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

An analysis of local monotonic attention variants

André Merboldt[1], *Albert Zeyer*[1,2], *Ralf Schlüter*[1], *Hermann Ney*[1,2]

[1]Human Language Technology and Pattern Recognition, Computer Science Department, RWTH Aachen University, 52062 Aachen, Germany
[2]AppTek GmbH, 52062 Aachen, Germany

andre.merboldt@rwth-aachen.de
{zeyer, schlueter, ney}@cs.rwth-aachen.de

Abstract

Speech recognition using attention-based models is an effective approach to transcribing audio directly to text within an integrated end-to-end architecture. Global attention approaches compute a weighting over the complete input sequence, whereas local attention mechanisms are restricted to only a localized window of the sequence. For speech, the latter approach supports the monotonicity property of the speech-text alignment. Therefore, we revise several variants of such models and provide a comprehensive comparison, which has been missing so far in the literature. Additionally, we introduce a simple technique to implement windowed attention. This can be applied on top of an existing global attention model. The goal is to transition into a local attention model, by using a local window for the otherwise unchanged attention mechanism, starting from the temporal position with the most recent most active attention energy. We test this method on Switchboard and LibriSpeech and show that the proposed model can even be trained from random initialization and achieve results comparable to the global attention baseline.

Index Terms: encoder-decoder-attention, end-to-end, local attention, monotonic attention

1. Introduction

Deep neural networks (NN) have been used in various forms for automatic speech recognition. First as a discriminator in hybrid hidden Markov models (HMM) / NN models [1, 2], but more recently also combined with recurrent neural networks (RNN) to perform sequence transduction directly. The latter approach is referred to as end-to-end models where all steps in the speech recognition task are jointly optimized. One popular sequence-to-sequence model is an attention-based [3, 4] encoder-decoder [5, 6] architecture, which allocates attention to the encoder-processed input sequence during the output generation. Although first proposed for machine translation, the model has been applied to speech recognition, with promising results [7, 8], comparable to traditional hybrid HMM models given sufficient amount of training data.

In general, automatic speech recognition approaches need to handle the variation in speaking rate. This leads to different approaches to model the alignment between the speech audio and corresponding word sequences. Some end-to-end approaches such as CTC [9, 10] and RNN-Transducer [11, 12] treat the alignment as latent variables and then marginalize over all alignment possibilities. In contrast, attention-based models perform this implicitly by computing a soft alignment between input and output sequence.

If the model has access to the complete encoded input sequence, we refer to the attention mechanism as *global*. In the other case, where only a restricted subset of the sequence can be accessed, we speak of *local* attention [13]. In this work we study the latter approach, due to several benefits laid out below.

- Local attention can enable online decoding in contrast to globally computed attention, which requires the whole input sequence to be available.
- Variants which restrict the window over which the attention is distributed, consume less memory during both forward and backpropagation which enables us to train larger models.
- Due to the monotonic speech-text alignment, using a local attention variant, which enforces this constraint, may benefit the optimization process.
- Intuitively, any variant which is not trained on the full sequence, but only a limited subset of it can generalize better to longer sequences not seen during training.

So far, local attention variants for speech recognition have been mostly applied to small datasets such as TIMIT [14]. However, the results have been mixed with some improvements [15], but also worse results [13, 16] than their global attention counterparts.

We explore speech recognition using locally windowed attention-based models on the Switchboard 300h corpus (conversational English speech) [17] and on the Librispeech 960h corpus (read English books) [18]. Our approach can be interpreted as a transition from global to local attention without training a completely new model. We performed all experiments using the RETURNN framework [19]. All configs to train the models of this work are publicly available[1].

2. Global Attention Modeling

Sequence-to-sequence attention models [3, 13] typically consist of two distinct parts, the encoder and decoder networks. The encoder network is tasked with processing the audio sequence and building a representation the decoder can process. Encoder-decoder models typically employ recurrent neural networks (RNN) to process variable-length sequences. In terms of modeling, during the sequential output generation, each output depends on all previously produced tokens. Attention-based models expand the encoder-decoder approach by dynamically computing a distribution α_i over the encoded input $h_1^{T'}(x_1^T)$ in each decoding step i for encoder output of length T' and input sequence x_1^T. The weighted sum c_i represents the expectation over the encoder output with respect to α_i which in turn is computed by normalizing energy values $e_{i,t}$. Energy values

[1]https://github.com/rwth-i6/returnn-experiments/tree/master/2019-asr-local-attention

Copyright © 2019 ISCA

http://dx.doi.org/10.21437/Interspeech.2019-2879

intuitively relate the encoder output h_t and the current decoder state s_i. Various energy functions have been proposed, and we use the *additive* attention [3] as shown in Eq. (2).

$$c_i = \sum_{t=1}^{T'} \alpha_{i,t} \cdot h_t$$
$$\alpha_{i,t} = \text{softmax}_t(e_{i,t}) \quad (1)$$
$$e_{i,t} = v^\intercal \tanh(W_s s_i + W_h h_t) \quad (2)$$

Thus, the prediction for the word w_i includes the dynamically computed context vector c_i in each decoder step i.

$$p(w_1^N | x_1^T) = \prod_{i=1}^{N} p(w_i | w_1^{i-1}, c_i(x_1^T, w_1^{i-1}, c_1^{i-1}))$$

Global attention models may break both the monotonicity as well as locality property in the input-output relation. For machine translation [3], the former is a desired property, while for speech recognition the goal is to support this property as much as possible.

3. Related work

Our work on local heuristics is closely related to the *windowing* approach proposed by [4, 20]. However we extend and improve the idea by selecting a different heuristic and performing a systematic evaluation of the approach as well as comparing it to other proposed models. It can be observed that our proposed method provides a transition between global and a more local attention. This observation also allows for further refinement in the future where the approach is combined with more pretraining ideas to further improve both convergence and performance.

Segmental recurrent neural networks [21–23] divide the input sequence into a number of segments over which conditional probabilities are defined. They can be trained end-to-end and use the segmentation to explicitly model the alignment boundaries.

In another work [24], a model is proposed which alternates between processing input segments and producing output, thereby capitalizing on the aforementioned monotonicity property.

The *Neural Transducer* [25] operates on a non-overlaping fixed-size window block basis and predicts one or multiple labels for each. This was later also compared to global attention [26].

Hard attention [27] models use stochastic sampling to perform discrete decision whether to attend or not. This approach was extended in [28] by restricting the attention window locally to enforce monotonicity, while also enabling efficient training using backpropagation.

CTC [9] and the *RNN-Transducer* (RNN-T) [11,29] are also online-capable and strictly monotonic models. Among these models, attention models seems to perform slightly better, although the RNN-T gets close in performance [30]. A similar model to RNN-T is also the *Recurrent Neural Aligner* [31].

4. Local and Monotonic Attention

In this section, we examine some previously-proposed extensions to attention which typically aim at reducing computational cost or exploiting monotonicity properties. All next three publications are similar in nature as they employ local attention by computing energy values only on a window of the encoder outputs. Additionally all three learn a policy which should predict the next window position, either as relative position update or absolute position prediction. By using a Gaussian weighting of the window, the distribution parameters can be learned, in particular the mean which is then used to compute the next window center. Without this weighting, no gradient would be propagated to the policy predicting the next window position. The model descriptions include the notation of subscribed trainable vectors v or matrices W.

4.1. "Effective Approaches to Attention-based Neural Machine Translation"

Although proposed for machine translation, the proposed **local-p** model by [13] seems to fit speech recognition as well. The authors choose to use an absolute position prediction which predicts the window position with respect to the input length. This prevents the use of this model during online decoding as it requires the knowledge of the input sequence while also not accounting for sequences longer than those encountered during training. Additionally, it breaks the monotonicity property which we would like to use in speech recognition. The energy computation itself can be written as above in Eq. (2).

$$p_i = T' \cdot \text{sigmoid}\left(v_p^\intercal \tanh(W_p s_i)\right)$$
$$\alpha_{i,k_i} = \text{softmax}_k(e_{i,k_i}) \cdot \exp\left(-\tfrac{(k_i - p_i)^2}{2\sigma^2}\right)$$
$$k_i \in \{p_i - \tfrac{D}{2}, \dots, p_i + \tfrac{D}{2}\}$$

4.2. "Gaussian prediction based Attention"

Instead of computing the energy values between the current decoder state s_i and encoder states h_t within a defined window, the authors of [16] decided to compute the attention weights solely based on a re-normalized Gaussian weighting. By parameterizing the weighting function with a non-linear transformation of the decoder state, the relative position prediction can be learned. To allow online decoding, their energy computation truncates the weighting based on the predicted standard deviation times a constant K. Therefore their computation is limited to a window of encoder states h_k with $k \in \{1, \dots, \lfloor p_i + K\sigma_i \rfloor\}$ as:

$$p_0 = 0, \quad p_i = p_{i-1} + \Delta p_i$$
$$\Delta p_i = S_w \cdot \text{sigmoid}(v_p^\intercal \tanh(W_p s_i))$$
$$\sigma_i = D_w \cdot \text{sigmoid}(v_\sigma^\intercal \tanh(W_\sigma s_i))$$
$$e_{i,k} = \begin{cases} \exp\left(-\frac{(k-p_i)^2}{2\sigma_i^2}\right) & , k \in \{1, \dots, \lfloor p_i + K\sigma_i \rfloor\} \\ 0 & , \text{otherwise} \end{cases}$$
$$\alpha_{i,k} = \text{softmax}_k(e_{i,t})$$

4.3. "Local Monotonic Attention Mechanism for End-to-End Speech and Language Processing"

Even more expressive power is added by [15] which similarly to the previously-discussed models also employ a Gaussian weighting around the center p_i of the local attention window. Their computation window is then limited to h_k with $k \in \{p_i - 2\sigma, \dots, p_i + 2\sigma\}$ and the constant σ and $\alpha_{\mathcal{S},i}(k) = \alpha_{i,k}$ from Eq. (1).

$$\lambda_i = \exp\left(v_\lambda^\intercal \tanh(W_p s_i)\right)$$
$$\alpha_{\mathcal{N},i}(k) = \lambda_i \cdot \exp\left(-\tfrac{(k-p_i)^2}{2\sigma^2}\right)$$
$$c_i = \Sigma_{k=p_i-2\sigma}^{p_i+2\sigma} \left(\alpha_{\mathcal{N},i} \cdot \alpha_{\mathcal{S},i}\right) \cdot h_k$$

One limitation of the previous models is the limited relative position prediction, where the local window can only advance a limited number of encoder steps (S_w in Section 4.2) due to the use of a scaled sigmoid function. Thus they modified the prediction mechanism by using an unbounded function, for example $\exp(\cdot)$ for predicting the relative position advance.

5. Simple heuristic for local windowed attention

One contribution of this work is a simple and efficient heuristic to perform local attention. This method is based on the observation that our attention-based models tend to allocate attention mostly to the first frame of a spoken word in the case of BPE output units, as observed in [32]. The attention computation is restricted to a window of fixed size D, with its position determined by the $\arg\max$ of the previous attention weights α_{i-1}. This simple modification allows to use a conventionally trained global attention model to adjust to perform online decoding. Figure 1 provides a visualization of the proposed heuristic, which shows the adaptations the encoder had to learn in order to perform under the window constraint. We also show this modified version can be trained from random initialization, achieving similar performance to the globally-trained model.

$$p_0 = 0, \qquad p_i = \mathrm{argmax}_k(\alpha_{i-1,k})$$
$$\alpha_{i,k_i} = \mathrm{softmax}_k(e_{i,k_i}), \quad k_i \in \{p_i, \ldots, p_i + D - 1\}$$

5.1. Heuristic using median

Using the index of the median of the previous attention weights as heuristic has been proposed [4]. We denote $\bar{S}_i$ as the i-th value in the numerically ordered list of the finite sequence $\mathcal{S}$. Then the median value separating the higher and lower halves of $\mathcal{S}$ is defined as $\mathrm{median}(\mathcal{S}) := \bar{S}_{\lfloor |\mathcal{S}|/2 \rfloor}$. Consequently, $\mathrm{argmedian}(\mathcal{S})$ denotes the index of $\mathrm{median}(\mathcal{S})$ in the original sequence $\mathcal{S}$. Thus the attention computation can be written as follows:

$$\alpha_{i,k_i} = \mathrm{softmax}_k(e_{i,k_i}), \qquad k_i \in \{p_i - \tfrac{D}{2} + 1, \ldots, p_i + \tfrac{D}{2}\}$$
$$p_0 = 0, \quad p_i = \mathrm{argmedian}_k(\alpha_{i-1,k})$$

6. Experimental results

All models were implemented in the RETURNN framework [33]. Unless otherwise noted, the encoder is a 6-layered stack of bidirectional LSTM layers with each individual layer having 1000 units. The global attention baseline is derived from [8]. Byte-pair encoded (BPE) [34] sub-word units are used as targets, with a dictionary size of 2.000. In order to improve convergence and enable training of the wide and deep encoder, the same pretraining scheme is applied as described in the aforementioned publication. Via max-pooling in time between the encoder layers, we reduce the encoder hidden sequence length by a factor of 8. We have an input feature (either MFCC or Gammatone [35]) every 10 ms, i.e. an encoder output frame h_t represents 80 ms. The models were optimized using the Adam optimizer [36] with a learning rate scheduling starting at 10^{-3} which is halved once the loss is not decreasing on a held-out dataset.

6.1. Switchboard 300h

Switchboard [17] is a speech corpus with 300 hours of recorded conversational English telephone speech and their aligned transcriptions. We test on both subsets of Hub5'00, the easier Switchboard ("SWB") and the more difficult CallHome ("CH") part, as well as the Hub5'01 test set. We compare the different approaches in Table 1, where the window size is fixed at 10 encoder frames, corresponding to 0.8 seconds. Each variant was optimized individually, although their window size was fixed at 10 for a fair comparison. The variant **softplus** is the aforementioned local-p model using the softplus function as relative position prediction: $\mathrm{softplus}(x) = \log(1 + \exp x)$. Otherwise, it implements the local-p model mentioned in Section 4.1 with 256 hidden units for the position prediction. For the **additive combination**, the softplus activation is substituted with a scaled sigmoid with factor got exchanged for a scaled sigmoid and the Gaussian weighting is now added to the normalized energy values. **Gaussian** is the variant *only* employing Gaussian weighting of the encoder states, without any additional energy computation. **Scaled Gaussian** got inspired by the idea of scaling the Gaussian with an additionally learned parameter, however, it did not yield any improvements. We observe that with window size 10 (0.8 secs), all the local models perform worse than the global baseline. The closest ones are with the simple $\arg\max$ heuristic, and the *softplus* position prediction.

Table 2 shows results of both local heuristics using differ-

Table 1: *Attention variants trained from random initialization. Window size is 10 encoder frames (0.8 secs) across all windowed approaches. Each setting was tuned individually.*

Model	Variant	WER [%]			
		Hub5'00			Hub5'01
		Σ	SWB	CH	
Baseline	global	19.4	13.1	25.7	19.1
Heuristic	argmax	22.7	16.2	29.1	21.7
	argmedian	29.3	23.9	34.7	31.8
Position prediction	softplus	22.7	14.9	30.5	21.6
	additive comb.	38.7	35.6	41.8	42.7
	Gaussian	23.8	15.7	31.9	22.7
	scaled Gaussian	34.5	26.6	42.4	33.2

Table 2: *Models employing local heuristics.* Training *column indicates how much training was performed.* none *used pretrained weights from the global attention model.* retrain *models were trained for 20 more epochs using pretrained weights.* scratch *implies training from randomly initialized parameters.* − *indicates models which diverged.*

Model	Heuristic	Training	Hub5'00 WER [%]					
			Local attention window size					
			2	5	8	10	12	15
Hybrid		scratch	14.6					
Att.	global	scratch	20.4					
	argmax	none	−	41.8	25.6	23.0	21.8	20.9
		retrain	65.7	41.8	22.9	21.2	20.7	20.7
		scratch	−	27.9	26.0	25.1	20.4	**20.1**
	arg median	none	−	−	72.1	50.4	38.8	27.3
		retrain	74.6	69.3	54.1	45.4	41.4	26.2
		scratch	−	−	−	29.3	30.0	25.1

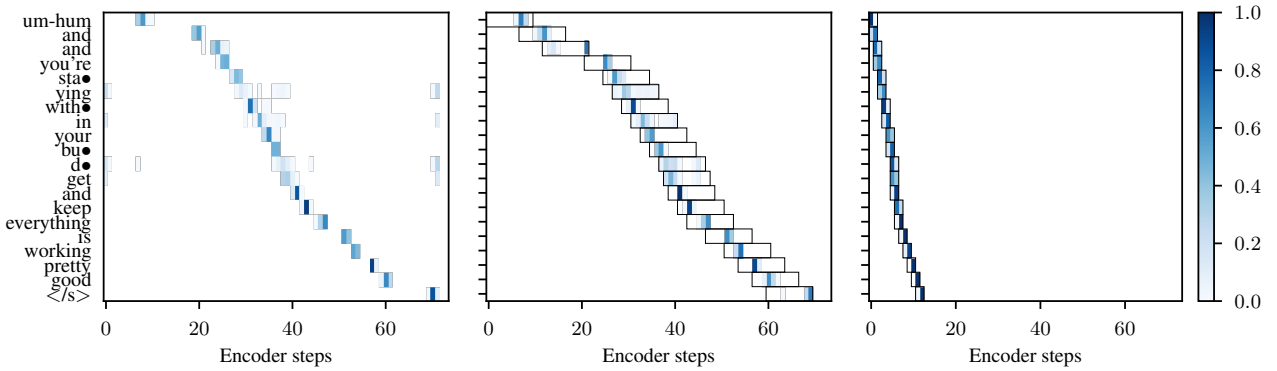

Figure 1: *Visualization of the attention weights comparing the* (left) *global attention model to* (middle) *the* $\arg\max$ *local heuristic using window size 10 (0.8 secs), and* (right) *with window size 2 (0.16 secs), all applied to the same sequence. The heuristic model was retrained using 20 epochs by employing the heuristic on top of the pretrained global attention model. The soft alignment was produced by enforcing the BPE outputs on the left and letting the model align those to the encoder frames.* • *indicate the BPE separator within words boundaries ("sta•" + "ying" → "staying").*

Table 3: *Experimental results for local heuristics on LibriSpeech.* recog *show models using pretrained weights of the global attention baseline, but performing recognition using the local heuristic.* retrain *indicate results of further training of 20 epochs employing the local heuristic (but not from scratch as in Table 2). A window size of 15 corresponds to 1.2 seconds.*

Model	Window size	WER [%]		
		test-clean	test-other	
		recog	recog	retrain
baseline (global)	∞	4.90	15.13	15.13
local (argmax)	8	14.59	26.16	22.75
	10	9.08	20.25	18.04
	12	6.69	17.63	16.69
	15	5.65	16.32	15.60
	20	5.65	15.96	15.64

ent window sizes and training variants. We observe that using the $\arg\max$ of the previous attention weighting as the left-most window position provides the best results. As expected, further expansions of the window improves the final performance, as well as retraining the model based on the global attention baseline. Interestingly, the local heuristic model trained from scratch using a larger window size (15 frames, 1.2 secs) outperforms the attention baseline by a small margin. It is also interesting to observe that the small window size 2 can work at all, because the attention process will never have a chance to see the full sequence. It can be seen in Fig. 1 that the encoder learned to compress all the relevant information to the initial few frames.

6.2. LibriSpeech 960h

LibriSpeech [18] is a large corpus developed for automatic speech recognition. It consists of 960 hours of read English speech, based on public domain audio books. We report results in Table 3 for experiments on the *test-clean* and the more challenging *test-other* dataset splits. The local heuristic performs robustly on LibriSpeech, and expectedly achieves higher word error rates using smaller window sizes.

7. Conclusion & Outlook

This work introduces a simple yet effective approach to local attention by employing a fixed-size window and using a simple $\arg\max$ heuristic for the start of the window. We show that this approach works well in practice and allows to facilitate previously well-performing global attention models in an online setting. In addition to reusing trained models, the windowing approach is also used for training from random initialization where it performs even slightly better than the global attention baseline.

On the challenging Switchboard 300h dataset for conversational telephone speech, we present promising results for a number of local attention models. Interestingly, the proposed simple local attention computation outperforms various previously explored more complex models.

The proposed approach has the drawback that it does not deal properly with silence. E.g. if the first spoken word appears after a longer time span than the window size, or some silence occurs in the sentence which is longer than the window size, no proper alignment is possible and it will be hard for the model to learn. Also, a hard decision is made on the window position, and this decision is not incorporated into the search procedure. A probabilistic formulation would solve this latter issue. A blank output label could also solve the first issue. We leave this open for future work.

8. Acknowledgements

This work has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 694537, project "SEQCLAS") and from a Google Focused Award. The work reflects only the authors' views and none of the funding parties is responsible for any use that may be made of the information it contains.

9. References

[1] H. Bourlard and N. Morgan, *Connectionist speech recognition: a hybrid approach.* Springer, 1994, vol. 247.

[2] A. J. Robinson, "An application of recurrent nets to phone probability estimation," *Neural Networks, IEEE Transactions on*, vol. 5, no. 2, pp. 298–305, 1994.

[3] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *Proceedings of the In-*

ternational Conference on Learning Representations (ICLR), San Diego, CA, 2015.

[4] D. Bahdanau, J. Chorowski, D. Serdyuk, P. Brakel, and Y. Bengio, "End-to-end attention-based large vocabulary speech recognition," in *ICASSP*, 2016, pp. 4945–4949.

[5] I. Sutskever, O. Vinyals, and Q. V. V. Le, "Sequence to sequence learning with neural networks," in *Advances in Neural Information Processing Systems 27*, Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Weinberger, Eds. Monteal, Canada: Curran Associates, Inc., 2014, pp. 3104–3112.

[6] K. Cho, B. van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder–decoder for statistical machine translation," in *EMNLP*. Doha, Qatar: Association for Computational Linguistics, October 2014, pp. 1724–1734. [Online]. Available: http://www.aclweb.org/anthology/D14-1179

[7] 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 *ICASSP*, 2018, pp. 4774–4778.

[8] A. Zeyer, K. Irie, R. Schlüter, and H. Ney, "Improved training of end-to-end attention models for speech recognition," in *Interspeech*, Hyderabad, India, Sep. 2018.

[9] A. Graves, S. Fernández, F. J. Gomez, and J. Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *Proceedings of the 23rd International Conference on Machine Learning*, 2006.

[10] A. Graves and N. Jaitly, "Towards end-to-end speech recognition with recurrent neural networks," in *ICML*, ser. Proceedings of Machine Learning Research, E. P. Xing and T. Jebara, Eds., vol. 32, Bejing, China, 22–24 Jun 2014, pp. 1764–1772. [Online]. Available: http://proceedings.mlr.press/v32/graves14.html

[11] A. Graves, "Sequence transduction with recurrent neural networks," *ICML: Representation Learning Workshop*, 2012.

[12] A. Graves, A. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *ICASSP*, May 2013, pp. 6645–6649.

[13] T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, 2015, pp. 1412–1421. [Online]. Available: http://aclweb.org/anthology/D/D15/D15-1166.pdf

[14] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren, "Darpa timit acoustic phonetic continuous speech corpus cdrom," 1993.

[15] A. Tjandra, S. Sakti, and S. Nakamura, "Local monotonic attention mechanism for end-to-end speech and language processing," in *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Asian Federation of Natural Language Processing, 2017, pp. 431–440. [Online]. Available: http://aclweb.org/anthology/I17-1044

[16] J. Hou, S. Zhang, and L.-R. Dai, "Gaussian prediction based attention for online end-to-end speech recognition," in *INTERSPEECH*, 2017.

[17] J. J. Godfrey, E. C. Holliman, and J. McDaniel, "Switchboard: Telephone speech corpus for research and development," in *ICASSP*. Washington, DC, USA: IEEE Computer Society, 2003, pp. 517–520. [Online]. Available: http://dl.acm.org/citation.cfm?id=1895550.1895693

[18] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An asr corpus based on public domain audio books," *ICASSP*, pp. 5206–5210, 2015.

[19] A. Zeyer, T. Alkhouli, and H. Ney, "RETURNN as a generic flexible neural toolkit with application to translation and speech recognition," in *Annual Meeting of the Assoc. for Computational Linguistics*, Melbourne, Australia, Jul. 2018.

[20] J. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, "Attention-based models for speech recognition," in *Advances in Neural Information Processing Systems 28*, 2015.

[21] L. Kong, C. Dyer, and N. A. Smith, "Segmental recurrent neural networks," *CoRR*, vol. abs/1511.06018, 2015.

[22] L. Lu, L. Kong, C. Dyer, N. A. Smith, and S. Renals, "Segmental recurrent neural networks for end-to-end speech recognition," in *INTERSPEECH*, 2016.

[23] E. Beck, M. Hannemann, P. Doetsch, R. Schlüter, and H. Ney, "Segmental encoder-decoder models for large vocabulary automatic speech recognition," in *Interspeech*, Hyderabad, India, Sep. 2018.

[24] L. Yu, J. Buys, and P. Blunsom, "Online segment to segment neural transduction," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2016, pp. 1307–1316. [Online]. Available: http://www.aclweb.org/anthology/D16-1138

[25] N. Jaitly, Q. V. Le, O. Vinyals, I. Sutskever, D. Sussillo, and S. Bengio, "An online sequence-to-sequence model using partial conditioning," in *Advances in Neural Information Processing Systems 29*, 2016.

[26] R. Prabhavalkar, T. N. Sainath, B. Li, K. Rao, and N. Jaitly, "An analysis of" attention" in sequence-to-sequence models." in *Interspeech*, 2017, pp. 3702–3706.

[27] C. Raffel, M.-T. Luong, P. J. Liu, R. J. Weiss, and D. Eck, "Online and linear-time attention by enforcing monotonic alignments," in *Proceedings of the 34th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, D. Precup and Y. W. Teh, Eds., vol. 70. International Convention Centre, Sydney, Australia: PMLR, 06–11 Aug 2017, pp. 2837–2846. [Online]. Available: http://proceedings.mlr.press/v70/raffel17a.html

[28] C.-C. Chiu and C. Raffel, "Monotonic chunkwise attention," in *ICLR*, 2018. [Online]. Available: https://openreview.net/forum?id=Hko85plCW

[29] E. Battenberg, J. Chen, R. Child, A. Coates, Y. G. Y. Li, H. Liu, S. Satheesh, A. Sriram, and Z. Zhu, "Exploring neural transducers for end-to-end speech recognition," in *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2017, pp. 206–213.

[30] R. Prabhavalkar, K. Rao, T. Sainath, B. Li, L. Johnson, and N. Jaitly, "A comparison of sequence-to-sequence models for speech recognition," 2017. [Online]. Available: http://www.isca-speech.org/archive/Interspeech_2017/pdfs/0233.PDF

[31] H. Sak, M. Shannon, K. Rao, and F. Beaufays, "Recurrent neural aligner: An encoder-decoder neural network model for sequence to sequence mapping." in *Interspeech*, 2017, pp. 1298–1302.

[32] E. Beck, A. Zeyer, P. Doetsch, A. Merboldt, R. Schlüter, and H. Ney, "Sequence modeling and alignment for lvcsr-systems," in *ITG Conference on Speech Communication*, 2018.

[33] A. Zeyer, T. Alkhouli, and H. Ney, "RETURNN as a generic flexible neural toolkit with application to translation and speech recognition," in *Annual Meeting of the Assoc. for Computational Linguistics*, Melbourne, Australia, Jul. 2018.

[34] 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, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*, 2016. [Online]. Available: http://aclweb.org/anthology/P/P16/P16-1162.pdf

[35] R. Schlüter, I. Bezrukov, H. Wagner, and H. Ney, "Gammatone features and feature combination for large vocabulary speech recognition," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Honolulu, HI, USA, Apr. 2007, pp. 649–652.

[36] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proceedings of the International Conference on Learning Representations (ICLR)*, San Diego, CA, 2015.