networks to estimate alignment information and labels. The label prediction networks have the same architecture as the neural LMs. This structural feature allows NSR model to be learned using text-only data. Hybrid autoregressive transducer (HAT) models [8] consist of two separate sub-networks for blank and label predictions, respectively. The purpose of HAT is to estimate internal LM scores corresponding to the prior probability of the transducer-based NSR models and replace it with external LM scores. HAT has attracted considerable research attention because it serves as the baseline system not only for contextual speech recognition [9, 10, 11, 12] but also for general speech recognition [13, 14]. HAT algorithms can be justified only under a special condition that their output scores are decomposed to acoustic and linguistic scores. However, HAT cannot encourage the models to satisfy the condition.

To tackle the limitation above, we introduce a novel constrained learning method for HAT models. Specifically, the mean squared error (MSE) between HAT logits and the sum of acoustic and linguistic logits in the label prediction networks is used as an additional loss. This training mechanism is called HAT+MSE, which overcomes the limitations of selecting various network structures for estimating labels that were imposed by the existing HAT. The proposed methods can be easily applied to existing HAT-based NSR systems without requiring new hyperparameters. HAT+MSE significantly enhanced prior estimations compared to the original HAT methods. It also improved recognition accuracy across various joint network setups for label prediction. Pre-trained HAT models can be improved by post training with HAT+MSE.

We review the previous studies on LM adaptation and HAT in the next section. The proposed model architecture and HAT+MSE are explained in Section 3. The proposed method is evaluated in Section 4. We conclude with a summary of this work and future work in Section 5.

2. Preliminaries

2.1. Language Model Adaptation

ASR decoding problems can be formulated according to the Bayes' Theorem as follows:

$$\begin{aligned} Y^* &= \underset{\hat{Y}}{\operatorname{argmax}}\, P(\hat{Y}|X) \\ &= \underset{\hat{Y}}{\operatorname{argmax}}\, P(X|\hat{Y}) \cdot P(\hat{Y}), \end{aligned} \tag{1}$$

where a posterior $P(\hat{Y}|X)$ of a hypothesis $\hat{Y}$ for a given speech signal X is factorized into an acoustic likelihood $P(X|\hat{Y})$ and

*Equal contribution.

10.21437/Interspeech.2023-213

a prior probability $P(\hat{Y})$. To boost the probability of certain output label sequences with external LMs, $P(X|\hat{Y})$ is computed and multiplied with $P_{LM}(\hat{Y})$ for validated LM adaptation. Weighted-finite state transducer (WFST)-based ASR systems [15], a kind of conventional ASR systems, are trained separately to estimate $P(\hat{Y})$, $P(X|\hat{Y})$, and alignment information is then composed into one graph. Therefore, WFST-based ASR systems [16, 17] are suitable for computing $P(X|\hat{Y})$ in an on-the-fly manner. LMs could also be easily applied to neural acoustic encoder-only ASR systems [18, 19].

However, the same mechanism is inapplicable for NSR systems because they directly learn how to maximize the probability of a label sequence Y for given X.Among various LM adaptation methods [20, 21, 22, 23], one of the popular methods for streaming NSR systems is simply to add the log probability of LMs for the predicted text, $\log P_{LM}(\hat{Y})$, to $\log P(\hat{Y}|X)$ with a scaling factor λ during decoding [2, 20, 24] such that

$$Y' = \underset{\hat{Y}}{\operatorname{argmax}} \log P(\hat{Y}|X) + \lambda \log P_{LM}(\hat{Y}) + \gamma \mathcal{R}(X, Y), \tag{2}$$

where $\mathcal{R}$ is an optional term scaled by γ such as a penalization term for incomplete transcripts [21].

2.2. Hybrid Autoregressive Transducer (HAT)

HAT [8] is a variant of the recurrent neural network transducer (RNNT) [25]. HAT models are developed with a pair of sub-networks consisting of transcription, prediction, and joint networks to separately compute posteriors of blanks $< b >$ and labels $\mathbf{Y} = \{\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_3, ..., \mathbf{y}_k, ..., \mathbf{y}_{K-1}\}$, whereas RNNT models calculate the posteriors of $\tilde{\mathbf{Y}} = \mathbf{Y} \cup \{< b >\}$ through a single sub-network. The posterior at each node of a lattice $P(\hat{Y}_u = \tilde{\mathbf{y}}_k | X, \hat{Y}_{1:u-1})$ is computed as follows:

$$\begin{cases} P_b^{t,u} = \sigma(J_b(\mathbf{f}_b^t + \mathbf{g}_b^u)), & \tilde{\mathbf{y}}_k =< b > \\ P_l^{t,u} = (1 - P_b^{t,u})\text{Softmax}(J_l(\mathbf{f}_l^t + \mathbf{g}_l^u))_k, & \tilde{\mathbf{y}}_k \neq< b >, \end{cases} \tag{3}$$

where subscripts b and l indicate blank and label networks, respectively. $\mathbf{f}$ and $\mathbf{g}$ depict a transcription and prediction network output vector, respectively, for the tth input speech frame and uth label. Here, J represents a joint network and σ indicates a Sigmoid activation function. The uth local ILM score is defined as $J_l(\mathbf{g}_l^u)$ and can be justified under special conditions when $J_l(\mathbf{f}_l^t + \mathbf{g}_l^u) \approx J_l(\mathbf{f}_l^t) + J_l(\mathbf{g}_l^u)$. The sequence-level log probability of ILMs, $\log P_{ILM}(Y)$, is computed by normalizing each local ILM score with a log_softmax function and summing them. The on-the-fly LM adaptation of HAT during decoding is formulated as follows:

$$\begin{aligned} \tilde{Y}^* = \underset{\tilde{Y}}{\operatorname{argmax}} & \lambda_1 \log P(\tilde{Y}|X) - \lambda_2 \log P_{ILM}(\mathcal{B}(\tilde{Y})) \\ & + \lambda_3 \log P_{LM}(\mathcal{B}(\tilde{Y})), \end{aligned} \tag{4}$$

where $\mathcal{B}$ is a function to convert alignment paths to label sequences [25]. λ_1 is set to 1 in this study. This inference algorithm is mathematically justified as long as the aforementioned conditions are satisfied.

3. More accurate ILM score estimation

3.1. Model Architecture

The HAT model architecture depicted in Figure 1 is described. A pair of speech and label sequences, X and Y, are input to both sub-networks, that is, blank and label networks. In this study, J_b is developed with two linear layers and a rectified linear unit (ReLU) activation function in-between them. The configuration of J_l can be varied by modifying the kinds of activation functions and the number of pairs of linear layers and activation functions N_{J_l}. When $N_{J_l} = 0$, J_l consists of an activation function and a linear layer without the first linear layer marked with the dashed line.

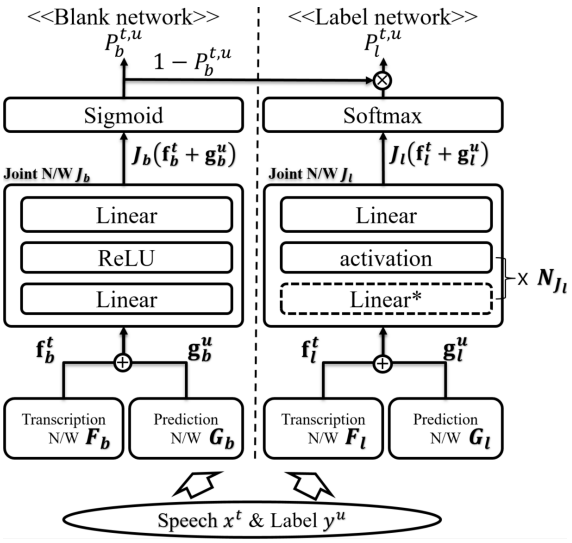


Figure 1: *Schematic of hybrid autoregressive transducer (HAT) model architecture. The layer with * is not used when $N_{J_l} = 0$. Transcription network (N/W) and prediction N/W can be shared for the blank and label networks.*

3.2. Constrained Learning with MSE loss

As explained in Section 2.2, $J_l(\mathbf{f}_l^t + \mathbf{g}_l^u)$ should be approximately equal to $J_l(\mathbf{f}_l^t) + J_l(\mathbf{g}_l^u)$ to estimate ILM scores accurately. Therefore, we devised the novel training method to encourage the output vectors of J_l satisfy the condition. MSE is used as an additional loss to minimize the difference between $J_l(\mathbf{f}_l^t + \mathbf{g}_l^u)$ and $J_l(\mathbf{f}_l^t) + J_l(\mathbf{g}_l^u)$ and is computed as follows:

$$\mathcal{L}_{\text{MSE}}^{t,u} = \frac{1}{|\mathbf{Y}|} \sum_{d=1}^{|\mathbf{Y}|} (J_l(\mathbf{f}_l^t + \mathbf{g}_l^u)_d - (J_l(\mathbf{f}_l^t) + J_l(\mathbf{g}_l^u))_d)^2 \tag{5}$$

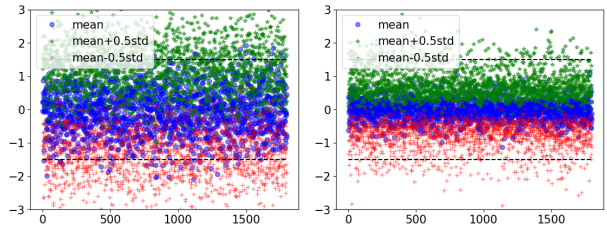
The sequence-level MSE loss is computed with the arithmetic average over the speech feature length T and text label sequence length U.

$$\mathcal{L}_{\text{MSE}} = \frac{1}{T} \frac{1}{U} \sum^{T} \sum^{U} \mathcal{L}_{\text{MSE}}^{t,u} \tag{6}$$

$\mathcal{L}_{\text{MSE}}$ is added to the HAT transducer loss as follows.

$$\mathcal{L}_{\text{HAT+MSE}} = \mathcal{L}_{\text{HAT}} + \mathcal{L}_{\text{MSE}} \tag{7}$$

We empirically investigated whether the output logits of HAT+MSE models satisfy the special condition for being mathematically-justified, more than that of HAT models. Figure 2 depicts a plot of $\mathbf{f}_l + \mathbf{g}_l$ vectors by dimension for the *in-house* speech recognition tasks when a Tanh function is used as an activation function of J_l and N_{J_l} is set to 0. Most $\mathbf{f}_l + \mathbf{g}_l$ mean values marked with blue dots are within the linear range



(a) *HAT* (b) *HAT+MSE*

Figure 2: *Mean and mean±0.5×standard deviation of* $\mathbf{f}_l + \mathbf{g}_l$ *vectors by dimension for the in-house (ott-contents) test set. Tanh is used as an activation function in the label joint networks and* $N_{J_l} = 0$*. Dashed lines indicate the linear range of a Tanh activation function.*

of a Tanh function [8], that is, [-1.5, 1.5]. However, the values of the HAT+MSE model tend to gather more densely as in Figure 2b. Specifically, the rate at which the $\mathbf{f}_l + \mathbf{g}_l$ values fit in the linear range has significantly increased from 43.88% to 69.58% by applying the proposed method. Moreover, a perplexity (PPL) of ILM in the models was reduced from 9.47 to 6.97. We explain the ILM performance in detail in Section 4.2.

4. Experiments

4.1. Experimental setup

NSR models were trained using the 1K-hours of Korean speech corpus, which was recorded at sample rate of 16 kHz with 16 bits of bit depth (*in-house*). The *in-house* corpus consists of voice command-related 1.1M utterances for various smart devices, such as mobiles, televisions, among others. We randomly sampled 1,000 utterances as the validation set. We also used the Librispeech corpus [26], which contains 960-hours of English speech data. We trained all NSR models using 40-dimensional mel-frequency filterbank features plus the log-energy and their delta and delta-delta features. The hop size was set to 10 ms, and the signal was windowed by 25 ms. The features were normalized by their means and variances either per utterance or per speaker for the *in-house* and Librispeech corpus, respectively.

We used eight layers of bidirectional long short-term memories [27] with 600 hidden units for transcription networks and three layers of unidirectional LSTMs with 512 hidden units for prediction networks. The sum of both network output vectors were used as an input of the joint network. For HAT, we used the shared transcription network and prediction network for J_b and J_l, and divided the output vectors by 1:9 for blank and label outputs. The output dimension of linear layers in J_b and J_l were 2,000, with the exception of last linear layers. J_l predicts 71 Korean graphemes for the *in-house* dataset and 2,001 subword units by byte-pair encoding [28] for Librispeech. RNNT models were developed using the same number of layers and hidden units for the transcription and prediction network as HAT models, and their joint networks consist of a linear layer with 2,000 units, a ReLU layer, a linear layer with vocabulary size units, and a softmax layer. The LMs were constructed with three hidden layers of 1,500 LSTM cells, resulting in a total number of parameters of 54M and 60M for the *in-house* and Librispeech corpus, respectively. Models were initialized by a Xavier uniform initializer with a fan_avg mode at scale 1.0. We used the ADAM optimizer [29] without learning schedulers. Models for the *in-house* data were trained for 25 epochs, and models for Librispeech were trained for 50 epochs. Beam search decoding [25] was used with the beam widths of 4 and 8 for the *in-house* and Librispeech corpus, respectively. When decoding the HAT and HAT+MSE models with eq (4), we used the range of λ_2 as [0.01, 0.65] and λ_3 as [max(λ_2 + 0.1, 0.5), 1.0] for the cases with LM adaptation. RNNT models were also decoded with eq (2) and [0.1, 0.9] was used for λ on applying LM adaptation. Experiment results in the following subsections were evaluated using λ_2, λ_3 and λ showing the best performance in each case. The penalization term in eq (2) was not used for any cases, that is, $\gamma = 0$. All experiments were conducted on NVIDIA™ A100 graphics processing units (GPUs).

For the *in-house* corpus, four test sets and the corresponding text corpus were used as evaluation tasks. The two test sets "stv-random" and "stv-difference," and their task specific text corpora were recorded from smart televisions. Here, "stv-random" is a test set randomly selected 1,002 utterances. "stv-difference" consists of 1,102 utterances that are differently recognized using a couple of WFST-based ASR models on the conventional ASR system [17]. The text corpus to construct LMs for the two sets contains 1.7M sentences. The other two test sets, "stv-command" and "ott-contents" are internally recorded utterances and consist of 1,000 television command-related utterances and 400 utterances related to content titles on streaming services, respectively. To develop LMs for the two sets, we sampled television-command-related 7K sentences and content-related 54K sentences each from the text corpus collected within the company. In the tables of Section 4.2 and 4.3, the "o" and "x" symbols marked in the LM section indicate the cases where LM adaptation has been used or not used, respectively.

4.2. Results on the *in-house* corpus

We compared the proposed models with RNNT and original HAT models in terms of word error rates (WERs) by varying the setups of J_l as in Table 1. The "Average" row exhibits the average of the 9 WERs of HAT and HAT+MSE models for each column. A training batch consists of 46K frames, and approximately 193K weight updates were performed according to the validation loss. We used 8 GPUs to train each model, and set the learning rate to 1.5e-4. The HAT and HAT+MSE models were composed of 56M, 60M and 64M parameters respectively, depending on the N_{J_l} values set to 0, 1, and 2. Additionally, the training time also increased to 13, 15, and 18 hours, respectively. The RNNT models were built with 60M of parameters and they required approximately 14 hours to train. For all cases, the proposed method achieved the lowest WERs with LMs and have shown at most 32.09% and 28.60% relative WER reductions compared with RNNT and HAT, respectively, for "ott-contents." We could not observe a notable relationship between N_{J_l} and WERs. We also examined ILM PPLs of HAT and HAT+MSE models as in Table 2. When measuring PPLs, transcriptions with value outside the 1.5 × interquartile range were excluded since the outliers could distort the result values of both HAT and HAT+MSE. The ILM PPLs of our models are significantly lower and more stable than those of HAT models.

As depicted in Table 3, we applied HAT+MSE for the pre-trained HAT models to investigate the possible application of our method for the existing HAT-based ASR systems. The pre-trained models were constructed by setting $N_{J_l} = 0$ and the post HAT+MSE (HAT+PMSE) method was used for 10 epochs. In most cases, HAT+PMSE improved the recognition accuracy of pre-trained HAT models, but the accuracy of HAT+PMSE models could not reach that of HAT+MSE.

Table 1: *Word error rates (WERs) of language model (LM) adaptation for in-house test sets according to label joint network J_l setups (Acti.: an activation function in J_l, N_{J_l}: the number of first layer block in J_l) and whether the constrained training is applied*

Model		stv-random				stv-different				stv-command				ott-contents			
LM		×		o		×		o		×		o		×		o	
RNNT		6.39		4.40		14.55		13.04		8.62		3.88		24.32		7.26	
Acti.	N_{J_l}	HAT	+MSE	HAT	+MSE	HAT	+MSE	HAT	+MSE	HAT	+MSE	HAT	+MSE	HAT	+MSE	HAT	+MSE
Sigmoid	0	6.10	5.02	4.16	3.87	15.16	14.27	12.45	12.22	8.55	8.09	3.67	3.62	24.92	23.40	6.01	5.74
	1	6.05	5.60	4.21	3.73	15.69	15.07	12.56	12.37	8.23	8.02	3.88	3.69	25.19	24.97	7.42	5.63
	2	6.03	5.31	4.21	3.78	15.37	14.74	12.66	12.22	8.48	7.66	3.82	3.53	24.81	24.49	5.96	5.85
ReLU	0	5.96	5.86	4.14	4.02	16.30	14.71	12.88	11.88	8.50	7.65	3.58	3.53	23.89	23.84	6.72	5.63
	1	6.34	5.26	4.09	3.76	15.56	14.23	12.96	11.54	7.99	7.34	3.76	3.45	24.27	23.73	6.39	4.93
	2	8.04	5.74	4.57	4.35	17.27	14.69	13.51	12.39	8.67	7.87	4.10	3.69	24.38	24.05	8.88	6.34
Tanh	0	5.57	5.09	3.83	3.71	16.13	15.01	12.41	12.35	8.40	7.80	3.82	3.58	25.35	23.24	7.10	5.85
	1	6.79	5.38	4.83	3.54	16.00	15.22	13.32	12.22	8.86	7.44	3.76	3.38	25.03	24.38	7.69	5.90
	2	6.58	5.84	4.40	3.97	15.88	15.81	12.88	12.16	8.55	8.43	3.76	3.60	25.35	24.97	7.48	6.34
Average	-	6.38	5.46	4.27	3.86	15.93	14.86	12.85	12.15	8.47	7.81	3.79	3.56	24.80	24.12	7.07	5.80

Table 2: *Perplexity of internal language model $J_l(g_l)$ with in-house test sets depending on label joint network J_l setups (Acti.: an activation function in J_l, N_{J_l}: the number of first linear layers in J_l) and whether the constrained learning is applied*

	HAT			HAT+MSE		
Acti. \ N_{J_l}	0	1	2	0	1	2
stv-random						
Sigmoid	4.90	10.37	4.78	3.14	3.02	3.60
ReLU	5.39	5.86	14.45	3.10	3.05	5.79
Tanh	3.32	3.87	3.99	3.03	3.11	3.71
stv-different						
Sigmoid	6.37	15.09	6.12	3.62	3.41	4.01
ReLU	6.77	7.86	25.30	3.54	3.45	5.93
Tanh	3.95	4.64	4.66	3.53	3.62	4.15
stv-command						
Sigmoid	5.73	11.42	6.64	3.40	3.35	3.97
ReLU	7.07	8.09	21.41	3.40	3.34	5.86
Tanh	3.79	4.33	4.30	3.41	3.41	3.96
ott-contents						
Sigmoid	19.99	79.12	19.71	6.72	6.35	6.49
ReLU	31.68	35.50	178.50	6.54	5.88	7.83
Tanh	9.47	11.81	11.74	6.97	6.89	7.27

Table 3: *Word error rates (WERs) of language model (LM) adaptation for the in-domain corpus when constrained learning is applied for pre-trained hybrid autoregressive transducer (HAT) models when the number of first layer blocks in the label joint networks is set to 0*

Acti.	LM	stv-rand.	stv-diff.	stv-comm.	ott-cont.
Sigmoid	×	5.64	15.14	8.06	24.21
	o	4.11	12.16	3.93	5.85
ReLU	×	5.76	14.57	8.11	24.05
	o	3.68	12.28	3.82	6.39
Tanh	×	5.52	15.10	8.09	24.32
	o	3.73	12.62	3.52	6.12

Table 4: *Word error rates (WERs) of language model (LM) adaptation for the Librispeech corpus depending on whether the constrained learning is applied*

Model	LM	dev-clean	dev-other	test-clean	test-other
RNNT	×	4.28	13.19	4.52	13.33
	o	3.72	11.38	3.75	11.60
HAT	×	4.38	13.35	4.66	13.59
	o	3.54	10.96	4.06	10.94
+MSE	×	4.36	12.97	4.43	13.16
	o	3.43	9.92	3.37	10.04

4.3. Results on the Librispeech corpus

Our models were also evaluated on the Librispeech corpus as in Table 4. Weight updates were conducted about 478K times according to the loss measured on "dev-clean" and "dev-other." A training batch contains approximately 36K frames and about 74-hours were required to train HAT and HAT+MSE models. Training RNNT models took about 69-hours. 5 GPUs were utilized to train each model, and the learning rate was set to 1.2e-4. We set N_{J_l}=1 and ReLU as an activation function for J_l. Each NSR model consists of 65M of parameters. HAT+MSE models exhibited lower WERs compared with HAT models over all evaluation sets and simultaneously minimize the accuracy degradation from RNNT models when LMs are not applied.

5. Conclusion

We proposed HAT+MSE as a novel training method. A MSE loss was used in addition to a HAT loss to encourage justified LM adaptation. Compared to related work, our method does not need structural changes of HAT models. Thus, it can be successfully applied to HAT models either from scratch or after regular HAT training. The prior estimation can be improved by devising a new structure of RNNT variant models.

6. References

[1] Y. He, T. N. Sainath, R. Prabhavalkar, I. McGraw, R. Alvarez, D. Zhao, D. Rybach, A. Kannan, Y. Wu, R. Pang, Q. Liang, D. Bhatia, Y. Shangguan, B. Li, G. Pundak, K. C. Sim, T. Bagby, S.-y. Chang, K. Rao, and A. Gruenstein, "Streaming end-to-end speech recognition for mobile devices," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 6381–6385.

[2] K. Kim, K. Lee, D. Gowda, J. Park, S. Kim, S. Jin, Y.-Y. Lee, J. Yeo, D. Kim, S. Jung, J. Lee, M. Han, and C. Kim, "Attention based on-device streaming speech recognition with large speech corpus," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2019, pp. 956–963.

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

[4] N. Morgan and H. Bourlard, "Continuous speech recognition using multilayer perceptrons with hidden Markov models," in *International Conference on Acoustics, Speech, and Signal Processing*, 1990, pp. 413–416 vol.1.

[5] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012.

[6] 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)*, 2019, pp. 434–441.

[7] X. Chen, Z. Meng, S. Parthasarathy, and J. Li, "Factorized neural transducer for efficient language model adaptation," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 8132–8136.

[8] 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)*, 2020, pp. 6139–6143.

[9] 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)*, 2021, pp. 243–250.

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

[11] Z. Meng, Y. Gaur, N. Kanda, J. Li, X. Chen, Y. Wu, and Y. Gong, "Internal language model adaptation with text-only data for end-to-end speech recognition," in *Proc. Interspeech 2022*, 2022, pp. 2608–2612.

[12] Z. Meng, T. Chen, R. Prabhavalkar, Y. Zhang, G. Wang, K. Audhkhasi, J. Emond, T. Strohman, B. Ramabhadran, W. R. Huang, E. Variani, Y. Huang, and P. J. Moreno, "Modular hybrid autoregressive transducer," in *2022 IEEE Spoken Language Technology Workshop (SLT)*, 2023, pp. 197–204.

[13] T. N. Sainath, R. Prabhavalkar, A. Bapna, Y. Zhang, Z. Huo, Z. Chen, B. Li, W. Wang, and T. Strohman, "JOIST: A joint speech and text streaming model for ASR," in *2022 IEEE Spoken Language Technology Workshop (SLT)*, 2023, pp. 52–59.

[14] C. Peyser, R. Huang, T. Sainath, R. Prabhavalkar, M. Picheny, and K. Cho, "Dual learning for large vocabulary on-device ASR," in *2022 IEEE Spoken Language Technology Workshop (SLT)*, 2023, pp. 245–251.

[15] M. Mohri, F. Pereira, and M. Riley, "Weighted finite-state transducers in speech recognition," *Computer Speech & Language*, vol. 16, no. 1, pp. 69–88, 2002. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0885230801901846

[16] J. Kim, J. Chong, and I. Lane, "Efficient on-the-fly hypothesis rescoring in a hybrid GPU/CPU-based large vocabulary continuous speech recognition engine," in *Proc. Interspeech 2012*, 2012, pp. 1035–1038.

[17] K. Lee, C. Park, I. Kim, N. Kim, and J. Lee, "Applying GPGPU to recurrent neural network language model based fast network search in the real-time LVCSR," in *Proc. Interspeech 2015*, 2015, pp. 2102–2106.

[18] E. Variani, E. McDermott, and G. Heigold, "A Gaussian mixture model layer jointly optimized with discriminative features within a deep neural network architecture," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 4270–4274.

[19] N. Kanda, X. Lu, and H. Kawai, "Minimum Bayes risk training of CTC acoustic models in maximum a posteriori based decoding framework," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 4855–4859.

[20] Ç. Gülçehre, O. Firat, K. Xu, K. Cho, L. Barrault, H. Lin, F. Bougares, H. Schwenk, and Y. Bengio, "On using monolingual corpora in neural machine translation," *CoRR*, vol. abs/1503.03535, 2015. [Online]. Available: http://arxiv.org/abs/1503.03535

[21] 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.

[22] A. Sriram, H. Jun, S. Satheesh, and A. Coates, "Cold fusion: Training seq2seq models together with language models," in *Proc. Interspeech 2018*, 2018, pp. 387–391.

[23] C. Shan, C. Weng, G. Wang, D. Su, M. Luo, D. Yu, and L. Xie, "Component fusion: Learning replaceable language model component for end-to-end speech recognition system," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 5361–5635.

[24] J. Park, S. Jin, J. Park, S. Kim, D. Sandhyana, C. Lee, M. Han, J. Lee, S. Jung, C. Han, and C. Kim, "Conformer-based on-device streaming speech recognition with KD compression and two-pass architecture," in *IEEE Spoken Language Technology Workshop, SLT 2022, Doha, Qatar, January 9-12, 2023*. IEEE, 2022, pp. 92–99. [Online]. Available: https://doi.org/10.1109/SLT54892.2023.10023291

[25] A. Graves, "Sequence transduction with recurrent neural networks," in *Representation Learning Workshop of International Conference on Machine Learning (ICML)*, 2012.

[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)*, 2015, pp. 5206–5210.

[27] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, p. 1735–1780, nov 1997. [Online]. Available: https://doi.org/10.1162/neco.1997.9.8.1735

[28] R. Sennrich, B. Haddow, and A. Birch, "Neural machine translation of rare words with subword units," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Aug. 2016, pp. 1715–1725.

[29] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Y. Bengio and Y. LeCun, Eds., 2015. [Online]. Available: http://arxiv.org/abs/1412.6980