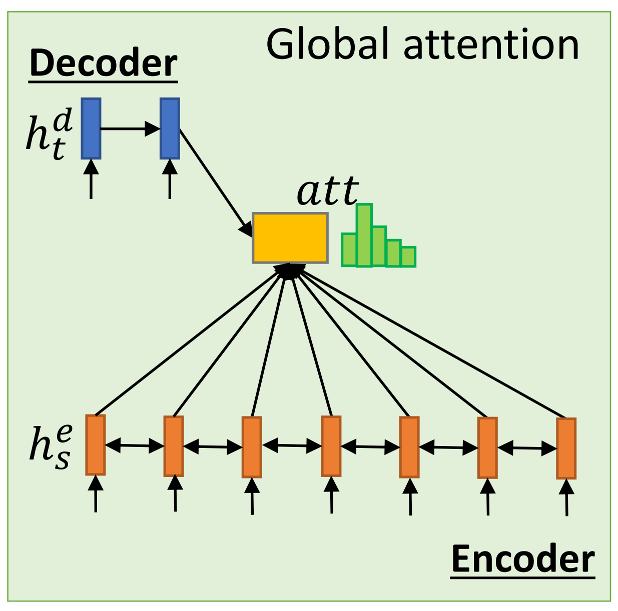
End-to-End Speech Recognition with Local Monotonic Attention

Andros Tjandra, Sakriani Sakti, Satoshi Nakamura

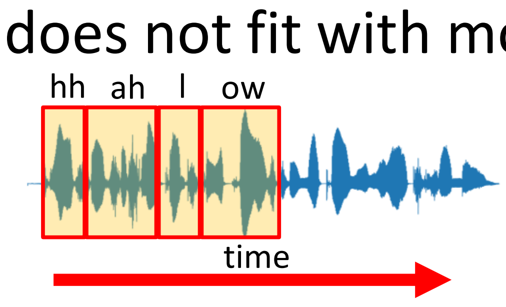
Nara Institute of Science and Technology (**NAIST**), Japan

Motivation

- Most attentional mechanism in encoder-decoder neural network has a "global" property
  - → Attend whole input sequence and calculate the expected context



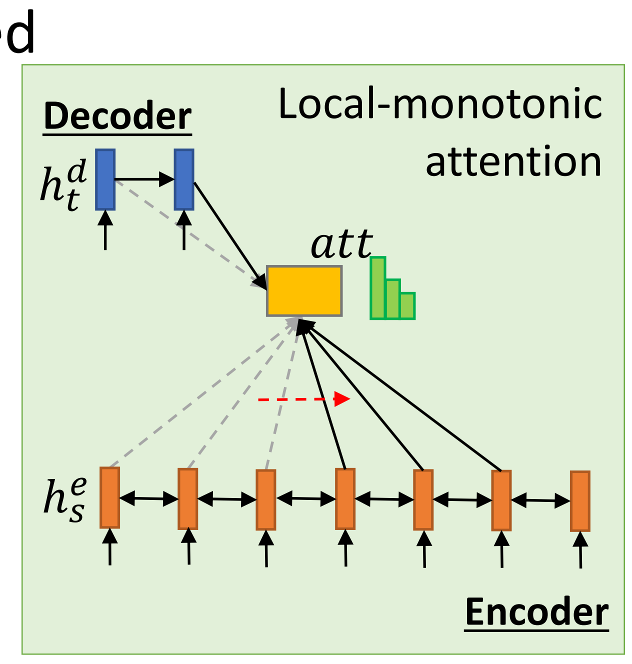
- However, it does not fit with monotonous nature in ASR



We need:

Attention Mechanism with Local & Monotonicity Properties

- **Monotonicity**: current expected position ≥ than previous expected position
- **Locality**: attention score outside certain distance will be ignored



Designing Local & Monotonic Attention

1. Monotonicity-based Prediction of Central Position

- Predict $\Delta p_t, \lambda_t$ with an MLP(h_t^d)
- New position $p_t = p_{t-1} + \Delta p_t$
- Generate Gaussian attention $a_t^{\mathcal{N}}(s) = \lambda_t * \exp\left(-\frac{(s-p_t)^2}{2\sigma^2}\right)$

Two different formulations for Δp_t:

a. Constrained

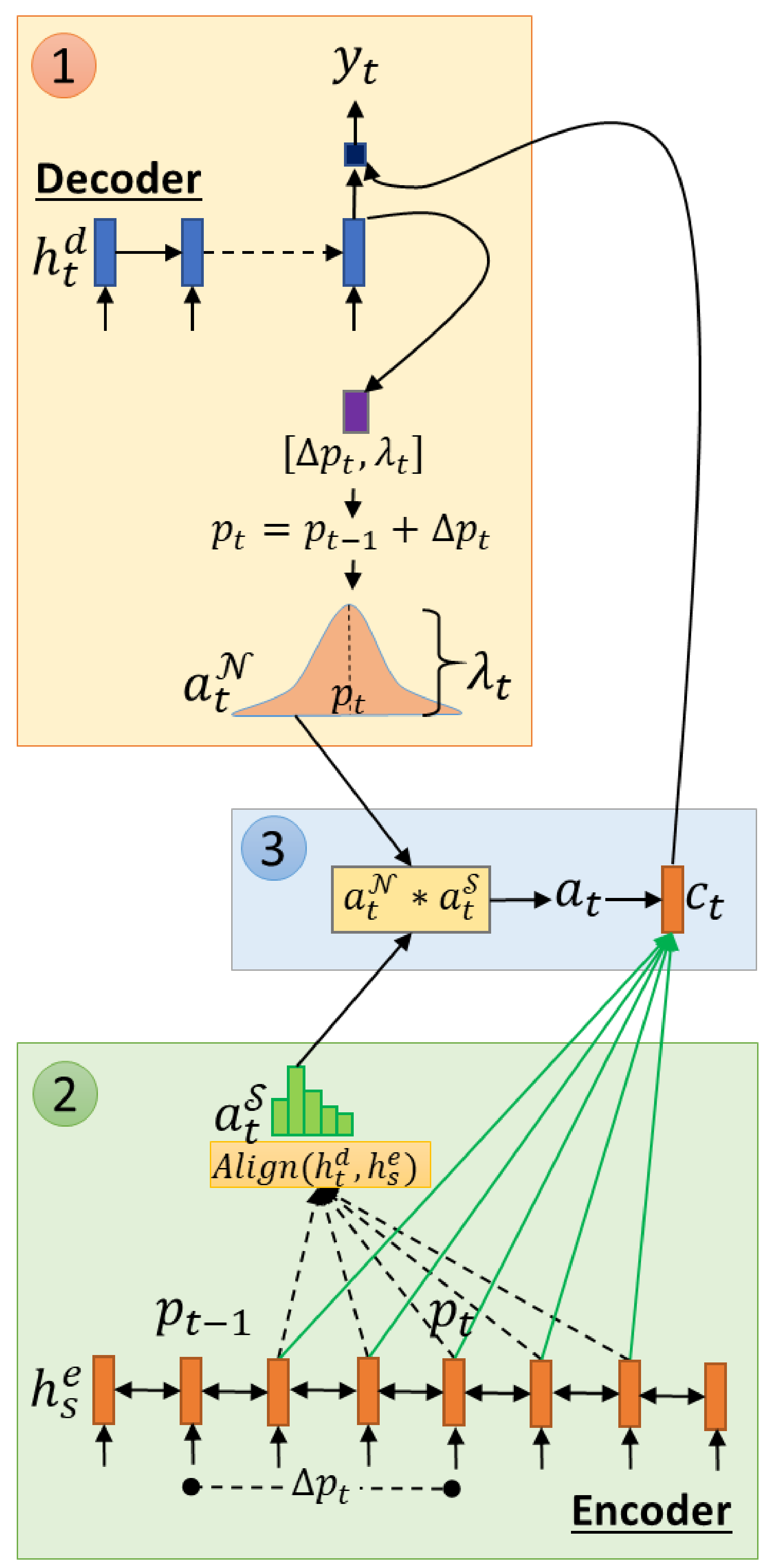
$$\Delta p_t = C_{max} * f\left(MLP\left(h_t^d\right)\right)$$

- $f: \mathbb{R} \rightarrow [0,1]$ (e.g., sigmoid)
- C_{max} is hyperparameter (max. jump range)

b. Unconstrained

$$\Delta p_t = f\left(MLP\left(h_t^d\right)\right)$$

- $f: \mathbb{R} \rightarrow \mathbb{R}_0^+$ (e.g., exp, softplus)



2. Locality-based Alignment Generation

- Select a subset of encoder states h_e^s where $\forall s \in [p_t - 2\sigma, p_t + 2\sigma]$
- Calculate the scorer attention $a_t^{\mathcal{S}} = Align(h_s^e, h_t^d)$ only based on a subset of h_s^e

$$Align(h_s^e, h_t^d) = \begin{cases} h_s^{e\intercal} W h_t^d & \text{bilinear} \\ W_1 \sigma(W_2[h_s^e, h_t^d]) & \text{MLP} \end{cases}$$

Advantage : reduce the complexity from $O(T * S) \rightarrow O(T * \sigma)$

3. Context Calculation

Combine attention $a_t^{\mathcal{N}}, a_t^{\mathcal{S}}$ and calculate weighted sum c_t

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

Experimental Results

Setup:

- **Dataset:** TIMIT
- **Features:** 40-dims log-fbank $+\Delta + \Delta^2$ (total 120dims)
- **Output:** 39 phns + eos symbol
- **Metric:** Phoneme Error Rate (PER)
- **Model:** Encoder: 3 BiLSTM (hidden 512) Decoder: 2 LSTM (hidden 512)

Model			Test PER(%)
Baseline			
Att Enc-Dec Global MLP Scorer			23.8
Att Enc-Dec *local-m* (Loung et al., 2015)			(not converged)
Proposed			
Monotonicity	**Locality ($Align$)**		
Pos Prediction Δp_t	**Scorer**	**Function**	
Constrained (*sigmoid*)	No	-	23.2
	Yes	Bilinear	21.9
		MLP	21.7
Unconstrained (*exp*)	No	-	23.1
	Yes	Bilinear	**20.9**
		MLP	21.4

Conclusion

- Demonstrated a novel attention mechanism that ensures **monotonicity** & **locality** properties
- Explained various ways to control those properties
- Experimental results showed that **unconstrained position prediction + local alignment** produced best result
- Can be applied to other tasks: G2P & MT on certain language pairs

Link to full paper: https://goo.gl/GGqYiT

Contact: andros.tjandra.ai6@is.naist.jp