# End-to-End Speech Recognition with Local Monotonic Attention

**Andros Tjandra, Sakriani Sakti, Satoshi Nakamura**
Graduate School of Information Science
Nara Institute of Science and Technology
{andros.tjandra.ai6, ssakti, s-nakamura}@is.naist.jp

## Abstract

Most attention mechanism in sequence-to-sequence model is based on a global attention property which requires a computation of a weighted summarization of the whole input sequence generated by encoder states. However, it is computationally expensive and often produces misalignment on the longer input sequence. Furthermore, it does not fit with monotonous or left-to-right nature in speech recognition task. In this paper, we propose a novel attention mechanism that has local and monotonic properties. Various ways to control those properties are also explored. Experimental results demonstrate that encoder-decoder based ASR with local monotonic attention could achieve significant performance improvements and reduce the computational complexity in comparison with the one that used the standard global attention architecture.

## 1 Introduction

Recently, end-to-end ASR based model approach allows the model to directly learn the mapping between variable-length speech to text [Chorowski et al., 2014, Chan et al., 2016]. It allow us to replaces the conventional ASR component such as a acoustic model, a pronunciation lexicon, and a language model, into a single integrated neural network. There are three main modules for end-to-end training for sequence-to-sequence task: (1) Encoder module which represent source information, (2) Decoder module which produces output sequence, (3) Attention module which help decoder to extract related information from encoder representation.

However, mostly these attention module used today has a “global" property [Bahdanau et al., 2014, Luong et al., 2015]. Every time the decoder needs to predict the output given the previous output, it must compute a weighted summarization of the whole input sequence generated by the encoder states. However, although the global attention mechanism has often improved performance in some tasks, it is very computationally expensive. Furthermore, global attention does not fit with monotonous or left-to-right natures in speech recognition tasks and focus on the audio’s specific timing. Therefore, the attention needs two important characteristics to address these problems: local and monotonicity properties. The local property helps our attention module focus on certain parts that the decoder wants to transcribe, and the monotonicity property strictly generates alignment left-to-right from beginning to the end of speech.

In this paper, we propose a novel attention module that has local and monotonicity properties[1]. We explore various ways to control these properties on the attention-based encoder-decoder model. Experimental results demonstrate that an encoder-decoder based ASR with local monotonic attention significantly improved performance and reduced the computational complexity more than one that used the standard global attention architecture.

[1] The long version of this paper with title "Local Monotonic Attention Mechanism for End-to-End Speech and Language Processing" has been accepted at 8th International Joint Conference on Natural Language Processing (IJCNLP) 2017

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

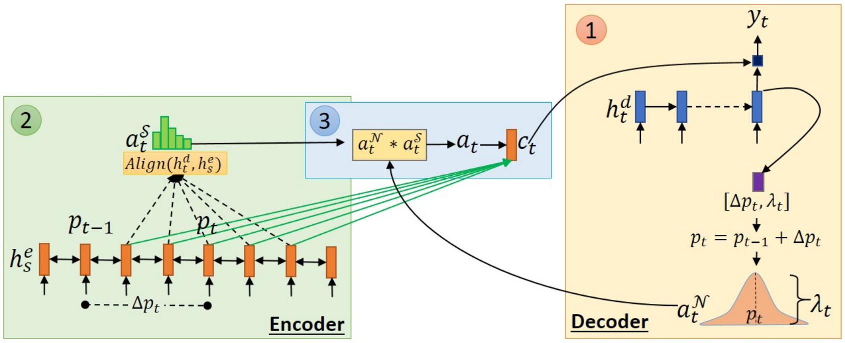

Figure 1: Local monotonic attention.

## 2 Attention-based Encoder Decoder Neural Network

The encoder-decoder model is a neural network that directly models conditional probability $p(\mathbf{y}|\mathbf{x})$, where $\mathbf{x} = [x_1, ..., x_S]$ is the source sequence with length $S$ and $\mathbf{y} = [y_1, ..., y_T]$ is the target sequence with length $T$. The encoder task processes input sequence $\mathbf{x}$ and outputs representative information $\mathbf{h^e} = [h_1^e, ..., h_S^e]$ for the decoder. The attention module produces context information $c_t$ at the time $t$ based on the encoder and decoder hidden states:

$$c_t = \sum_{s=1}^{S} a_t(s) * h_s^e \qquad \text{where} \qquad a_t(s) = \text{Align}(h_s^e, h_t^d) = \frac{\exp(\text{Score}(h_s^e, h_t^d))}{\sum_{s=1}^{S} \exp(\text{Score}(h_s^e, h_t^d))} \tag{1}$$

There are several variations for score functions:

$$\text{Score}(h_s^e, h_t^d) = \begin{cases} \langle h_s^e, h_t^d \rangle, & \text{dot product} \\ h_s^{e\intercal} W_s h_t^d, & \text{bilinear} \\ V_s^\intercal \tanh(W_s[h_s^e, h_t^d]), & \text{MLP} \end{cases} \tag{2}$$

where Score : $(\mathbb{R}^M \times \mathbb{R}^N) \to \mathbb{R}$, $M$ is the number of hidden units for encoder and $N$ is the number of hidden units for decoder. Finally, the decoder task, which predicts the target sequence probability at time $t$ based on previous output and context information $c_t$ can be formulated $\log p(\mathbf{y}|\mathbf{x}) = \sum_{t=1}^{T} \log p(y_t|y_{<t}, c_t)$.

## 3 Locality and Monotonicity Properties

Figure 1 illustrates the overall mechanism of our proposed local monotonic attention, and details are described blow.

1. **Monotonicity-based Prediction of Central Position**
   First, we define how to predict the next central position of the alignment illustrated in Part (1) of Figure 1. Assume we have source sequence with length $S$, represented by $S$ encoded states $\mathbf{h^e} = [h_1^e, ..., h_S^e]$. At time $t$, we want to decode the $t$-th target output given the source sequence, $y_{t-1}$, and current decoder states $h_t^d$. In standard approaches, we use hidden states $h_t^d$ to predict the position difference $\Delta p_t$ with a multilayer perceptron (MLP). We use variable $\Delta p_t$ to determine how far we should move the center of the alignment compared to previous center $p_{t-1}$.
   In this paper, we propose two different formulations for estimating $\Delta p_t$:
   - **Constrained position prediction:**
     We limit maximum range from $\Delta p_t$ with hyperparameter $C_{max}$ with the following equation:
     $$\Delta p_t = C_{max} * \text{sigmoid}(V_p^\intercal \tanh(W_p h_t^d)) \tag{3}$$
     However, it requires us to handle hyperparameter $C_{max}$.

   - **Unconstrained position prediction:**
     Compared to a previous formulation, since we do not limit the maximum range of $\Delta p_t$, here we can ignore hyperparameter $C_{max}$ and use exponential (exp) function instead of sigmoid. We formulate unconstrained position prediction with following equation:
     $$\Delta p_t = \exp(V_p^\intercal \tanh(W_p h_t^d)) \tag{4}$$

Both equations guarantee monotonicity properties since $\forall t \in [1..T], p_t \geq (p_{t-1} + \Delta p_t)$. Additionally, we also used scaling variable $\lambda_t$ to scale the unnormalized Gaussian distribution that depends on $h_t$. The main objective of this step is to generate a scaled Gaussian distribution $a_t^{\mathcal{N}}$ in Eq.5.

$$a_t^{\mathcal{N}}(s) = \lambda_t * \exp\left(-\frac{(s - p_t)^2}{2\sigma^2}\right). \tag{5}$$

where $p_t$ is the mean and $\sigma$ is the standard deviation. In this paper, we treat $\sigma$ as a hyperparameter.

2. **Locality-based Alignment Generation**
After calculating new position $p_t$, we generate locality-based alignment, as shown in Part (2) of Figure 1.Here, we just need calculate the scores (Eq. 2) and generate alignment $a_t^{\mathcal{S}}$ only within $[p_t - 2\sigma, p_t + 2\sigma]$:

$$a_t^{\mathcal{S}}(s) = \text{Align}(h_s^e, h_t^d), \quad \forall s \in [p_t - 2\sigma, p_t + 2\sigma]. \tag{6}$$

Compared to the standard global attention, we can reduce the decoding computational complexity $O(T * S)$ into $O(T * \sigma)$ where $\sigma \ll S$ and $\sigma$ is constant, $T$ is total decoding step, $S$ is the length of the encoder states.

3. **Context Calculation**
In the last step, we calculate context $c_t$ with alignments $a_t^{\mathcal{N}}$ and $a_t^{\mathcal{S}}$, as shown in Part (3) of Figure 1:

$$c_t = \sum_{s=(p_t-2\sigma)}^{(p_t+2\sigma)} \left(a_t^{\mathcal{N}}(s) * a_t^{\mathcal{S}}(s)\right) * h_s^e \tag{7}$$

Context $c_t$ and current hidden state $h_t^d$ will later be utilized

Overall, we can rephrase the first step as generating “prior" probabilities $a_t^{\mathcal{N}}$ based on the previous $p_{t-1}$ position and the current decoder states. Then the second step task generates “likelihood" probabilities $a_t^{\mathcal{S}}$ by measuring the relevance of our encoder states with the current decoder states. In the third step, we combine our “prior" and “likelihood" probability into an unnormalized “posterior" probability $a_t$ and calculate expected context $c_t$. for calculating current output $y_t$. For more detailed explanation, please refer to the long version of our paper Tjandra et al. [2017][2].

## 4 Experiment and Result Discussion

We conducted our experiments on the TIMIT [Garofolo et al., 1993] dataset with the same set-up for training, development, and test sets as defined in the Kaldi s5 recipe [Povey et al., 2011]. We used 40-dimensional fbank as the speech features and 39 phoneme class for our target. On the encoder sides, we used a linear layer followed by three Bi-LSTM with 512 hidden units. We reduced the source sequence length by a factor of 4 with subsampling. On the decoder sides, we used two LSTMs with 512 hidden units. Hyperparameter $\sigma$ was set to 1.5, and $C_{max}$ for constrained position prediction (see Eq. 3) was set to 5. Both hyperparameters were empirically selected and generally gave consistent results across various settings in our proposed model. In the recognition phase, we generated transcriptions with best-1 (greedy) search from the decoder. We did not use any language model in this work to rescore our transcription result.

[2] `https://arxiv.org/abs/1705.08091`

Table 1: Result from baseline and proposed models on TIMIT.

| Model | | | Test PER (%) |
|---|---|---|---|
| **Global Attention Model (Baseline)** | | | |
| Att Enc-Dec [Pereyra et al., 2017] | | | 23.2 |
| Att Enc-Dec [Luo et al., 2016] | | | 24.5 |
| Att Enc-Dec with MLP Scorer (Our baseline) | | | 23.8 |
| **Local Attention Model (Proposed)** | | | |
| **Monotonicity** | **Locality** | | **Test** |
| **Pos Prediction** $\Delta p_t$ | **Alignment Score**$(h^e_s, h^d_t)$ | **Function Type** | **PER (%)** |
| Const (*sigmoid*) | No | - | 23.2 |
| Const (*sigmoid*) | Yes | Bilinear | 21.9 |
| Const (*sigmoid*) | Yes | MLP | 21.7 |
| Unconst (*exp*) | No | - | 23.1 |
| Unconst (*exp*) | Yes | Bilinear | 20.9 |
| Unconst (*exp*) | Yes | MLP | 21.4 |

In Table 1, we summarizes the experiments on our proposed local attention models and compares them to the baseline model using several possible scenarios. We found that it is more beneficial to use the unconstrained position prediction formulation since it gives better performance and we do not need to handle the additional hyperparameter $C_{max}$. We also found out that the scorer function is essential for our proposed models. Overall, our proposed encoder-decoder model significantly improved the performance and reduced the computational complexity in comparison with one that used standard global attention mechanism. The best performance achieved by our proposed model with unconstrained position prediction and bilinear scorer, and provided 12.2% relative error rate reduction to our baseline.

## 5 Related Works

Aharoni and Goldberg [2016] proposed a monotonic attention by introducing an additional step symbol to control the latest attention position. However, they need the complete target alignment to train the model. Compared to our approach, we do not require any explicit alignment. Raffel et al. [2017] also proposed a method for producing a monotonic alignment by using Bernoulli random variable to control when the attention mechanism should stop and generate output. Compare to our proposed method, we did not restrict the area of the attention window. In the other hand, it is also possible to combine their method with ours to predict the central position $p_t$ location.

## 6 Conclusion

This paper demonstrated a novel attention mechanism for encoder decode model that ensures monotonicity and locality properties. We explored various ways to control these properties, including monotonicity-based position prediction and locality-based alignment generation. We test our proposed model with speech recognition task. Our result revealed that we significantly improved the performance and reduced the computational complexity more than one that used standard global attention architecture.

## Acknowledgements

Part of this work was supported by JSPS KAKENHI Grant Numbers JP17H06101 and JP17K00237.

## References

R. Aharoni and Y. Goldberg. Sequence to sequence transduction with hard monotonic attention. *arXiv preprint arXiv:1611.01487*, 2016.

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

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*, pages 4960–4964. IEEE, 2016.

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.

J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett. Darpa TIMIT acoustic-phonetic continous speech corpus cd-rom. NIST speech disc 1-1.1. *NASA STI/Recon technical report n*, 93, 1993.

Y. Luo, C.-C. Chiu, N. Jaitly, and I. Sutskever. Learning online alignments with continuous rewards policy gradient. *arXiv preprint arXiv:1608.01281*, 2016.

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

G. Pereyra, G. Tucker, J. Chorowski, Ł. Kaiser, and G. Hinton. Regularizing neural networks by penalizing confident output distributions. *arXiv preprint arXiv:1701.06548*, 2017.

D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely. The Kaldi speech recognition toolkit. In *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society, Dec. 2011. IEEE Catalog No.: CFP11SRW-USB.

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.

A. Tjandra, S. Sakti, and S. Nakamura. Local monotonic attention mechanism for end-to-end speech recognition. *arXiv preprint arXiv:1705.08091*, 2017.