Sequence Modeling and Alignment for LVCSR-Systems

Eugen Beck, Albert Zeyer, Patrick Doetsch, André Merboldt, Ralf Schlüter, Hermann Ney

Lehrstuhl Informatik 6, RWTH Aachen University, Germany
Email: {beck,zeyer,doetsch,schlueter,ney}@cs.rwth-aachen.de

Abstract

Today, modeling automatic speech recognition (ASR) systems using deep neural networks (DNNs) has led to considerable improvements in performance, with word error rates being approximately halved compared to the status we had 10 to 15 years ago. Current state-of-the-art systems, at least if they are trained on moderate to medium amounts of training data, still follow the classical separation into language models and generative acoustic models. Acoustic modeling in these systems follows the so-called hybrid HMM approach. However, in the last years, many efforts were started to derive end-to-end models for ASR, which naturally follow the discriminative structure of neural networks. These include alternative solutions for the alignment problem underlying ASR, which in classical systems has been solved using hidden Markov models (HMMs). In this work we discuss and analyze two novel approaches to DNN-based ASR, the attention-based encoder–decoder approach, and the (segmental) inverted HMM approach. Experimental results are presented on the well-known Switchboard corpus and are compared against the standard hybrid approach, with specific focus on the sequence alignment behavior of the different approaches.

1 Introduction

The classic approach to sequence modeling in ASR has been hidden Markov models (HMMs) with a variety of emission distributions. Early approaches used mixtures of Gaussian distributions (GMMs) (e.g. [1]), but also neural networks (NNs) [2] to model the emission distributions. Prominent approaches to neural network based acoustic modeling comprise the hybrid HMM [2] and tandem [3, 4] approaches. The tandem approach augments the input features to standard Gaussian mixture HMMs by neural network outputs, in the hybrid HMM approach emission distributions are replaced by renormalized neural network outputs. For more details the reader is referred to [5]. Both tandem and hybrid HMM still are embedded in the standard HMM approach for sequencemodeling, i.e. for handling variations in speaking rate.

In contrast to this, connectionist temporal classification (CTC) [6] was a first approach to move away from the standard HMM paradigm. Technically it still is an HMM model with a special topology. In CTC an additional *blank*-label is added to the output of the neural network which has to be hypothesized between any two labels. However, unlike the hybrid approach, CTC is derived directly as a discriminative model and does not necessitate a renormalization using a state prior like in the hybrid approach. This causes CTC to output blank symbols predominantly. The actual labels in turn are most often reduced to single frames. This so-called peaky behaviour is analyzed and discussed in more detail in [7].

The *encoder-decoder framework with attention* has become the standard approach for machine translation [8–10] and many other domains. Recent investigations have shown promising results by applying the same approach for speech recognition [11–16]. The model is composed of two parts. An encoder network transforms the input features into a more useful representation. Afterwards a decoder network reads the output from the encoder to produce output symbols one-by-one. Often this is done with the attention mechanism. However attention does not require monotonicity in its implicit alignments. There are attempts to restrict the attention to become monotonic in various ways [17–24]. In this work we revisit and analyze our results from [16]. These models are without the modifications and extensions mentioned above.

Both CTC and attention do not (explicitly) hypothesize label (e.g. word or phoneme) boundaries, as done with HMMs. The last approach discussed here, the segmental, or inverted HMM approach introduces a discriminative model that still handles alignment using latent variables, similar to standard HMMs. To this end the input features are split into a sequence of segments and exactly one label is assigned to each segment. Often label-probabilities for a segment are extracted by sampling the encoder-output within the segment boundaries [25–28]. The publications [29] and [30] use a stochastic alignment. In [31] higher order features are integrated. The decoder sees a constant number of features per output label and is recurrent w.r.t. to the output labels. In previous work from our group [32, 33] inverted alignments were introduced as a form of deep segmental modeling. For these studies, the length model was static and did not incorporate information from the input features. In this work we will also revisit and analyze our results from [34].

All of the models discussed here introduce different approaches to handle the alignment problem in sequence modeling - either explicitly, as in HMM and inverted HMM / segmental modeling, or implicitly as in CTC or encoder-decoder approaches with attention. HMM based models with acoustic models that do not effectively have unlimited context or are trained with a given alignment, tend to align closely to the acoustic signals. In contrast to this, models that are trained from scratch and utilize unlimited input context tend to exhibit "peaky" behavior [7]. This is especially true for CTC based models. An attention system does not produce alignments in the usual sense, but provides attention weights on the encoder output frames to be considered for a label hypothesis. As will be shown, these are in general (in our system) short intervals at the beginning of each word. Segment-based systems produce an alignment that is similar to that of a classic HMM system. The only difference is that in a segment-based system boundaries between labels are hypothesized instead of framewise HMM-state-identities.

Whereas hybrid HMMs today still provide the state-of-the-art acoustic modeling technique in ASR irrespective of the amount of training data available, first attention-based encoder-decoder ASR systems where shown to outperform hybrid HMMs on large tasks [13, 16]. Current best systems for all of the approaches discussed here have in common that they include deep neural network modeling using bi-directional LSTMs. Although these certainly differ to some extent, e.g. in depth and width, or choice of target labels, a prominent difference between these is the alignment approach they rely on. Therefore, in this work we will present results for an attention-based and a segment-based system on a large vocabulary English telephony task. We will compare and discuss their alignments to the alignment produced by a well trained HMM-system.

2 Attention

Our model is similar to the architecture used in machine translation [35], except for encoder time reduction. The *encoder* operates on the input audio feature sequence and its purpose is to encode a higher-level representation of the input sequence. The *decoder* purpose is to generate output symbol by symbol to generate the whole output sequence. For each output symbol, it can access the encoder via the *attention* mechanism. Our models use *subword units* as output symbols, bypassing the use of a pronunciation lexicon. The subword units are created via *byte-pair encoding (BPE)* [36]. We use deep bidirectional long short-term memory (LSTM) neural networks [37] for the encoder and a single LSTM layer for the decoder.

2.1 Model

After every layer in the encoder, we optionally do max-pooling in the time dimension to reduce the encoder length. I.e. for the input sequence x_1^T, we end up with the encoder state

$$h_1^{T'} = \text{LSTM}_{\#\text{enc}} \circ \cdots \circ \text{max-pool}_1 \circ \text{LSTM}_1(x_1^T),$$

where $T' = \text{red} \cdot \text{T}$ for the time reduction factor red, and #enc is the number of encoder layers, with #enc ≥ 2. We use the MLP attention [8, 9, 17, 18, 38]. Our model closely follows the machine translation model presented by Bahar et al. [38] and Bahdanau et al. [8] and we use a variant of attention weight / fertility feedback [39], which is inverse in our case, to use a multiplication instead of a division, for better numerical stability. More specifically, the attention energies $e_{i,t} \in \mathbb{R}$ for encoder time-step t and decoder step i are defined as

$$e_{i,t} = v^\top \tanh(W[s_i, h_t, \beta_{i,t}]),$$

where v is a trainable vector, W a trainable matrix, s_i the current decoder state, h_t the encoder state, and $\beta_{i,t}$ is the attention weight feedback, defined as

$$\beta_{i,t} = \sigma(v_\beta^\top h_t) \cdot \sum_{k=1}^{i-1} \alpha_{k,t},$$

where v_β is a trainable vector. Then the attention weights are defined as

$$\alpha_i = \text{softmax}_t(e_i)$$

and the attention context vector is given as

$$c_i = \sum_t \alpha_{i,t} h_t.$$

The decoder state is recurrent function implemented as

$$s_i = \text{LSTMCell}(s_{i-1}, y_{i-1}, c_{i-1})$$

and the final prediction probability for the output symbol y_i is given as

$$p(y_i | y_{i-1}, x_1^T) = \text{softmax}(\text{MLP}_{\text{readout}}(s_i, y_{i-1}, c_i)).$$

In our case we use $\text{MLP}_{\text{readout}} = \text{linear} \circ \text{maxout} \circ \text{linear}$.

For a more comprehensive description of the encoder-decoder-attention model pursued here, the reader is referred to [16].

3 Inverted HMM

Just as in encoder-decoder frameworks with attention, we can define a hidden Markov model (HMM) that operates in a label-synchronous fashion [30, 33]. Instead of assigning a generalized triphone state to each of the observations by introducing a latent variable describing the state sequence through the HMM, we now use the latent variable to model the segmentation end-points (or label boundaries) t_0^N. Each output label w_n is hereby encoded over the frames within its segment $x_{t_{n-1}}^{t_n}$, where we assume $t_0 = 0, t_N = T$ (coverage) and $t_n < t_m \Leftrightarrow n < m$ (monotonicity). The model is derived as follows from Bayes' decision rule:

$$\begin{aligned}
\arg\max_{w_1^N} p(w_1^N | x_1^T) &= \arg\max_{w_1^N} \left\{ \sum_{t_0^N} p(w_1^N, t_0^N | x_1^N) \right\} \\
&= \arg\max_{w_1^N} \left\{ \sum_{t_0^N} \prod_n p(w_n, t_n | w_1^{n-1}, t_0^{n-1}, x_1^T) \right\} \\
&\quad \text{(first order Markov assumption)} \\
&= \arg\max_{w_1^N} \left\{ \sum_{t_0^N} \prod_n p(w_n, t_n | w_{n-1}, t_{n-1}, x_1^T) \right\} \\
&\quad \text{(maximum approximation for segmentation)} \\
&= \arg\max_{w_1^N, t_0^N} \left\{ \prod_n p(w_n, t_n | w_{n-1}, t_{n-1}, x_1^T) \right\}
\end{aligned}$$

From this base form we can derive two variants, which we will call Variant A / B. The first variant models the joint distribution without label feedback:

$$p(w_n, t_n | w_{n-1}, t_{n-1}, x_1^T) = p(w_n, t_n | t_{n-1}, x_1^T)$$

The second variant uses a label dependent length model with a framewise class posterior distribution:

$$\begin{aligned}
p(w_n, t_n | w_{n-1}, t_{n-1}, x_1^T) = p(w_n | t_{n-1}, x_1^T) \\
\cdot p(t_n | w_n, t_{n-1}, x_1^T)
\end{aligned}$$

In the second variant (variant B) we use a geometric-like factorization of the length distribution using a termination probability $p(end | t, w_n, x_1^T)$ with a single sigmoid unit per label:

$$p(t_n | w_n, t_{n-1}, x_1^T) = \left(\prod_{t=t_{n-1}+1}^{t_n - 1} 1 - p(end | t, w_n, x_1^T) \right) \cdot p(end | t_n, w_n, x_1^T)$$

During decoding we search for the most likely sentence hypothesis, where we have to take different sentence lengths N into account:

$$\begin{aligned} r : x_1^T \rightarrow w_1^N &= \arg\max_{N,w_1^N} p(N, w_1^N | x_1^T) \\ &= \arg\max_{N,w_1^N} p(N|x_1^T) \cdot p(w_1^N|N, x_1^T) \end{aligned}$$

We would like to point out that in tasks like speech or handwriting recognition it is useful to keep the language model $p(w_1^N)$ on the level of output symbols (words), but use smaller units (characters, mono-/(tied) triphones) for the acoustic model. Thus, the language model can be trained on a larger text corpus and during decoding, the acoustic model scores and language model scores are combined in a log-linear fashion (mixture of experts).

4 Experimental results

We show results on the 300h Switchboard English Telephony task. We use 40-dimensional Gammatone input-features [40], extracted by RASR [41].

4.1 Attention

We use a *pretraining scheme* applied on the encoder, which grows the encoder in layer depth, as well as decreases the initial high encoder time reduction factor [16]. Results are shown in Table 1. We observe that our attention model performs better on the easier Switchboard subset of the dev set Hub5'00, where it is the best end-to-end model we know. On the harder CallHome part, it also performs well compared to other end-to-end models but the relative difference is not as high.

model	LM	label unit	WER[%]	
			SWB	CH
LF MMI, 2016 [42]	4-gram	CDp	9.6	19.3
hybrid	4-gram	CDp	9.8	19.0
hybrid	LSTM	CDp	**8.3**	**17.3**
CTC[1], 2014 [43]	RNN	chars	20.0	31.8
CTC, 2015 [44]	none	chars	38.0	56.1
	RNN	chars	21.4	40.2
CTC, 2017 [45]	none	chars	24.7	37.1
	n-gram	chars	19.8	32.1
CTC[2], 2017 [45]	word RNN	chars	14.0	**25.3**
attention, 2016 [46]	none	chars	32.8	52.7
	5-gram	chars	30.5	50.4
	none	words	26.8	48.2
	3-gram	words	25.8	46.0
attention, 2017 [15]	none	chars	23.1	40.8
attention	none	BPE 10K	13.5	27.1
		BPE 1K	13.1	26.1
	LSTM	BPE 1K	**11.8**	25.7

Table 1: Comparisons on Switchboard 300h. The hybrid HMM/NN model is a 6 layer deep bidirectional LSTM. The attention model has a 6 layer deep bidirectional LSTM encoder and a 1 layer LSTM decoder. CDp are (clustered) context-dependent generalized triphone states. Byte-pair encoding (BPE) are sub-word units. SWB and CH are from Hub5'00. [1]added noise from external data. [2]added the lexicon, i.e. also additional data.

4.2 Segmental model

The systems in this work were trained over fixed segment boundaries that were obtained from the Viterbi alignment that was also shown in the comparison.

In Table 2 we see results for The Hub5'00 evaluation corpus. As expected our baseline system performs better when output labels are more fine-grained. A similar trend can be observed for segment-based models. The newly presented models do not yet reach the performance of our baseline system, so more work has to be done to examine possible improvements. One limiting factor for us was training time. The training time increases proportionally to the maximum window size (for Variant A).

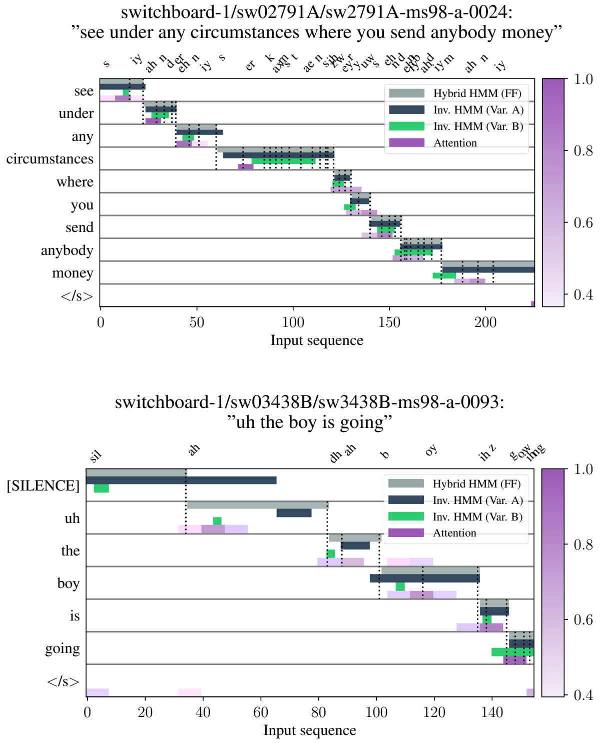
Model	label unit	DS	LN	CH	SWB	Σ/2
hybrid	1-st. monophn.	no	N/A	27.3	14.0	20.6
	1-st. CART 1501	no	N/A	23.6	12.1	17.8
	3-st. CART 9001	no	N/A	21.6	11.0	16.3
Var. A	1-state monophn.	no	no	42.5	22.1	32.3
		yes	yes	34.5	17.6	26.1
	3-st. monophn.	no	yes	34.9	19.3	27.4
	1-st. CART 1501	yes	no	29.6	15.0	22.3
Var. B	3-st. CART 9001	no	no	25.6	13.0	19.3

Table 2: Results on the Hub5'00 evaluation corpus in terms of %WER, the number behind CART is its number of leafs, X-st. = X state, DS=downsampling, LN=length normalization, CH=Callhome, SWB=Switchboard

5 Analysis of alignment behavior

All three models presented here can be regarded as different ways to structure information that is spread out over variable-length frame sequences w.r.t. the output / target labels. Classic HMM systems traverse an automaton by assigning state-occupation probabilities to each frame. The segmental model / inverted HMM stays very close to this model, but does not allow state repetitions. Instead boundaries between states are hypothesized explicitly. The attention model represents the most radical departure from the classic HMM model. It gathers evidence from the encoder by computing a soft alignment. As these attention weights are computed independent of the next output label identity, they cannot adapt to differences in length of the evidence for that particular output label. Thus the encoder in an attention system has to "cooperate" with the decoder to move evidence into fixed sized time intervals.

In Figure 1 we computed a forced alignment for different systems on two utterances from the training corpus. These two were picked because they exhibit behavior that we found was characteristic also for other utterances. Nonetheless this is mainly anecdotal evidence for the most part. One clear observation is that the segmental model A aligns very similarly to the hybrid system, which is unsurprising since it was trained with this alignment. In the second utterance we can observe that "uh" was aligned with a much shorter segment. This is due to the fact that


switchboard-1/sw02791A/sw2791A-ms98-a-0024:
”see under any circumstances where you send anybody money”

switchboard-1/sw03438B/sw3438B-ms98-a-0093:
”uh the boy is going”

Figure 1: Alignments by different systems of two utterances from the training-set

the window size is limited and thus the model cannot align very long single-phoneme labels. Segment model B was also trained to imitate the hybrid alignment, but it seems that this is only partially successful. We do not have a clear explanation for this yet. As mentioned before, the alignment model cannot discriminate between short and long output labels when it computes the attention weights. Thus we see that even for long words like “circumstances" it only attends to a short part of the input. But for other cases (“the" in the second utterance) the attention weights are more widespread. Even wider than real length of the acoustic realization as suggested by the alignment of the hybrid system.

In addition we analyzed the relation between the length of the attention window and the length of the recognized word. In Figure 2 we show 2 statistics depending on the recognized word. One is the standard deviation within the attention weights σ_i, the other is the distance between the preceding and following words. Both are presented in form of a contour plot.

$$\mu_i = \sum_{t=1}^{T'} t \cdot \alpha_{i,t},\ \sigma_i = \sqrt{\sum_{t=1}^{T'} (t-\mu_i)^2 \cdot \alpha_{i,t}},\ d_i = \frac{\mu_{i+1} + \mu_{i-1}}{2}$$

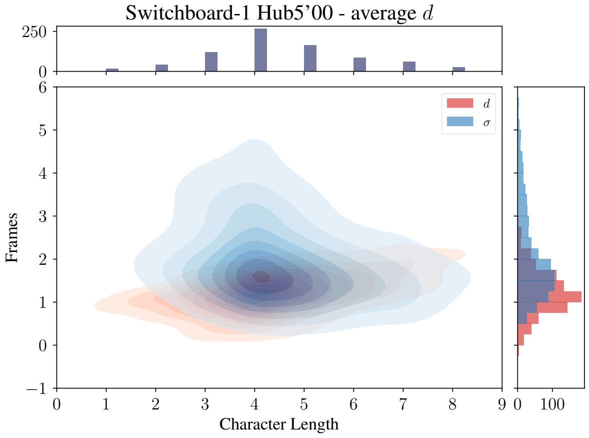


Figure 2: Word length in characters vs. attention size and word distance

As suggested by the anecdotal evidence above, and as expected by the fact that the attention interval computation does not use knowledge about the hypothesized label, we can observe here that the length of the attention window is not strongly influenced by the length of the recognized word in terms of its number of characters. For Figure 2 we computed the average for each sub-word unit in the Switchboard-1 Hub5’00 evaluation set over 4461 sequences. However, it should be noted that shorter subwords counter-intuitively can show longer attention intervals, than longer subwords.This observation is not yet well-understood and will be furthe investigated. In contrast to the average attention inverval length, the average distance between the attention intervals of adjacent subwords do seem to follow the actual word length in terms of characters, i.e. the positioning of the attention center seems to compensate for the relative constancy of the attention interval length. But it is important to note here that silence is not hypothesized by the attention system and thus might distort the results. Another detail worth mentioning is that we used the character length of words for the measurement above due to simplicity, even though it may not be an accurate representation of the spoken word length.

6 Conclusions

We have presented new results for two recent approaches to sequence-to-sequence modeling: Attention based systems and segmental models. These both promise to be a more elegant approach, but in the field of speech recognition have not yet reached parity on all tasks. More work needs to be done and we think analysis of the alignment behavior is one key aspect that can guide future development as the contribution of the attention model in the encoder-decoder approach, as well as the length modeling in the segmental/inverted HMM approach are more powerful components than the corresponding transition models in the standard/hybrid HMM approach. We have thus compared the alignments generated by the different systems and in case of attention analyzed its behavior in some aspects more systematically. We have also shown that some variants of segmental models perform alignment in a similar way as hybrid systems.

7 Acknowledgments

The research was partially supported by a Google PhD fellowship grant and has received funding from the European Research Council (ERC) under the European Union’s Horizon 2020 research and innovation programme (grant agreement No 694537) and under the Marie Skłodowska-Curie grant agreement No 644283. The work reflects only the authors’ views and neither the European Research Council Executive Agency nor Google is not responsible for any use that may be made of the information it contains.

References

[1] F. Jelinek, *Statistical methods for speech recognition*. MIT press, 1997.

[2] H. Bourlard and C. J. Wellekens, "Links between Markov Models and Multilayer Perceptrons," in *Advances in Neural Information Processing Systems I* (D. Touretzky, ed.), pp. 502–510, San Mateo, CA: Morgan Kaufmann, 1989.

[3] H. Hermansky, D. P. W. Ellis, and S. Sharma, "Tandem connectionist feature extraction for conventional hmm systems," in *2000 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.00CH37100)*, vol. 3, pp. 1635–1638 vol.3, 2000.

[4] F. Grezl, M. Karafiat, S. Kontar, and J. Cernocky, "Probabilistic and bottle-neck features for lvcsr of meetings," in *2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07*, vol. 4, pp. IV–757–IV–760, April 2007.

[5] R. Schlüter, P. Doetsch, P. Golik, M. Kitza, T. Menne, K. Irie, Z. Tüske, and A. Zeyer, "Neural networks in automatic speech recognition - a paradigm change?," in *Fortschritte der Akustik - DAGA 2018, 44. Jahrestagung für Akustik*, (Munich, Germany), Mar. 2018.

[6] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *ICML*, pp. 369–376, ACM, 2006.

[7] A. Zeyer, E. Beck, R. Schlüter, and H. Ney, "CTC in the context of generalized full-sum HMM training," in *Interspeech*, (Stockholm, Sweden), pp. 944–948, Aug. 2017.

[8] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

[9] M.-T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," *arXiv preprint arXiv:1508.04025*, 2015.

[10] Y. Wu, M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey, *et al.*, "Google's neural machine translation system: Bridging the gap between human and machine translation," *arXiv preprint arXiv:1609.08144*, 2016.

[11] W. Chan, N. Jaitly, Q. V. Le, and O. Vinyals, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," in *ICASSP*, 2016.

[12] P. Doetsch, A. Zeyer, and H. Ney, "Bidirectional decoder networks for attention-based end-to-end offline handwriting recognition," in *International Conference on Frontiers in Handwriting Recognition*, (Shenzhen, China), pp. 361–366, Oct. 2016.

[13] C.-C. Chiu, T. N. Sainath, Y. Wu, R. Prabhavalkar, P. Nguyen, Z. Chen, A. Kannan, R. J. Weiss, K. Rao, K. Gonina, *et al.*, "State-of-the-art speech recognition with sequence-to-sequence models," *arXiv preprint arXiv:1712.01769*, 2017.

[14] E. Battenberg, J. Chen, R. Child, A. Coates, Y. Gaur, Y. Li, H. Liu, S. Satheesh, A. Sriram, and Z. Zhu, "Exploring neural transducers for end-to-end speech recognition," in *ASRU*, (Okinawa, Japan), pp. 206–213, Dec. 2017.

[15] S. Toshniwal, H. Tang, L. Lu, and K. Livescu, "Multitask learning with low-level auxiliary tasks for encoder-decoder based speech recognition," in *Proc. Interspeech*, pp. 3532–3536, 2017.

[16] A. Zeyer, K. Irie, R. Schlüter, and H. Ney, "Improved training of end-to-end attention models for speech recognition," *arXiv preprint arXiv:1805.03294, submitted to Interspeech 2018*, 2018.

[17] J. Chorowski, D. Bahdanau, K. Cho, and Y. Bengio, "End-to-end continuous speech recognition using attention-based recurrent nn: first results," *arXiv preprint arXiv:1412.1602*, 2014.

[18] 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*, pp. 5067–5075, 2016.

[19] R. Aharoni and Y. Goldberg, "Morphological inflection generation with hard monotonic attention," *arXiv preprint arXiv:1611.01487*, 2016.

[20] C. Raffel, T. Luong, P. J. Liu, R. J. Weiss, and D. Eck, "Online and linear-time attention by enforcing monotonic alignments," *arXiv preprint arXiv:1704.00784*, 2017.

[21] C.-C. Chiu and C. Raffel, "Monotonic chunkwise attention," *arXiv preprint arXiv:1712.05382*, 2017.

[22] 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)*, vol. 1, pp. 431–440, 2017.

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

[24] J. Hou, S. Zhang, and L. Dai, "Gaussian prediction based attention for online end-to-end speech recognition," in *Proc. Interspeech*, pp. 3692–3696, 2017.

[25] G. Zweig and P. Nguyen, "A segmental CRF approach to large vocabulary continuous speech recognition," in *2009 IEEE Workshop on Automatic Speech Recognition Understanding*, pp. 152–157, Nov 2009.

[26] O. Abdel-Hamid, L. Deng, D. Yu, and H. Jiang, "Deep segmental neural networks for speech recognition," in *Interspeech 2013*, pp. 1849–1853, 2013.

[27] Y. He and E. Fosler-Lussier, "Efficient segmental conditional random fields for one-pass phone recognition," in *Interspeech 2012*, pp. 1898–1901, 2012.

[28] Y. He and E. Fosler-Lussier, "Segmental conditional random fields with deep neural networks as acoustic models for first-pass word recognition," in *Interspeech 2015*, pp. 2640–2644, 2015.

[29] L. Kong, C. Dyer, and N. A. Smith, "Segmental recurrent neural networks," in *Intern. Conf. on Learning Representations (ICLR)*, 2016.

[30] 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*, pp. 385–389, 2016.

[31] M. Ratajczak, S. Tschiatschek, and F. Pernkopf, "Frame and segment level recurrent neural networks for phone classification," in *Proc. Interspeech 2017*, pp. 1318–1322, 2017.

[32] P. Doetsch, S. Hegselmann, R. Schlüter, and H. Ney, "Inverted HMM - a Proof of Concept," in *Neural Information Processing Systems Workshop*, (Barcelona, Spain), Dec 2016.

[33] P. Doetsch, M. Hannemann, R. Schlüter, and H. Ney, "Inverted alignments for end-to-end automatic speech recognition," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, pp. 1265–1273, Dec 2017.

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

[35] A. Zeyer, T. Alkhouli, and H. Ney, "RETURNN as a generic flexible neural toolkit with application to translation and speech recognition," *arXiv preprint arXiv:1805.05225, published on ACL 2018*, 2018.

[36] R. Sennrich, B. Haddow, and A. Birch, "Neural machine translation of rare words with subword units," in *ACL*, (Berlin, Germany), pp. 1715–1725, August 2016.

[37] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[38] P. Bahar, J. Rosendahl, N. Rossenbach, and H. Ney, "The RWTH Aachen machine translation systems for IWSLT 2017," in *Int. Workshop on Spoken Language Translation*, (Tokyo, Japan), pp. 29–34, Dec. 2017.

[39] Z. Tu, Z. Lu, Y. Liu, X. Liu, and H. Li, "Modeling coverage for neural machine translation," in *ACL*, 2016.

[40] R. Schlüter, I. Bezrukov, H. Wagner, and H. Ney, "Gammatone features and feature combination for large vocabulary speech recognition," in *ICASSP*, (Honolulu, HI, USA), pp. 649–652, Apr. 2007.

[41] S. Wiesler, A. Richard, P. Golik, R. Schlüter, and H. Ney, "RASR/NN: The RWTH neural network toolkit for speech recognition," in *ICASSP*, (Florence, Italy), pp. 3313–3317, May 2014.

[42] D. Povey, V. Peddinti, D. Galvez, P. Ghahremani, V. Manohar, X. Na, Y. Wang, and S. Khudanpur, "Purely sequence-trained neural networks for ASR based on lattice-free MMI," in *Interspeech*, pp. 2751–2755, 2016.

[43] A. Hannun, C. Case, J. Casper, B. Catanzaro, G. Diamos, E. Elsen, R. Prenger, S. Satheesh, S. Sengupta, A. Coates, *et al.*, "DeepSpeech: Scaling up end-to-end speech recognition," *arXiv preprint arXiv:1412.5567*, 2014.

[44] A. L. Maas, Z. Xie, D. Jurafsky, and A. Y. Ng, "Lexicon-free conversational speech recognition with neural networks," in *Proc. NAACL*, 2015.

[45] G. Zweig, C. Yu, J. Droppo, and A. Stolcke, "Advances in all-neural speech recognition," in *ICASSP*, pp. 4805–4809, IEEE, 2017.

[46] L. Lu, X. Zhang, and S. Renais, "On training the recurrent neural network encoder-decoder for large vocabulary end-to-end speech recognition," in *ICASSP*, pp. 5060–5064, IEEE, 2016.