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, the 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:

\[ Y^* = \arg \max_Y P(\hat{Y}|X) = \arg \max_Y P(X|\hat{Y}) \cdot P(\hat{Y}), \]  

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...
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 \( \lambda \) during decoding [2, 20, 24] such that

\[
Y' = \underset{\hat{Y}}{\arg \max} \log P(\hat{Y}|X) + \lambda \log P_{LM}(\hat{Y}) + \gamma R(X,Y),
\]

where \( R \) is an optional term scaled by \( \gamma \) 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 \( Y = \{y_1, y_2, \ldots, y_k, \ldots, y_{K-L} \} \), whereas RNNT models calculate the posteriors of \( \hat{Y} = Y \cup \{< b > \} \) through a single sub-network. The posterior at each node of a lattice \( P(\hat{Y}_u) = \overline{\gamma}_u(X, Y_{1:u-1}) \) is computed as follows:

\[
\begin{align*}
P^{b,u}_t &= \sigma(J_b(t_u^b + g^b_u)), \\
P^{l,u}_t &= (1 - P^{b,u}_t)\text{Softmax}(J_l(t_u^l + g^l_u)), ~ \bar{y}_u = < b >, \\
\end{align*}
\]

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

\[
\hat{Y}' = \underset{\hat{Y}}{\arg \max} \lambda_1 \log P(\hat{Y}|X) - \lambda_2 \log P_{LM}(B(\hat{Y}))
\]

\[
+ \lambda_3 \log P_{LM}(B(\hat{Y}))
\]

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, \( J_l \) 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.

#### 3.2. Constrained Learning with MSE loss

As explained in Section 2.2, \( J_l(t_u^l + g^l_u) \) should be approximately equal to \( J_l(t_u^l + g^l_u) \) 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 \( J_l(t_u^l + g^l_u) \) and \( J_l(t_u^l + g^l_u) \) and is computed as follows:

\[
\begin{align*}
\mathcal{L}_{MSE} &= \frac{1}{|Y|} \sum_{d=1}^{|Y|} (J_l(t_u^l + g^l_u) - J_l(t_u^l + g^l_u))^2 \\
\end{align*}
\]

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

\[
\mathcal{L}_{MSE} = \frac{1}{T} \sum_{t=1}^T \sum_{u=1}^U \mathcal{L}_{MSE}
\]

\[
\mathcal{L}_{HAT-MSE} = \mathcal{L}_{HAT} + \mathcal{L}_{MSE}
\]

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 \( f_l + g_l \) vectors by dimension for the in-house speech recognition tasks where a Tanh function is used as an activation function of \( J_l \) and \( N_{J_l} \) is set to 0. Most \( f_l + g_l \) mean values marked with blue dots are within the linear range.

---

*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.*
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 $f_l + g_l$ 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 $J_3$ and $J_1$, and divided the output vectors by 1.9 for blank and label outputs. The output dimension of linear layers in $J_3$ and $J_1$ were 2,000, with the exception of last linear layers. $J_3$ 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 $\lambda_3$ as $[0.01, 0.65]$ and $\lambda_3$ as $[\max(\lambda_3 + 0.1, 0.5), 1.0]$ for the cases with LM adaptation. RNNT models were also decoded with eq (2) and eq (2) and used for $\lambda$ on applying LM adaptation. Experiment results in the following subsections were evaluated using $\lambda_3$, $\lambda_4$, and $\lambda$ 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 WSTF-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. “stv-difference” is the result 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 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 $N_{J_3}$ 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 $N_{J_3} = 0$ 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

<table>
<thead>
<tr>
<th>Model</th>
<th>stv-random</th>
<th>stv-different</th>
<th>stv-command</th>
<th>ott-contents</th>
</tr>
</thead>
<tbody>
<tr>
<td>LM</td>
<td>x</td>
<td>o</td>
<td>x</td>
<td>o</td>
</tr>
</tbody>
</table>

Acti. \( N_{J_l} \)

<table>
<thead>
<tr>
<th></th>
<th>HAT</th>
<th>+MSE</th>
<th>HAT</th>
<th>+MSE</th>
<th>HAT</th>
<th>+MSE</th>
<th>HAT</th>
<th>+MSE</th>
<th>HAT</th>
<th>+MSE</th>
<th>HAT</th>
<th>+MSE</th>
<th>HAT</th>
<th>+MSE</th>
<th>HAT</th>
<th>+MSE</th>
<th>HAT</th>
<th>+MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sigmoid</td>
<td>0</td>
<td>6.10</td>
<td>5.02</td>
<td>4.16</td>
<td>3.87</td>
<td>15.16</td>
<td>14.27</td>
<td>12.45</td>
<td>12.22</td>
<td>8.55</td>
<td>8.09</td>
<td>3.67</td>
<td>3.62</td>
<td>24.92</td>
<td>23.40</td>
<td>6.01</td>
<td>5.74</td>
<td></td>
</tr>
<tr>
<td>ReLU</td>
<td>0</td>
<td>5.96</td>
<td>5.86</td>
<td>4.14</td>
<td>4.02</td>
<td>16.30</td>
<td>14.71</td>
<td>12.88</td>
<td>11.88</td>
<td>8.50</td>
<td>7.65</td>
<td>3.58</td>
<td>3.53</td>
<td>23.89</td>
<td>23.84</td>
<td>6.72</td>
<td>5.63</td>
<td></td>
</tr>
<tr>
<td>Tanh</td>
<td>0</td>
<td>5.57</td>
<td>5.09</td>
<td>3.83</td>
<td>3.71</td>
<td>16.13</td>
<td>15.01</td>
<td>12.41</td>
<td>12.35</td>
<td>8.40</td>
<td>7.80</td>
<td>3.82</td>
<td>3.58</td>
<td>25.35</td>
<td>23.24</td>
<td>7.10</td>
<td>5.85</td>
<td></td>
</tr>
<tr>
<td>Average</td>
<td>-</td>
<td>6.38</td>
<td>5.46</td>
<td>4.27</td>
<td>3.86</td>
<td>15.93</td>
<td>14.86</td>
<td>12.85</td>
<td>12.15</td>
<td>8.47</td>
<td>7.81</td>
<td>3.79</td>
<td>3.56</td>
<td>24.80</td>
<td>24.12</td>
<td>7.07</td>
<td>5.80</td>
<td></td>
</tr>
</tbody>
</table>

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


