

# IMPROVED MULTI-STAGE TRAINING OF ONLINE ATTENTION-BASED ENCODER-DECODER MODELS

*Abhinav Garg, Dhananjaya Gowda, Ankur Kumar, Kwangyoung Kim, Mehul Kumar, Chanwoo Kim*

Speech Processing Lab, AI Center, Samsung Research, Korea

## ABSTRACT

In this paper, we propose a refined multi-stage multi-task training strategy to improve the performance of online attention-based encoder-decoder (AED) models. A three-stage training based on three levels of architectural granularity namely, character encoder, byte pair encoding (BPE) based encoder, and attention decoder, is proposed. Also, multi-task learning based on two-levels of linguistic granularity namely, character and BPE, is used. We explore different pre-training strategies for the encoders including transfer learning from a bidirectional encoder. Our encoder-decoder models with online attention show  $\sim 35\%$  and  $\sim 10\%$  relative improvement over their baselines for smaller and bigger models, respectively. Our models achieve a word error rate (WER) of **5.04%** and **4.48%** on the Librispeech test-clean data for the smaller and bigger models respectively after fusion with long short-term memory (LSTM) based external language model (LM).

**Index Terms**— Attention based encoder-decoder models, online attention, multi-stage training, multi-task learning

## 1. INTRODUCTION

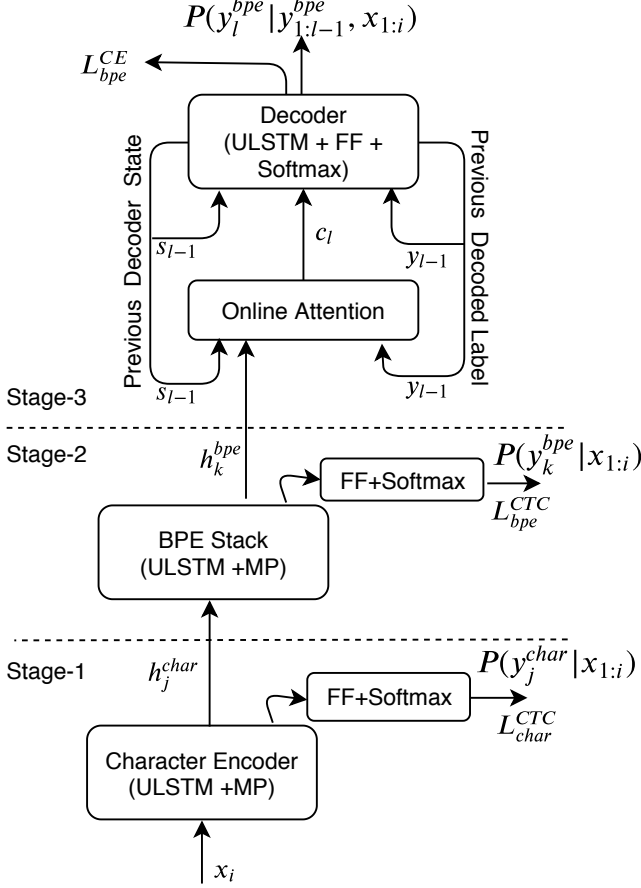
Recently, attention-based encoder-decoder (AED) models have gained popularity for developing end-to-end neural network based automatic speech recognition (ASR) systems [1, 2, 3]. One of the primary advantages of AED models is that the language information is tightly coupled into the decoder, obviating the need for an external language model (LM). AED models have been shown to perform better than other end-to-end models, namely, connectionist temporal classification (CTC) and recurrent neural network transducer (RNN-T) models [4]. With large amounts of transcribed data ( $>1000$  hrs), the AED models perform even better than the conventional DNN-HMM systems [1, 3]. This makes the AED models one of the best candidates for both server-side and on-device ASR applications. However, most of the best performing AED models use a bidirectional long short-term memory (BLSTM) based encoder and attend to the entire input sequence, which makes them not suitable for online decoding.

Conventional HMM-DNN based ASR systems which use

unidirectional long short-term memory (ULSTM), time-delay neural network (TDNN) or convolutional neural network (CNN) based architectures, as well as other end-to-end architectures namely, CTC and RNN-T models, are capable of online decoding. However, both the HMM-DNN and CTC models need an external LM to achieve good performance, especially for large vocabulary ASR systems. RNN-T models also need a separate predictor network to capture the language information. RNN-T models with small footprints which are also capable of online decoding have been recently shown to perform extremely well for on-device applications [5]. AED models with monotonic attention have also been proposed earlier for online decoding [6, 7]. A hard monotonic attention based encoder-decoder was first proposed in [6]. A much improved soft monotonic chunkwise attention (MoChA) based model was proposed in [7]. However, the performance of these monotonic attention models lags behind their full attention [8] counterparts, as well as other model architectures capable of online decoding. In this paper, our primary focus is to improve the performance of encoder-decoder models with online attention.

Online end-to-end models (CTC or attention-based) with deep neural architectures, when trained with random initialization, often have had difficulties in converging well [9, 10]. Several initialization strategies have been proposed to avoid this problem and improve model convergence. In [11], it was shown that using a pre-training strategy improves the generalization capability and hence performance of models in several seq2seq problems such as machine translation and abstractive summarization. Initializing a word-based CTC model with a pre-trained phone-based CTC model was found to be useful in [12]. Similarly, multi-task learning on hierarchical models has also been found to be effective, as in [13]. Using teacher-student transfer learning, to train a CTC based online ULSTM student from a BLSTM teacher model has been explored in [9].

Motivated by these pre-training and initialization techniques, we propose a multi-stage training strategy for online AED models. In terms of contribution: (a) We propose a new multi-stage training strategy for improving the accuracy of online AED models. (b) We explore three different methods for training the ULSTM based encoder models at two levels of linguistic granularity, character and byte pair encoding



**Fig. 1.** Block schematic of the proposed multi-stage multi-task character-to-BPE online AED model.

(BPE) [14] units. (c) For each of the above three methods, we propose different pre-training strategies that help the models converge better. These strategies are not limited to our methods, but can also be applied to any LSTM based ASR model in general.

This work is organized in the following manner. Section 2 gives an overview of attention based encoder-decoder models. Section 3 describes our strategy. Section 4 contains experimental details and results. And we conclude in Section 5.

## 2. ATTENTION BASED ENCODER-DECODER MODELS

A block schematic of the multi-stage multi-task character-to-BPE (C2B) online AED model proposed in this paper is shown in Fig. 1. The model consists of three distinct stages, a character encoder, a BPE stack, and an attention decoder. The character encoder consists of several ULSTM layers with an optional maxpool (MP) layer in-between. It converts an input sequence  $\mathbf{x} = \{x_i\}_{i=1}^T$  of length  $T$  into a sequence of hidden character embeddings  $\mathbf{h}^{char} = \{h_j^{char}\}_{j=1}^{T_c}$  of length

$T_c$ , and  $j = \lfloor i/T_s \rfloor$  where  $T_s$  denotes the overall time sub-sampling (TS) factor introduced by the MP layers. The BPE stack consists of a few additional ULSTM layers again with an optional MP layer in-between. This converts the input character embeddings into a sequence of BPE embeddings  $\mathbf{h}^{bpe} = \{h_k^{bpe}\}_{k=1}^{T_b}$  of length  $T_b$ . A single feed-forward (FF) projection layer followed by a softmax is used on top of the character encoder and BPE stack to convert the embeddings ( $\mathbf{h}^{char}$  and  $\mathbf{h}^{bpe}$ ) into probability distributions ( $P(y_j^{char}|x_{1:i})$  and  $P(y_k^{bpe}|x_{1:i})$ ) over the corresponding character and BPE target labels, respectively, as shown in the figure. The character encoder and the BPE encoder are trained using two different CTC losses on character and BPE targets ( $L_{CTC}^{char}$  and  $L_{CTC}^{bpe}$ ), respectively [15]. The CTC loss function for any chosen target labels  $lab$  (can be either  $char$  or  $bpe$  in our case) is given by

$$L_{CTC}^{lab} = -\log \sum_{\pi \in A(\mathbf{y}^{lab*}, \mathbf{h}^{lab})} \prod_j P(\pi_j^{lab} | x_{1:i}). \quad (1)$$

Here,  $\pi = A(\mathbf{y}^{lab*}, \mathbf{h}^{lab})$  denotes one of all possible alignments between the ground-truth target sequence  $\mathbf{y}^{lab*}$  and the encoder embeddings  $\mathbf{h}^{lab}$  either by repeating the output labels or by inserting a blank symbol.

An MP layer has no trainable parameters, but helps primarily in reducing the length of an input sequence by a pre-determined pooling factor. It has been observed that reducing the input sequence length progressively through the encoder helps in better convergence and accuracy of the AED models [2, 8, 16]. In this work, we use an overall time reduction factor of 2 for character encoder, and an additional reduction factor of 4 in the BPE stack, taking the overall reduction factor for the BPE encoder to 8. An MP layer wherever used has a reduction factor of 2 along the time-dimension unless specified otherwise. In this work, we use BPE based sub-word units as the final target labels for the AED model. Character labels are used as intermediate targets in a multi-task training strategy.

The attention decoder stage consists of an attention block and a decoder block. The attention block computes a context vector  $c_l$  based on the encoder embeddings  $\{h_{1:k}^{bpe}\}$ , previous predicted label  $y_{l-1}$ , and the previous decoder state  $s_{l-1}$ . Where,  $l$  is the target label produced at output after processing inputs  $x_{1:i}$ . In this paper, we use MoChA for the attention block to ensure online or streaming decoding [7]. The decoder consists of one or more ULSTM layers followed by a few fully connected FF layers and a softmax layer. The decoder uses the attention context vector, the previous state of the decoder, and the previous decoded label to generate the BPE label probabilities  $P(y_l^{bpe} | y_{1:l-1}^{bpe}, x_{1:i})$  at the softmax output. The AED model is trained with a cross-entropy (CE) loss  $L_{bpe}^{CE}$  between these label probabilities and the ground

truth target labels given by

$$L_{bpe}^{CE} = - \sum_{l=1}^L \log(P(y_l^{bpe*} | y_{1:l-1}^{bpe*} x_{1:i})), \quad (2)$$

where  $*$  denotes ground truth BPE target labels and  $L$  is the length of the target sequence.

### 3. MULTI-STAGE MULTI-TASK TRAINING OF ONLINE ATTENTION MODELS

Training CTC based encoder models have often been found to be difficult with random initialization [13]. Initializing these models with weights pre-trained based on frame-wise cross-entropy using prior alignment between the input and output sequences has shown better convergence of these models [17, 18]. In [9], a teacher-student transfer learning has been used to improve the convergence as well as the performance of character based CTC encoder models. In the case of AED models, it has been shown that a carefully designed layerwise pre-training strategy helps the models to converge better [2]. Similarly, a joint multi-task training using a CTC loss on the encoder and a CE loss on the attention decoder was proposed to improve the convergence and accuracies of AED models [19, 20].

In view of this, we propose a multi-stage multi-task training strategy to train our character-to-BPE (C2B) online AED models. The three different stages of our proposed training strategy are outlined below. In our work, *pre-initialization* would mean borrowing weights from an already well-trained model and *pre-training* would mean modifying some aspects of the network architecture during the initial part of the training.

#### 3.1. Stage-1: Training the CTC-Char encoder

In the first stage, we train the CTC based character encoder (CTC-Char) model. We use a deep stack of ULSTM layers to predict character sequence  $y_{char}$  given the input feature sequence  $x$ . We use one maxpool layer after the first ULSTM layer to reduce the encoder output length by a factor of 2. The character encoder output or embeddings for any given input  $x$  is given by

$$h_{char} = \mathcal{L}_n^u \circ \dots \circ \mathcal{L}_2^u \circ \mathcal{P} \circ \mathcal{L}_1^u(x) \quad (3)$$

where  $\mathcal{L}_i^u$  denotes the  $i^{th}$  ULSTM layer, and  $\mathcal{P}$  denotes a maxpool layer with a pooling factor of 2. The character encoder embeddings  $h_{char}$  are passed through a linear projection or FF layer followed by a softmax layer which captures the distribution of the character labels along with the blank symbol used in the computation of CTC loss. This network is trained to minimize the CTC loss between the softmax output and the ground truth label sequence  $y_{char}$ .

In order to improve the convergence of this model, we apply a layer-wise pre-training strategy moving from a higher time reduction factor to a lower factor, similar to [21]. We start with three ULSTM layers and we then incrementally add a new ULSTM layer approximately every  $1/3^{rd}$  of an epoch. Also, we use fixed MP layer after the first ULSTM layer and a floating MP layer before the last ULSTM layer, in order to maintain an overall time-reduction factor of 4 during the pre-training. The floating MP layer is removed after adding all the ULSTM layers reducing the overall time-reduction factor to two. Starting the training with all the ULSTM layers at once with a reduction factor of 2 leads to sub-optimal results, around 10% higher relative error as compared to our approach. Using higher reduction factor (8 or 16) is not feasible for character based CTC encoders as the input sequences tend to become shorter than the target sequences.

#### 3.2. Stage-2: Training the CTC-BPE encoder

In Stage-2, we start with the character encoder model trained in Stage-1, and convert it into a BPE encoder (CTC-BPE) model by using one of the three methods depicted in Fig. 2.

##### 3.2.1. C2B-Replace: Char-to-BPE encoder by replacing losses

In this method, we convert the pre-trained character encoder to BPE encoder by replacing the character CTC loss with BPE CTC loss. Two additional MP layers are added after the  $2^{nd}$  and  $3^{rd}$  ULSTM layers, taking the overall time-reduction factor to 8 including a factor of 2 from the character encoder. The BPE encoder output is given by

$$h_{bpe} = \mathcal{L}_n^u \circ \dots \circ \mathcal{P} \circ \mathcal{L}_3^u \circ \mathcal{P} \circ \mathcal{L}_2^u \circ \mathcal{P} \circ \mathcal{L}_1^u(x). \quad (4)$$

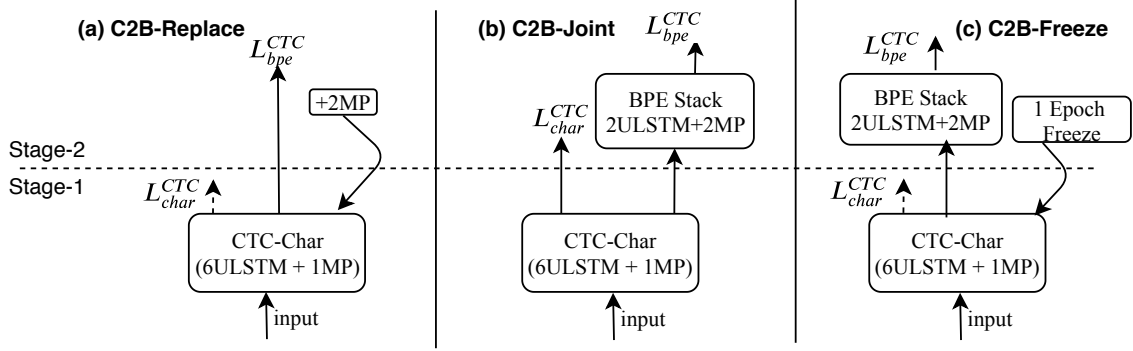
Similar to the pre-training strategy of the character encoder, we start with a high time-reduction factor of 32 using additional MP layers and gradually removing them one by one to a final factor of 8.

##### 3.2.2. C2B-Joint: Char-to-BPE encoder with joint losses

In this method, we convert the pre-trained character encoder to a BPE encoder by adding a BPE stack containing two ULSTM layers on top of character model and two additional MP layers to increase the overall reduction factor to 8. The BPE encoder output for this model is given by

$$h_{bpe} = \mathcal{L}_{bpe2}^u \circ \mathcal{P} \circ \mathcal{L}_{bpe1}^u \circ \mathcal{P} \circ (h_{char}). \quad (5)$$

The model is trained with the joint losses of character CTC and BPE CTC as shown in Fig. 2(b). There was no additional pre-training strategy applied for this method.



**Fig. 2.** Overview of methods to convert character encoder to BPE encoder. The dashed arrows for  $L_{char}^{CTC}$  in C2B-Replace and C2B-Freeze indicate that they will be removed during Stage-2.

### 3.2.3. C2B-Freeze: C2B-Joint with an initial freeze on character encoder

In this method, the conversion from character to BPE encoder is exactly similar to the C2B-Joint method, except for the absence of character CTC loss and freezing the character encoder for approximately one epoch after adding the additional BPE stack. An additional pretraining strategy was used by increasing the pooling factors of last two MP layers in the BPE stack from 2 to 4, making the overall reduction factor as 32 at the start of training. The overall reduction factor is gradually reduced to 8, by reducing the factors of the last two MP layers from 4 to 2, one at a time after half and one epoch of training, respectively. The BPE encoder output for this model is given by

$$h_{bpe} = \mathcal{L}_{bpe2}^u \circ \mathcal{P} \circ \mathcal{L}_{bpe1}^u \circ \mathcal{P} \circ \mathcal{F}(h_{char}), \quad (6)$$

where  $\mathcal{F}$  denotes a temporary freeze operation on the pre-trained character model.

### 3.3. Stage-3: Training the attention-decoder

In this stage, we use the BPE encoder trained from Stage-2 and build an end-to-end attention system by attaching a decoder with MoChA attention on top of it. Our decoder has a single ULSTM layer. We also use attention feedback and a maxout layer similar to [2]. This whole network is now trained with the cross-entropy loss between decoder output and the ground truth BPE target labels.

## 4. EXPERIMENTS AND RESULTS

### 4.1. Datasets and training

All our experiments are reported on the publicly available 1000 hours Librispeech corpus [22]. The *dev* and *test* sets consist of *clean* and *other* subsets. For our ULSTM encoder experiments (CTC-Char and CTC-BPE), we use combined dev set consisting of *dev-clean* and *dev-other* to evaluate our models. Our character vocabulary consists of 29 labels, all upper-case English alphabets, an unknown symbol, an end of sentence (eos) symbol, and space symbol to denote word boundaries. For BPE, we use a vocabulary of size 10K for all

**Table 1.** Performance (CER in %) on combined *dev* set of our CTC-Char model trained using C2B-Joint multi-stage method, as against the baseline CTC-Char models trained with and without pre-training (PT) as in Section 3.1, and BLSTM-to-ULSTM (B2U) teacher-student (TS) approach.

MODEL	CER(%) w/o LM
CTC-CHAR BASELINE	18.38
CTC-CHAR BASELINE + PT	16.40
B2U TS TRAINING	16.22
CTC-CHAR FROM C2B-JOINT	<b>13.61</b>

our experiments. The word error rate (WER) results are reported for our online AED models on the *test-clean* and *test-other* subsets.

We use 40-dimensional MFCC features extracted with librosa [23], which are then combined into variable batch sized chunks so as to keep the number of frames processed per training step constant. We use 6 ULSTM layers for the CTC-Char models each with 256 units, and a dropout of 0.3 [24]. We also use batch normalization after each ULSTM layer of the encoder [25]. Our decoder consists of a single ULSTM layer with 1024 units. We use RETURNN framework for all our experiments [26, 2, 27]. In all our experiments we use Adam optimizer [28] with an initial learning rate of 0.0008 and learning rate scheduling using a cross-validation set. We also use linear learning rate warm-up [1] along with gradient clipping. For generating the best hypothesis during the test time, we use beam search with beam size 12.

### 4.2. Effect of multi-stage training on CTC-Char model

In this section, we study the effect of our multi-stage training on the CTC character encoder model. For this, we use the CTC-Char model obtained after Stage-2 training of the C2B-Joint strategy, after removing the BPE stack, as our candidate model. The CTC-Char models from the other two strategies (C2B-Replace and C2B-Freeze) could not be evaluated owing to the removal of the CTC-Char losses during Stage-2 training. We compare our model against two baselines. (1) A 6 layer ULSTM baseline trained without any pre-training. (2) A 6 layer ULSTM baseline trained using similar pre-training strategy as ours in Section 3.1.

We also compare our multi-stage training strategy with

**Table 2.** Performance of the CTC-BPE models trained with different architectures and strategies in terms of BER(%) on the combined *dev* set. DNC denotes ‘Did not converge’.

MODEL	LOSS		INIT	PRE-TRAINING	BER(%) w/o LM
	$L_{CTC}^{char}$	$L_{CTC}^{bpe}$			
SINGLE-STAGE TRAINED MODELS					
BASELINE (8L ULSTM + MP)	NO	YES	RANDOM	NO	DNC
	YES	YES	RANDOM	NO	33.15
	NO	YES	RANDOM	PT	26.96
MULTI-STAGE TRAINED MODELS					
C2B-REPLACE	NO	YES	CTC-CHAR	NO	28.61
	NO	YES	CTC-CHAR	PT	24.91
C2B-JOINT	YES	YES	CTC-CHAR	NO	24.08
	YES	YES	CTC-CHAR	PT	DNC
C2B-FREEZE	NO	YES	CTC-CHAR	FREEZE	26.76
	NO	YES	CTC-CHAR	PT	DNC
	NO	YES	CTC-CHAR	FREEZE+PT	<b>23.63</b>

previously proposed teacher-student transfer learning based approach [9]. We train a BLSTM based CTC character encoder, exactly similar to training a ULSTM CTC character encoder as outlined in Section 3.1), as the teacher model. The teacher model had 256 BLSTM units in each layer in each direction, and achieved a character error rate (CER) of 7.6%. We then use frame-wise Kullback-Liebler (KL) divergence based teacher-student transfer learning to train the student model.

Table 1 shows the performance of different character encoder models in terms of CER on the combined *dev* set. Just with pre-training, the vanilla or baseline CTC-Char model achieves 10% relative improvement over the baseline. Our proposed multi-stage approach performs  $\sim 26\%$  better than the baseline without PT, and  $\sim 17\%$  better than the baseline with PT. During the training of C2B-Joint, character loss increases initially but ultimately settles down to a much lower value (26% relative) due to guidance from the BPE loss. Also, our model performs  $\sim 16\%$  better than model trained using teacher-student training without the need of any teacher model.

### 4.3. Performance of CTC-BPE encoder model

The performance of various CTC-BPE encoder models in terms of BPE label error rate (BER) on the combined *dev* set is given in Table 2. As a baseline, we use an encoder model with 8 layer ULSTM stack and 3 MP layers, trained in a single stage. When trained with a single BPE CTC loss, random initialization and without any pre-training, the model failed to converge. This may be due to the lack of sufficient gradient flow in the backward direction for the update of weights in an otherwise deep network. When the same model is trained with an additional character CTC loss after 6<sup>th</sup> layer, it converged to a BER of  $\sim 33.15\%$ . We obtain the best results of  $\sim 27\%$  BER when we trained this baseline model with only the BPE CTC loss, but with a pre-training similar to that

outlined in Section 3. We use this model as the baseline for our CTC-BPE experiments.

To show the effectiveness of pre-training techniques we also calculate BER for models trained with and without pre-training (only applicable for C2B-Replace and C2B-Freeze). We tried pre-training C2B-Joint with a strategy similar to C2B-Freeze but it did not converge for us. A more carefully crafted pre-training might be required for well convergence of joint loss models having CTC-Char initialization. Also, not freezing the CTC-Char weights leads to non-convergence as the initial epochs of CTC-BPE training adversely affect the well trained CTC-Char weights. Table 2 shows the corresponding results. For each method, pre-training gives better performance as compared to normal training. While training baseline with BPE CTC loss without pre-training does not converge, training similar architecture, C2B-Freeze with pre-initialization converges well even without pre-training suggesting that CTC-Char weights are indeed a very good initializer for CTC-BPE models. We achieve  $\sim 10.6\%$  relative improvement using C2B-Joint,  $\sim 7.6\%$  relative improvement using C2B-Replace and  $\sim 12.35\%$  relative improvement using C2B-Freeze over the baseline.

We note that each of our methods has some advantage over the others. C2B-Joint has one big advantage that it also improves the performance of CTC-Char online model. Joint loss between character and BPE helps both of them to converge to a better optimum. For applications where we need character outputs, we can simply discard the two additional layers and just use the character stack. C2B-Replace has an obvious advantage of being smaller in size and performing comparably to others. C2B-Freeze is the best performing method amongst the three.

### 4.4. Performance of online C2B-AED models

#### 4.4.1. Baseline

For comparison with our models, we train equivalent models without pre-initialization (i.e. random initialization). These

**Table 3.** WER(%) for our baseline AED online models trained with different strategies. Here PT-CTC refers to "CTC loss used only during pre-training"

MODEL	WER(%) w/o LM	
	CLEAN	OTHER
CE BASELINE	11.56	28.17
CE + CTC BASELINE	10.20	26.88
PT-CTC + CE BASELINE	<b>8.35</b>	<b>22.87</b>

models contain 8 ULSTM layers as the encoder, with a time reduction of two after each of the first three ULSTM layers. We use MoChA online attention and decoder architectures the same as mentioned before. Along with that, we use layer-wise pre-training similar to those in Section 3, where we start with an initial reduction factor of 32 and decrease it incrementally to 8 after one epoch.

For better convergence we try different training techniques including (1) Training with only CE loss (2) Training with combined CTC and CE loss [19] (3) Using combined loss for the first pass over the data (i.e. pre-training stage) and training with CE loss afterward. CTC loss is computed between the outputs of BPE encoder and target labels. And CE loss is computed between the outputs of decoder and target labels. We compare the performance of these baselines in Table 3. As we can see using CTC during initial epochs helps as it ensures better initial convergence for ULSTM encoder weights.

#### 4.4.2. Our proposed C2B online AED model

Table 4 shows the performance of our C2B online AED model as compared to the baselines mentioned above. We use the C2B-Joint model from Stage-2 with a BER of 24.08% to initialize our Stage-3 AED model primarily for its simplicity in training and good performance. Also, following our observations during training of baseline models, we discard the CTC losses used in C2B-Joint BPE encoder.

We also combine our models with an external RNN language model (RNN-LM) obtained from [21]. This RNN-LM was trained on officially available 800M word dataset of LibriSpeech and has a perplexity score of  $\sim 65.9$ . We use shallow fusion [29] for combining RNN-LM and our model. For test-clean, we achieve  $\sim 30\%$  relative improvement without external LM and over 35% relative with external LM. Our final C2B online AED model has only a total of around 20M parameters. And achieves **5.04** WER on LibriSpeech. This is the best WERs reported in literature on Librispeech corpus for online attention based encoder-decoder models, to the best of our knowledge.

#### 4.4.3. Experiments with bigger model

We also try our approaches on bigger models with ULSTM cell size as 1024 instead of 256 having around 4 times more parameters ( $\sim 86$ M parameters). The baseline for 1024 mod-

**Table 4.** WER(%) comparison of proposed C2B AED online model with baseline

MODEL	WER(w/o LM)		WER (w/ LM)	
	CLEAN	OTHER	CLEAN	OTHER
CE+CTC BASELINE	10.20	26.88	7.88	24.82
C2B AED	<b>7.24</b>	<b>20.80</b>	<b>5.04</b>	<b>16.61</b>

**Table 5.** WER(%) comparisons of bigger AED models (having encoder cell size 1024) with other architectures

MODEL	WER(w/o LM)		WER (w/ LM)	
	CLEAN	OTHER	CLEAN	OTHER
CE+CTC BASELINE	6.78	20.12	4.89	16.23
C2B AED	<b>6.25</b>	<b>18.33</b>	<b>4.48</b>	<b>15.93</b>
FCE (w/ EXT ALIGN)	<b>6.19</b>	<b>17.76</b>	<b>4.35</b>	<b>15.17</b>
OTHER COMPETITIVE ONLINE ARCHITECTURES				
TDNN-LSTM [32]	-	-	3.65	8.79
CTC/ASG [33]	3.86	11.95	2.95	8.79

els has same architecture and training as the baseline for 256 MoChA model used in table 4 except for the cell size.

We also explored initializing lower layers of the encoder with a model trained on an auxiliary loss. Specifically, we trained a 3 layer ULSTM model with frame level CE loss and phone target for each frame. These layer weights were then used to initialize lower layers of 5 layer ULSTM model which had an MP layer (pooling factor of 2) after the third ULSTM layer. The model is trained with CTC loss and phone sequence as the target. We now use these weights for initializing encoder which is 7 layer ULSTM model with MP layers after the third ULSTM layer (pooling factor of 2) and fifth ULSTM layer (pooling factor of 4). We also attach decoder and MoChA attention. The model is then trained with CE loss with the sequence of BPE as the target. We also use pre-training where we add a MP layer (pooling factor of 4) after the sixth ULSTM layer till half the data is seen. Frame level target and overall phone sequence for given audio were obtained from HMM/DMM systems using Montreal Forced Aligner [30] and Kaldi toolkit [31]. We call these experiments as FCE experiments.

Results for above 1024 sized models are summarized in table 5. Here also our strategy achieves over 7% relative improvement over the baseline without using LM. And about 8.4% relative improvement after doing shallow fusion using the same external LM as Section 4.4.2.

## 5. CONCLUSION

We present an effective multi-stage training strategy for online attention encoder-decoder models that achieves significant improvements over its baseline. All three stages of our 3-staged training strategy are themselves online end-to-end system at three different levels of granularity. Our strategy achieves significant improvements for all the stages involved. The performance of our online C2B AED models are the best WERs reported on Librispeech corpus for online attention based encoder-decoder models, to the best of our knowledge.

## 6. REFERENCES

- [1] Chung-Cheng Chiu, Tara N Sainath, Yonghui Wu, Rohit Prabhavalkar, Patrick Nguyen, Zhifeng Chen, Anjuli Kannan, Ron J Weiss, Kanishka Rao, Ekaterina Gonnina, 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.
- [2] Albert Zeyer, Tamer Alkhouli, and Hermann Ney, “Returnn as a generic flexible neural toolkit with application to translation and speech recognition,” in *ACL*, 2018.
- [3] Daniel S Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D Cubuk, and Quoc V Le, “SpecAugment: A simple data augmentation method for automatic speech recognition,” *arXiv preprint arXiv:1904.08779*, 2019.
- [4] Rohit Prabhavalkar, Kanishka Rao, Tara N Sainath, Bo Li, Leif Johnson, and Navdeep Jaitly, “A comparison of sequence-to-sequence models for speech recognition,” in *Interspeech*, 2017, pp. 939–943.
- [5] Yanzhang He, Tara N Sainath, Rohit Prabhavalkar, Ian McGraw, Raziel Alvarez, Ding Zhao, David Rybach, Anjuli Kannan, Yonghui Wu, Ruoming Pang, et al., “Streaming end-to-end speech recognition for mobile devices,” in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6381–6385.
- [6] Colin Raffel, Minh-Thang Luong, Peter J Liu, Ron J Weiss, and Douglas Eck, “Online and linear-time attention by enforcing monotonic alignments,” in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, 2017, pp. 2837–2846.
- [7] Chung-Cheng Chiu and Colin Raffel, “Monotonic chunkwise attention,” in *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. 2018, OpenReview.net.
- [8] William Chan, Navdeep Jaitly, Quoc Le, and Oriol Vinyals, “Listen, attend and spell: A neural network for large vocabulary conversational speech recognition,” in *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*. IEEE, 2016, pp. 4960–4964.
- [9] Suyoun Kim, Michael L. Seltzer, Jinyu Li, and Rui Zhao, “Improved training for online end-to-end speech recognition systems,” in *INTERSPEECH*, 2017.
- [10] Chengzhu Yu, Chunlei Zhang, Chao Weng, Jia Cui, and Dong Yu, “A multistage training framework for acoustic-to-word model,” in *Interspeech*, 2018, pp. 786–790.
- [11] Prajit Ramachandran, Peter J. Liu, and Quoc V. Le, “Unsupervised pretraining for sequence to sequence learning,” in *EMNLP*, 2016.
- [12] Kartik Audhkhasi, Bhuvana Ramabhadran, George Saon, Michael Picheny, and David Nahamoo, “Direct acoustics-to-word models for english conversational speech recognition,” in *INTERSPEECH*, 2017.
- [13] Kartik Audhkhasi, Brian Kingsbury, Bhuvana Ramabhadran, George Saon, and Michael Picheny, “Building competitive direct acoustics-to-word models for english conversational speech recognition,” in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 4759–4763.
- [14] Rico Sennrich, Barry Haddow, and Alexandra 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)*, Berlin, Germany, Aug. 2016, pp. 1715–1725, Association for Computational Linguistics.
- [15] Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber, “Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks,” in *Proceedings of the 23rd International Conference on Machine Learning*, New York, NY, USA, 2006, ICML ’06, pp. 369–376, ACM.
- [16] Daniel Povey, Vijayaditya Peddinti, Daniel Galvez, Pegah Ghahremani, Vimal Manohar, Xingyu Na, Yiming Wang, and Sanjeev Khudanpur, “Purely sequence-trained neural networks for asr based on lattice-free mmi,” in *Interspeech*, 2016, pp. 2751–2755.
- [17] Andrew Senior, Hasim Sak, and Kanishka Rao, “Flatstart-ctc: a new acoustic model training procedure for speech recognition,” in *ICASSP 2016*, 2016.
- [18] Hasim Sak, Andrew W. Senior, Kanishka Rao, Ozan Irsoy, Alex Graves, Françoise Beaufays, and Johan Schalkwyk, “Learning acoustic frame labeling for speech recognition with recurrent neural networks,” *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4280–4284, 2015.
- [19] Takaaki Hori, Shinji Watanabe, Yu Lin Zhang, and William Chan, “Advances in joint ctc-attention based end-to-end speech recognition with a deep cnn encoder and rnn-lm,” in *INTERSPEECH*, 2017.

- [20] Suyoun Kim, Takaaki Hori, and Shinji Watanabe, “Joint ctc-attention based end-to-end speech recognition using multi-task learning,” in *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2017, pp. 4835–4839.
- [21] Albert Zeyer, Kazuki Irie, Ralf Schlüter, and Hermann Ney, “Improved training of end-to-end attention models for speech recognition,” in *Interspeech*, Hyderabad, India, Sept. 2018.
- [22] 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)*, April 2015, pp. 5206–5210.
- [23] Brian McFee, Colin Raffel, Dawen Liang, Daniel PW Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto, “librosa: Audio and music signal analysis in python,” in *Proceedings of the 14th python in science conference*, 2015, vol. 8.
- [24] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov, “Dropout: A simple way to prevent neural networks from overfitting,” *Journal of Machine Learning Research*, vol. 15, pp. 1929–1958, 2014.
- [25] Sergey Ioffe and Christian Szegedy, “Batch normalization: Accelerating deep network training by reducing internal covariate shift,” in *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, Francis R. Bach and David M. Blei, Eds. 2015, vol. 37 of *JMLR Workshop and Conference Proceedings*, pp. 448–456, JMLR.org.
- [26] Patrick Doetsch, Albert Zeyer, Paul Voigtlaender, Ilia Kulikov, Ralf Schlüter, and Hermann Ney, “Returnn: The rwth extensible training framework for universal recurrent neural networks,” in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 5345–5349.
- [27] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al., “Tensorflow: A system for large-scale machine learning,” in *12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16)*, 2016, pp. 265–283.
- [28] Diederik Kingma and Jimmy Ba, “Adam: A method for stochastic optimization,” *International Conference on Learning Representations*, 12 2014.
- [29] Çağlar Gülçehre, Orhan Firat, Kelvin Xu, Kyunghyun Cho, Loïc Barrault, Huei-Chi Lin, Fethi Bougares, Holger Schwenk, and Yoshua Bengio, “On using monolingual corpora in neural machine translation,” *CoRR*, vol. abs/1503.03535, 2015.
- [30] Michael McAuliffe, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger, “Montreal forced aligner: trainable text-speech alignment using kaldi,” in *Proceedings of interspeech*, 2017, pp. 498–502.
- [31] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, Jan Silovsky, Georg Stemmer, and Karel Vesely, “The kaldi speech recognition toolkit,” in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. Dec. 2011, IEEE Signal Processing Society, IEEE Catalog No.: CFP11SRW-USB.
- [32] Kyu J. Han, Akshay Chandrashekar, Jungsuk Kim, and Ian R. Lane, “The CAPIO 2017 conversational speech recognition system,” *CoRR*, vol. abs/1801.00059, 2018.
- [33] Jason Li, Vitaly Lavrukhin, Boris Ginsburg, Ryan Leary, Oleksii Kuchaiev, Jonathan M. Cohen, Huyen Nguyen, and Ravi Teja Gadde, “Jasper: An end-to-end convolutional neural acoustic model,” 2019.