INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Multi-task multi-resolution char-to-BPE cross-attention decoder for end-to-end speech recognition

Dhananjaya Gowda, Abhinav Garg, Kwangyoun Kim, Mehul Kumar, Chanwoo Kim

Speech Processing Lab, AI Center, Samsung Research, Korea
{d.gowda, abhinav.garg, ky85.kim, mehul3.kumar, chanw.com}@samsung.com

Abstract

In this paper we present a new hierarchical character to byte-pair encoding (C2B) end-to-end neural network architecture for improving the performance of attention based encoder-decoder ASR models. We explore different strategies for building the hierarchical C2B models such as building the individual blocks one at a time, as well as training the entire model as a monolith in a single step. We show that C2B model trained simultaneously with four losses, two for character and two for BPE sequences help regularize the learning of character sequences as well as BPE sequences. The proposed multi-task multi-resolution hierarchical architecture improves the WER of a small footprint bidirectional full-attention E2E model on the 960 hours LibriSpeech corpus by around 15% relative and is comparable to the state-of-the-art performance of an almost 3 times bigger model on the same dataset.

1. Introduction

Attention-based encoder-decoder (AED) models have recieved a lot of attention in recent years compared to the conventional hybrid hidden markov model (HMM) and deep neural network (DNN) based automatic speech recognition (ASR) systems [1–7]. One of the main advantages of these end-to-end (E2E) AED models is that they capture the language information implicitly in their decoder, obviating the need for an external large language model (LM). This reduces the overall memory footprint of the ASR system making them the best candidates for on-device applications.

The performance of AED models are comparable or even better than the conventional HMM-DNN systems when trained on a large ($>$ 10000 hrs) corpus [3, 7]. However, the performance of E2E models are quite poor compared to conventional system when trained on smaller corpuses [1,2]. One of the main reasons for this is that the language information captured by E2E models gets better with more data. It has been recently shown that the E2E models can perform comparable to conventional HMM-LSTM systems on the 960h LibriSpeech corpus [5]. It has also been shown that the performance of these AED models can be further improved by using data augmentation methods and an external LM model [7, 8]. One of the main challenges for E2E AED models is to leverage on their smaller memory footprint and improve their performance to match that of larger models.

Several end-to-end model architectures have been proposed in the literature namely the connectionist temporal classifier (CTC) models [9] and recurrent neural network transducer (RNN-T) models [10]. On large scale models trained on large corpuses AED models perform significantly better than these two alternative E2E models. However, for small footprint ASR models trained on smaller transcribed corpuses there is no clear winner. CTC models can be used in conjunction with an external RNN-LM models trained on large text-only corpus which can give competitive results for on-device applications. RNN-T models also incorporate a predictor network analogous to an RNN-LM and is loosely coupled to its encoder-decoder architecture unlike the AED models with a tightly coupled LM. This gives them an additional advantage of using an externally pretrained predictor network using large text corpuses to boost it's performance. However, when there is no restriction on the amount of transcribed training data available AED models can be a better bet to achieve best performance at smallest possible memory footprint.

Traditionally, conventional HMM-DNN ASR systems have been modeling the distributions or emission probabilities of phonemes, or senones (tied-states of triphones) which are used along with word-level LMs to decode any given speech utterance. CTC-based models primarily use phoneme or character labels which are decoded using an external word-based LM. These models can be trained directly on word labels, but the model size becomes untenable for a large vocabulary ASR and has to be contrained to a manageable size of vacabulary as well as model. On the other hand, character based models keep the output layer size and hence the model size to the smallest, but the long-term dependencies on the output labels captured by the model tend to be weak. Recently, the use of intermediate subword units namely bype-pair encoded (BPE) labels or word-piece (WP) labels has shown a very good compromise between the model size and ability to capture long-term dependencies. BPE labels with a vocab size of 1K to 30K have been successfully used and show significant improvement over the character based AED models.

In this paper, we propose a new multi-task, multi-resolution, multi-head decoder based E2E AED architecture. In particular, the proposed architecture uses a hierarchical character-to-subword unit based E2E architecture for improving the performance of AED models. The use of character and BPE target labels at different levels of the E2E model provides for a multi-tasking approach. Combining encoder embeddings at different temporal resolutions that are ideal for character and BPE targets provide for a multi-resolutional approach.

The proposed architecture also combines a character-based AED model and a BPE based AED model with a common shared encoder stack which is jointly trained on four different losses, namely character-CTC encoder loss, BPE-CTC encoder loss, character-CE decoder loss and BPE-CE decoder loss. Apart from sharing a common character-level encoder stack, the proposed model has a separate BPE stack on top of the character-level encoder stack. Two different attention-decoder modules are attached to each of the encoder stacks. The embeddings from the two encoder stacks at different time resolutions and conditioned on two different tasks are combined into the BPE attention decoder. The embeddings are combined either as multi-head attention (MHA) or multi-head decoder (MHD)

http://dx.doi.org/10.21437/Interspeech.2019-3216

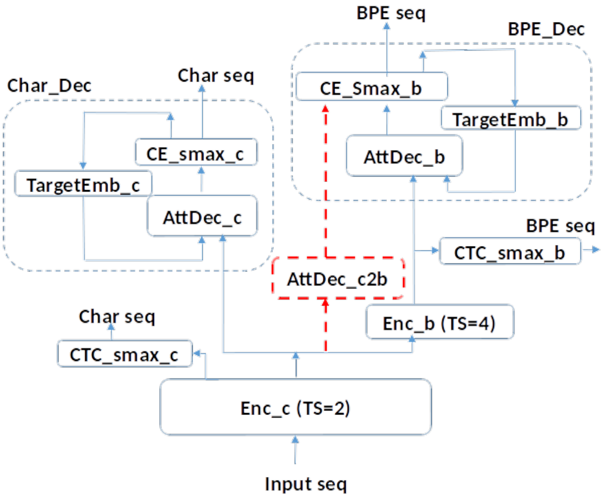

Figure 1: *Block schematic of the proposed hierarchical C2B E2E model.*

strategies [11, 12]. Several variants of the proposed architecture are explored with the primary goal of improving the performance of a small footprint model to match bigger models with marginal or no increase in model parameters.

2. Multi-task multi-resolution E2E models

In this section we give a brief overview of the proposed multi-task multi-resolution (MTMR) attention-based encoder-decoder (AED) model. A block schematic of the MTMR AED model is given in Fig. 1. The proposed architecture, based on the AED model used in [5], has a common encoder stack denoted 'Enc_c' shared between the character and BPE decoders ('Char_Dec' and 'BPE_Dec'). 'TS=2' denotes a temporal subsampling (TS) factor of two between the input and output sequences. This shared encoder is trained with a total of four different losses, namely the character CTC loss attached to this shared encoder through a softmax layer ('CTC_smax_c'), character CE loss on the character based attention-decoder, BPE CTC loss on the softmax layer ('CTC_smax_b') attached to the BPE encoder stack 'Enc_b', and the BPE CE loss on the BPE attention-decoder. At the decoder level, the embeddings from the character encoder stack which is at a different time resolution and trained on a different task is fed into a BPE decoder using a character-to-BPE (C2B) attention decoder (denoted 'AttDec_c2b' in Fig. 1).

The three main components of the proposed MTMR AED model are: 1. Char-to-BPE hierarchical encoders, 2. Multi-task attention decoders 3. Char-to-BPE cross-attention multi-resolution decoder.

2.1. Hierarchical char-to-BPE encoder stacks

An encoder stack consists of multiple layers of LSTM units with maxpool layers in-between any two LSTM layers. The maxpool layers are used to induce a temporal subsampling on the input feature sequence which helps the model to converge better [5]. The encoder stack can be pretrained independently, or jointly trained with the attention-decoder, by adding a softmax layer on top of the encoder stack and using a connectionist temporal classification (CTC) loss function given by $L_{ctc}(X,Y) = -\sum_{\pi \in B(Y,T)} \log P(\pi|X)$, where $X = [x_1, x_2, \ldots, x_T]$ is the input sequence of length T, $Y = [y_1, y_2, \ldots, y_L]$ is the output label sequence of length L. $B(Y,T)$ denotes a mapping function that provides all possible alignments between output Y and the input X of length T either by inserting a blank label or by repeating the output labels.

The encoder stack with a CTC softmax layer can by itself be used as a sequence-to-sequence model capable of generating or decoding sequences online with or without the help of an external language model. However, one major drawback with CTC models is the assumption of conditional independence of labels in computing $P(\pi|X)$ given by $P(\pi|X) = \prod_{t=1}^{T} p(\pi_t|x_t)$. Another major disadvantage of encoder-only CTC models is that the hidden embeddings at the output of encoder have a weak recurrence relation with the past feature vectors of the input sequence.

Motivated by the experiments on curriculum learning and multi-task learning, we propose to use two separate encoder stacks in an hierarchical manner. The bottom stack is trained to learn character sequences, while the upper stack learns to predict a BPE sequence. They use a temporal subsampling of 2 and 4, respectively, adding up to a total TS of 8 for the combined character-BPE stack.

2.2. Multi-task attention decoders

An attention based decoder partly alleviates the problem of conditional independence, by predicting the current output symbol based on the previous predicted symbol. It also uses an LSTM layer to capture recurrence relations in the output sequence and uses an attention mechanism to compute a weighted combination of the past and/or future encoder embeddings. An AED model tries to maximize the conditional probability $P(Y|X)$, by minimizing the cross-entropy loss given by $L_{CE} = -\sum_i \log P(y_i|X, y_{<i})$.

In the proposed MTMR architecture, we use two different attention decoders, each focusing on the task of decoding a character or a BPE sequence. Empirical evidence from our own experiments, and from the literature, suggests that the performance of a BPE based AED is better than a character based AED. However, the motivation for including a separate character based attention decoder stems from two aspects. One, empirical evidence suggests that the performance of the CTC-based encoder stack can improve signifcantly when it is trained as part of an AED model with a joint CTC and CE loss function. Another aspect which is not explored in this paper, but can be explored in the future, is to design a joint beam-search decoder which uses both the character as well as BPE attention decoders.

2.3. Char-to-BPE cross-attention decoder

Multi-head attention and multi-head decoders with heterogenous attentions seem to provide complimentary information for improving the decoder performance [3, 11, 12]. Motivated by the idea of MHA and MHDs, we propose a multi-resolution decoder to combine the embeddings from a character encoder and a BPE encoder which are trained to predict different granularities of linguistic information. This can be viewed as a variant of multi-head attention (MHA) [3, 11] and multi-head decoder (MHD) [12], but tries to combine information from components focusing on different tasks or linguistic information and also combine input feature embeddings at two different resolutions. We hypothesize that combining evidence at different resolutions of acoustic information as well as different linguistic granularities can improve the performance of the overall system. The proposed architecture is therefore a multi-task, multi-resolution, cross-attention (char-to-BPE in this case) based encoder-decoder model, and can be generalized easily to combine more number of tasks and resolutions.

3. Training strategies: curriculum vs multi-task learning

In moving from a character to BPE AED model there could be several possible strategies that can be explored. In this paper, we try to study the effectiveness of a purely curriculum based training strategy where the model learns one task at a time, as against a complex multi-task learning. Some strategies explored in this paper are briefly described below.

3.1. Curriculum learning: C2B with pretrained character model

This is a purely curriculum learning strategy, where a character AED model is trained from scratch using two joint CTC and CE character-level losses. The pretrained character AED is now used to initialize a BPE based AED model, with the two character-level losses being replaced by the two corresponding BPE-level losses. This is mainly to study the necessity for moving to a more complicated architecture proposed in this paper.

3.2. Curriculum + multi-task learning: C2B with a shared encoder stack and joint losses

This is a hybrid learning strategy where the model first learns to predict character sequences, followed by a joint multi-task learning to predict both character as well as BPE sequences. This architecture is similar to the block diagram shown in Fig. 1 but without the cross-attention block denoted 'AttDec_c2b'. In this case, the character AED model is pretrained with two character losses and then the BPE encoder stack as well as the BPE attention-decoder are added with two more additional BPE-level losses taking the total number of losses to four. Adding the C2B cross-attention decoder when moving from curriculum learning to multi-task learning makes the resulting model architecture exactly the same as shown in Fig. 1.

3.3. Multi-task multi-resolution cross-attention learning

This is a purely multi-task learning where there is no stage by stage curriculum learning. This model is exactly the same as depicted in Fig. 1 including the 'AttDec_c2b' cross-attention. This architecture combines information both at the encoder level and at the decoder level. The character encoder stack with a time subsampling factor of 2 is shared between the character and BPE AED models. At the decoder level, two different variantions are explored in utilizing the C2B cross-attention. One is similar to multi-head decoders where the information is combined at the final softmax layer as depicted in the figure. The other is to use a multi-head attention strategy where the context vectors from the BPE-attention and C2B cross-attention are combined into a single LSTM layer in the decoder.

4. Experiments and results

The experimental setup, dataset used and the results from various experiments carried out in this paper are presented in this section.

4.1. Dataset

All our experiments reported in this paper are carried out on the publicly available LibriSpeech corpus [13]. The training set consists of around 960 hours of transcribed audio data. The dev-clean and dev-other subsets each around 5 hours is used as our validation set. The final word error rate (WER) performance is calculated on the test-clean and test-other subsets of 5 hours each.

4.2. Experimental setup

All the experiments in this paper are carried out using the open source E2E ASR training toolkit ReturNN [14, 15]. The recipe and hyperparameters used for training our models are exactly same as the default recipe released by the authors of ReturNN unless otherwise specified [15]. The encoder stacks used in our experiments contain bidirectional LSTM layers with either 256 cells. A time subsampling (TS) factor of 2 is used for the character stack, and an additional TS factor of 4 is introduced in the BPE encoder stack. A BPE vocab size of 10K is used, while a character vocab size of 29 (digits are not included) is used. The decoder uses a single unidirectional LSTM layer with 1000 cells. A target embedding layer of dimention 621 is used on the one-hot target representations of the previous labels before being fed into the attention-decoder module. In all our experiments we use Adam Optimizer [16] with an initial learning rate of 0.0008 and a learning rate decay mechanism based on the model performance on a cross-validation set (dev-clean plus dev-other subsets from LibriSpeech). A linear learning rate warm-up [17] is used for the first few epochs, along with gradient norm clipping and gradient NaN filtering. A 40 dimensional mel-frequency cepstral coefficients (MFCC) input features are used where the zeroth cepstral coefficient is replaced with the RMS energy of the frame.

4.3. Results

The results from various hierarchical C2B experiments are discussed and presented in Tables 1 to 4.

4.3.1. Experiments with pretrained character AED model

The results from experiments which use a pretrained character level AED model to initialize the BPE-level AED model in Table 1. A character AED is trained with 256 units in each of the 6-layer encoder stack. A temporal subsampling factor of 2 is used after the bottom most layer. The model after training for around 12 epochs reaches a validation or dev-set character error rate (CER) of 2.75%. This model is now retrained as a BPE AED model by replacing the character CTC and CE losses with BPE CTC and BPE CE losses, respectively. The TS factor of the encoder stack was increased from 2 to 8 by adding time-pooling factors of 2 each to the top most two layers (2,2). A pretraining strategy with the time-pooling factors for the top two layers starting with (4,4), (4,2) and then (2,2) was used to help better convergence. Each of these time-pooling combinations were used for half an epoch. The models are then allowed to train a total of around 12 epochs. It was observed that the convergence of the models were much better when only the BPE-CE loss was used and the BPE-CTC loss was totally omitted during the retraining stage. This model is referred to as 'BFA-6L-C2B-2to1Loss' (M2).

The BPE-label error rates computed on the validation set which comprises of around 3000 utterances, 1500 randomly chosen from dev-clean and dev-other subsets. It can be seen

Table 1: *Performance of different C2B models on LibriSpeech dev-set in terms of BPE-label error rates (BER).*

Model	Dev BER %
M1: BFA-6L-Baseline	9.16
C2B with pretrained Char AED	
M2: BFA-6L-C2B-2to1loss	8.65
M3: BFA-6L-C2B-2to1loss (DecRinit)	8.79
C2B with an additional BPE stack	
M4: BFA-8L-C2B-4loss (no MHD)	8.58
M5: BFA-8L-C2B-MHD-4loss	8.75

Table 2: *Performance of different C2B models trained from scratch using cross-attention on LibriSpeech dev-set in terms of BPE-label error rates (BER).*

Model	Dev BER %
M1: BFA-6L-Baseline	9.16
C2B trained from scratch	
M6: BFA-6L-C2B-MHA-3loss	9.05
M7: BFA-6L-C2B-MHD-3loss	8.97
M8: BFA-6L-C2B-MHA-4loss	8.43
M9: BFA-6L-C2B-MHD-4loss	7.90
M10: BFA-8L-C2B-MHA-4loss	8.29
M11: BFA-8L-C2B-MHD-4Loss	**7.62**

Table 3: *Effect of adding more parameters (M12:BFA 8L no C2B multi-task learning) and vanilla MHD (M13, no multi- or cross-attention decoding) to the baseline model (M1).*

Model	Dev BER %
M1: BFA 6L Baseline	9.16
M12: BFA 8L	9.36
M13: BFA 6L MHD 2Loss	9.71

Table 4: *WER performance of the proposed 256u MTMR C2B models on LibriSpeech.*

Model	Test WER %	
	clean	other
M1: BFA 6L Baseline	5.31	16.92
M11: BFA 8L C2B MHD 4loss	4.54	13.96
+ RNN-LM SF	**3.45**	**11.25**
BFA 6L 1024u Baseline	4.26	14.36
+ RNN-LM SF	**3.34**	12.17

from the table that this model improves the baseline 6-layer bidirectional full-attention (BFA) model (M1) by ∼6.5% relative. As a variation, another BPE AED model (M3) was trained from the pretrained character AED model, by borrowing only the encoder stack while randomly initializing the BPE attention-decoder.

The bottom two rows in Table 1 correspond to experiments where the character encoder stack is left untouched and an additional BPE encoder stack of 2 layers is added on top with a TS factor of 4. Again a (4,4), (4,2) and (2,2) TS pretraining strategy is used to ensure better convergence of the models. In the first case (M4), there was no cross-attention from character embeddings to the BPE decoder. In the second case (M5), a cross-attention decoder was also present as shown in Fig. 1. It can be seen from the table that these experiments do improve the baseline performance from 5-6% relative.

4.3.2. Experiments with multi-task multi-resolution cross-attention

Based on empirical evidence it was seen that the models converged better when they are trained from scratch rather than using a pretrained character AED to initialize the MTMS cross-attention models. All the experiments described in this subsection are trained from scratch with either all 4 losses as shown in Fig. 1, or with only 3 losses by totally eliminating the character-based attention decoder. The character embeddings were fused into the BPE decoder either into the single LSTM layer as context vectors (referred to as multi-head attention (MHA)) or combined at the output softmax with separate LSTM layers for both the BPE attnetion and C2B cross-attention (referred to as multi-head decoder (MHD)). It can be seen from Table 2 that while the 3loss models (M6 & M7) are faster to train with fewer parameters and lesser training time computations, their performance is not as good as when using the character decoder and with all 4 losses (M8 & M9). Also, it can be seen that the MHD strategy gives significant improvements compared to the MHA strategy. The performance improvement over the baseline is even more impressive when a 8-layer encoder stack is used (M10 & M11) instead of 6 layers.

In order to investigate if the improvements are merely due to the addition of extra trainable parameters into the model, we carried out two more experiments. One is to train a new baseline with 8-layer encoder stack (M12). Another is to add a multi-head decoder feeding from the same BPE embeddings simulating primarily an ensemble effect (M13). It has been earlier reported in literature that such a strategy indeed improves the performance [12]. The results for these two experiments are given in Table 3. It can be seen that mere addition of extra trainable parameters does not always guarantee improved performance.

Table 4 summarizes the WER performance of the best C2B BFA model ('M11: BFA-8L-C2B-MHD-4Loss') with and without shallow fusion (SF) using an externally trained RNN-LM as used in [15]. It can be seen that the performance of our best 256u C2B AED model M11 (∼66M params) improves significantly compared to the baseline model (∼36M params) due to the addition of a multi-task multi-resolution C2B attention decoder. To the best of our knowledge these results are the new state-of-the-art for both with and without the LM for a smaller size BFA model of size 256 units in the encoder. After shallow fusion with an RNN-LM the performance of the proposed MTMR trained 256u smaller model is almost comparable to the ∼3 times bigger 1024u BFA model (∼187M params) for *test-clean* and even better for *test-other* test set.

5. Conclusions

In this paper, we proposed a new multi-task multi-resolution E2E ASR model that combines character and BPE level information in an heirarchical manner. Experimental results show that the proposed architecture gives an approximately 15% improvement in WER relative to the baseline BFA model with 256 LSTM cells. The results also show that the performance of the proposed small footprint MTMR AED model is comparable to the state-of-the-art performance on the LibriSpeech corpus given by an AED model with 1024 LSTM cells and is around 3 times larger than the proposed model. As far as we know this is new state-of-the-art for this smaller size model for LibriSpeech.

6. References

[1] J. K. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, "Attention-based models for speech recognition," in *Advances in neural information processing systems*, 2015, pp. 577–585.

[2] W. Chan, N. Jaitly, Q. Le, and O. 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.

[3] 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.

[4] T. N. Sainath, C.-C. Chiu, R. Prabhavalkar, A. Kannan, Y. Wu, P. Nguyen, and Z. Chen, "Improving the performance of online neural transducer models," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5864–5868.

[5] A. Zeyer, K. Irie, R. Schlüter, and H. Ney, "Improved training of end-to-end attention models for speech recognition," *CoRR*, vol. abs/1805.03294, 2018. [Online]. Available: http://arxiv.org/abs/1805.03294

[6] S. Sabour, W. Chan, and M. Norouzi, "Optimal completion distillation for sequence learning," *CoRR*, vol. abs/1810.01398, 2018. [Online]. Available: http://arxiv.org/abs/1810.01398

[7] C. Kim, M. Shin, A. Garg, and D. Gowda, "Improved vocal tract length perturbation for a state-of-the-art end-to-end speech recognition system," in *Interspeech*, 2019.

[8] 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," in *arXiv*, 2019. [Online]. Available: https://arxiv.org/abs/1904.08779

[9] 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*, ser. ICML '06. New York, NY, USA: ACM, 2006, pp. 369–376. [Online]. Available: http://doi.acm.org/10.1145/1143844.1143891

[10] K. Rao, H. Sak, and R. Prabhavalkar, "Exploring architectures, data and units for streaming end-to-end speech recognition with rnn-transducer," in *Automatic Speech Recognition and Understanding Workshop (ASRU), 2017 IEEE*. IEEE, 2017, pp. 193–199.

[11] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *NIPS*, 2017.

[12] T. Hayashi, S. Watanabe, T. Toda, and K. Takeda, "Multi-head decoder for end-to-end speech recognition," *CoRR*, vol. abs/1804.08050, 2018. [Online]. Available: http://arxiv.org/abs/1804.08050

[13] 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.

[14] P. Doetsch, A. Zeyer, P. Voigtlaender, I. Kulikov, R. Schlüter, and H. Ney, "RETURNN: the RWTH extensible training framework for universal recurrent neural networks," *CoRR*, vol. abs/1608.00895, 2016. [Online]. Available: http://arxiv.org/abs/1608.00895

[15] A. Zeyer, T. Alkhouli, and H. Ney, "RETURNN as a generic flexible neural toolkit with application to translation and speech recognition," *CoRR*, vol. abs/1805.05225, 2018. [Online]. Available: http://arxiv.org/abs/1805.05225

[16] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *CoRR*, vol. abs/1412.6980, 2014. [Online]. Available: http://arxiv.org/abs/1412.6980

[17] C.-C. Chiu, T. Sainath, Y. Wu, R. Prabhavalkar, P. Nguyen, Z. Chen, A. Kannan, R. J. Weiss, K. Rao, K. Gonina, N. Jaitly, B. Li, J. Chorowski, and M. Bacchiani, "State-of-the-art speech recognition with sequence-to-sequence models," 2018. [Online]. Available: https://arxiv.org/pdf/1712.01769.pdf