Streaming Automatic Speech Recognition with the Transformer Model

Niko Moritz, Takaaki Hori, Jonathan Le Roux

ICASSP
May, 2020

MITSUBISHI ELECTRIC RESEARCH LABORATORIES (MERL)
Cambridge, Massachusetts, USA
http://www.merl.com

Motivation

- End-to-end automatic speech recognition (ASR) has greatly simplified the pipeline for buidling and applying ASR systems.
- Offline end-to-end ASR systems have shown to surpass the performance of traditional hybrid DNN-HMM solutions.
- Streaming end-to-end architectures are still lacking behind this success.
- Encoder-decoder based architectures have demonstrated to achieve the best end-to-end ASR results but are difficult to apply in a streaming fashion.

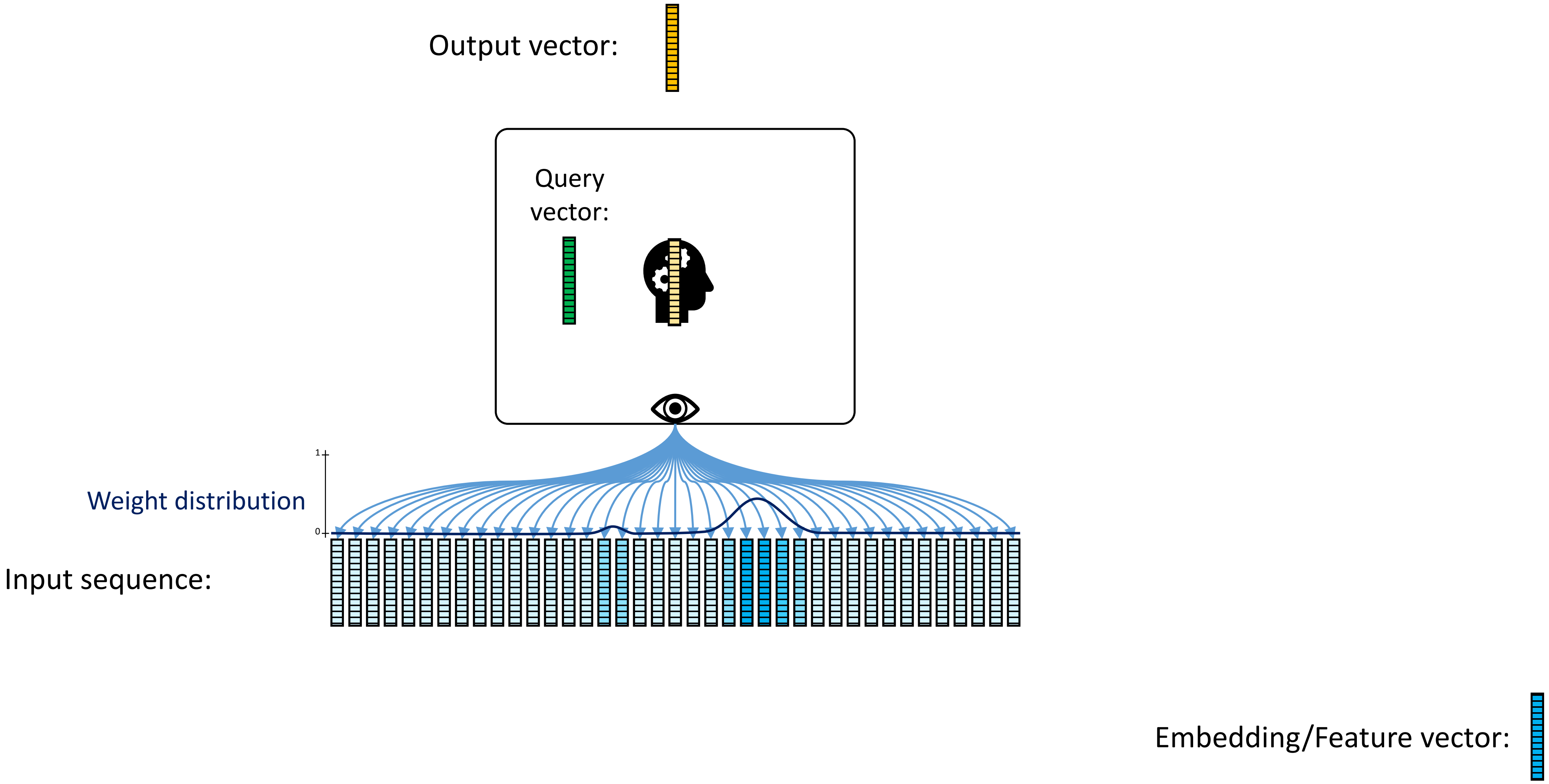
This work

- Our proposed triggered attention (TA) concept is used to overcome these difficulties.
- The TA concept is applied to the transformer architecture, achieving SOTA streaming end-to-end ASR results.

Outline

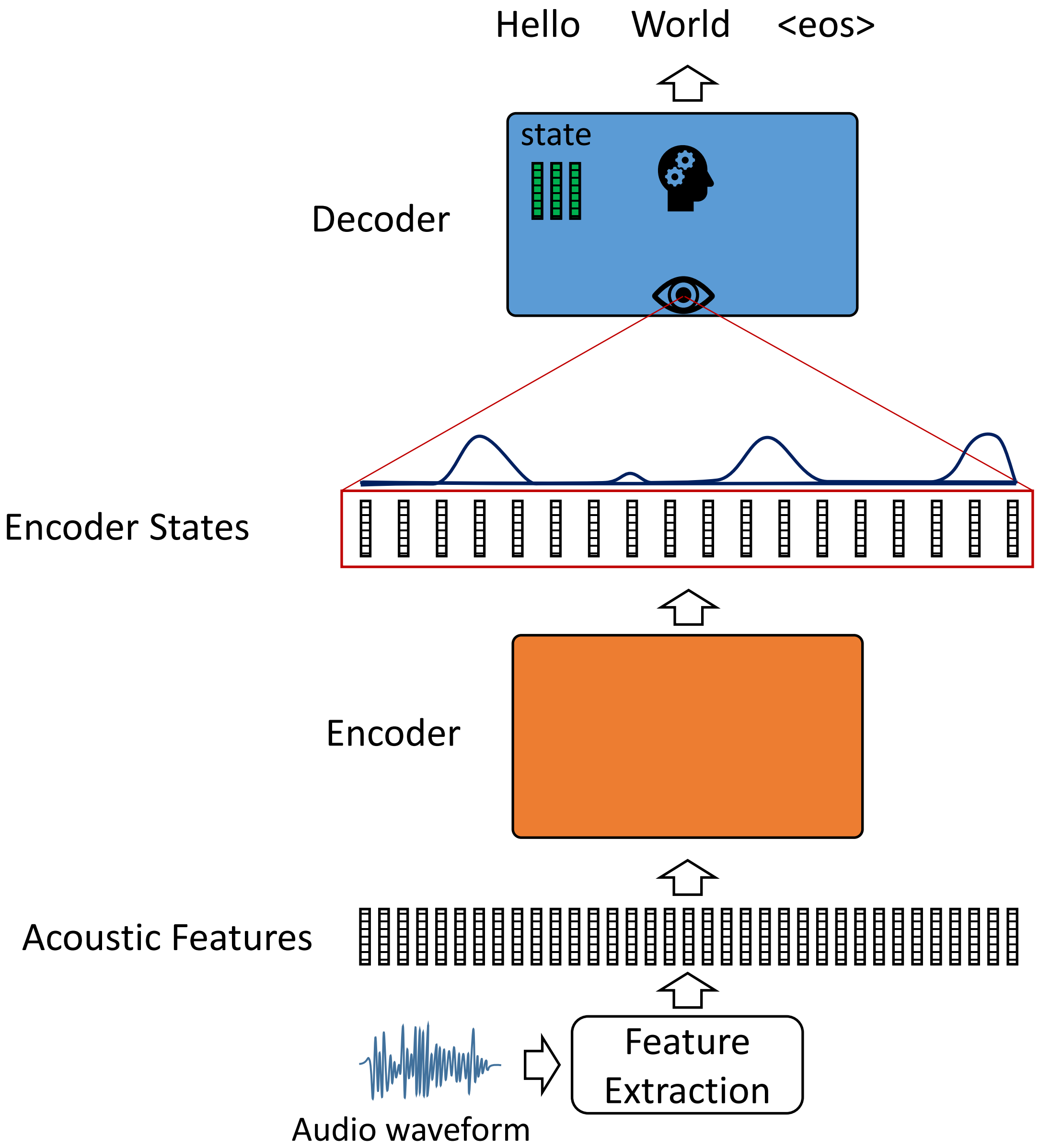
- Encoder-Decoder Neural Networks
  - Attention
  - Transformer
  - Self-attention
  - Time-Restricted Self-Attention
  - Streaming Encoder-Decoder Attention (prior work)
- Triggered Attention
  - Architecture
  - Frame-Synchronous Decoding Algorithm
- LibriSpeech Results

Attention

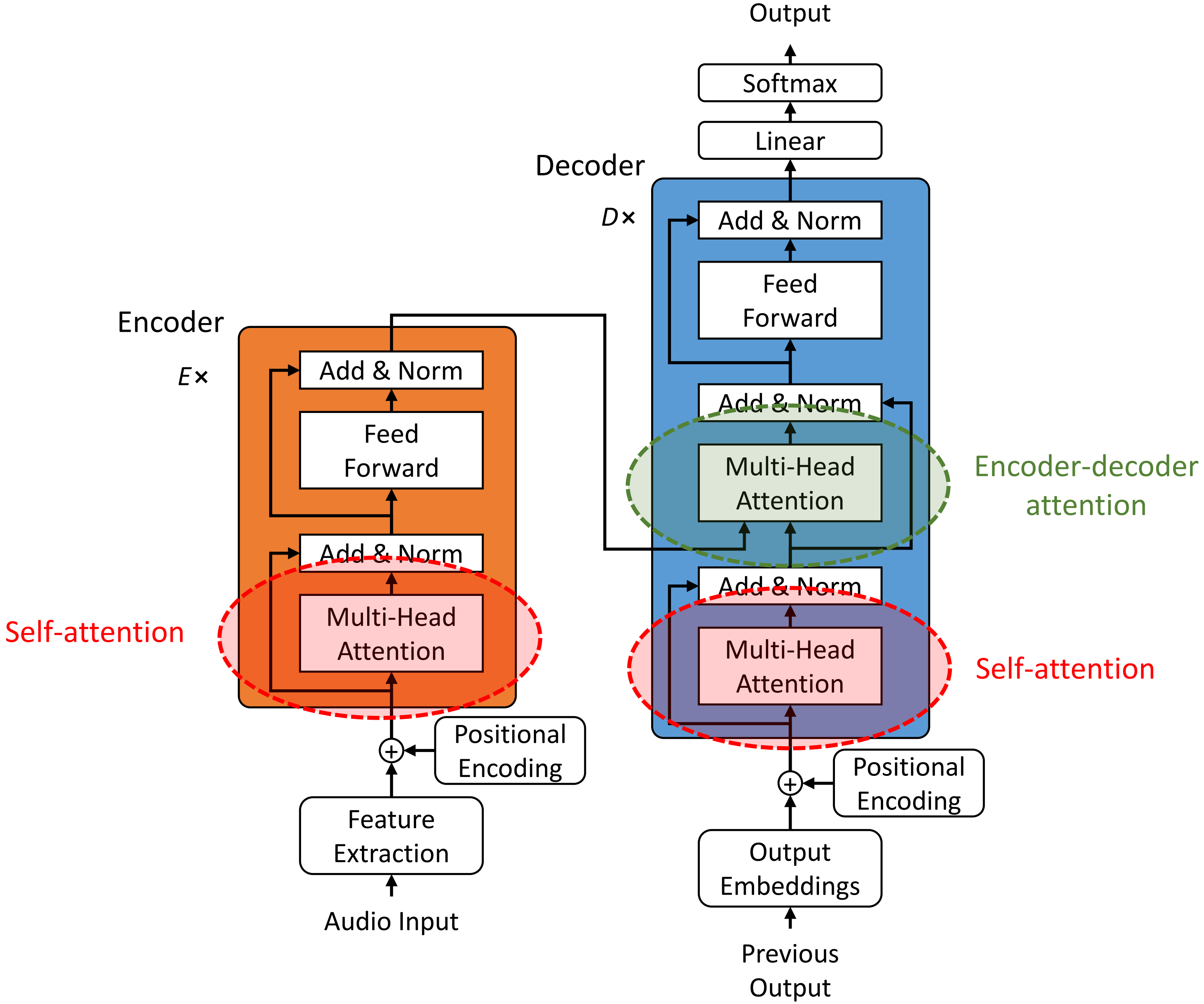


Output vector:
Query
vector:
1
0
Weight distribution
Input sequence:
Embedding/Feature vector:

Encoder-Decoder Attention

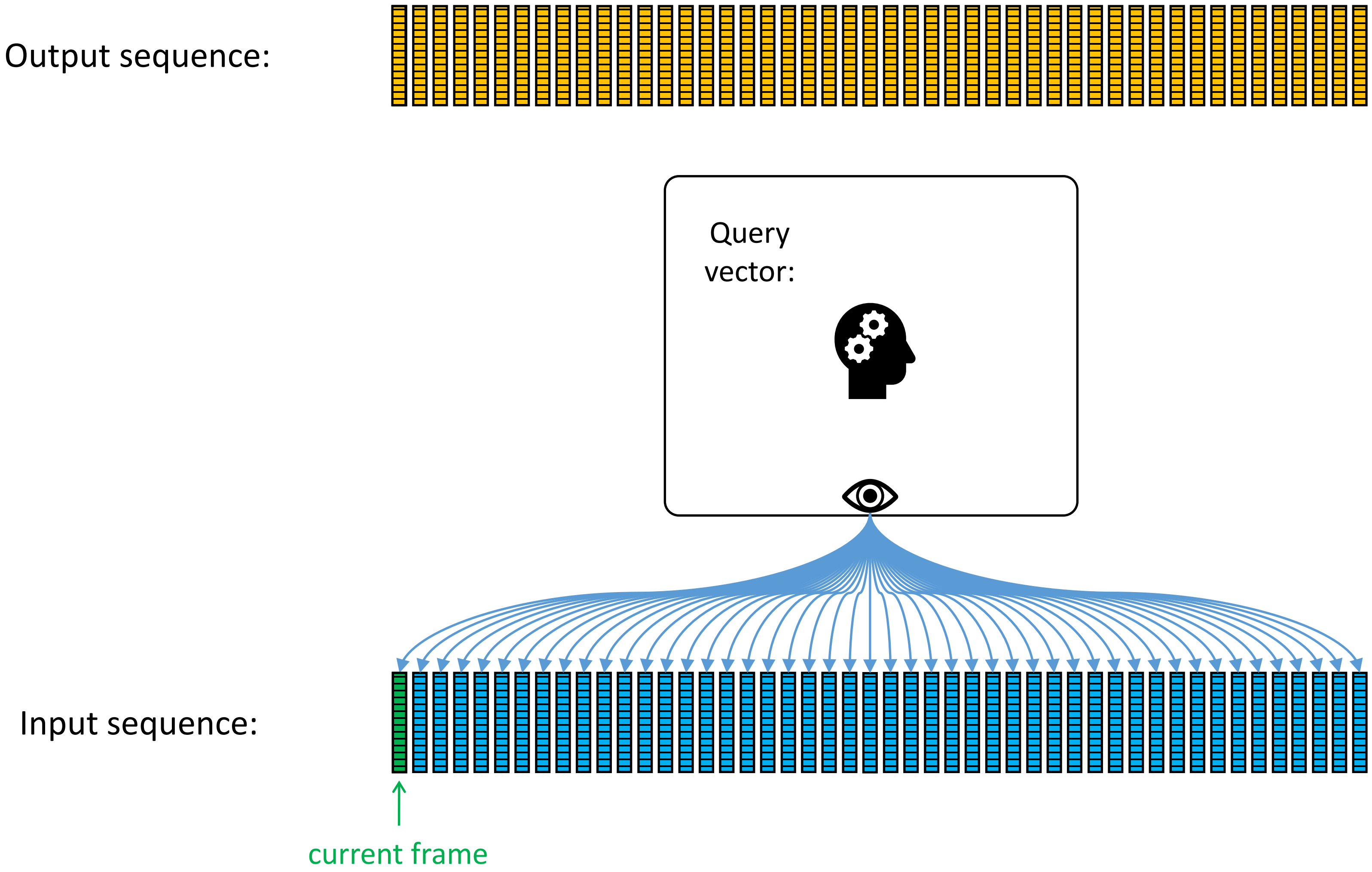


Transformer Architecture



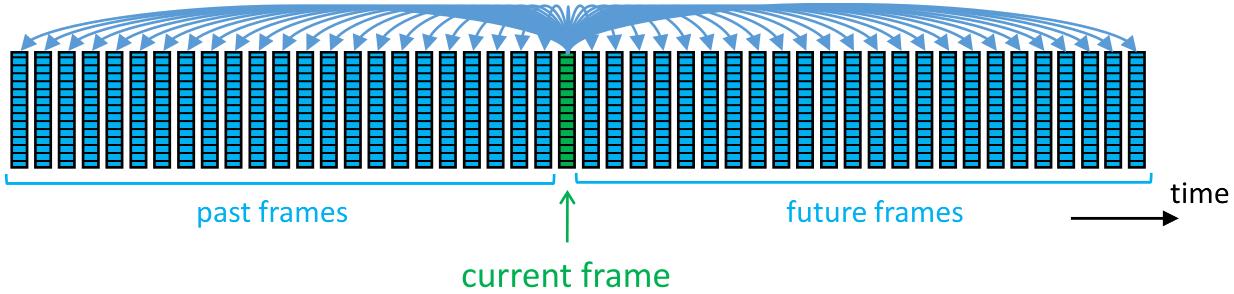
E=12
D=6

Self-Attention



Self-Attention

Input sequence:



Time-Restricted Self-Attention

Output sequence:

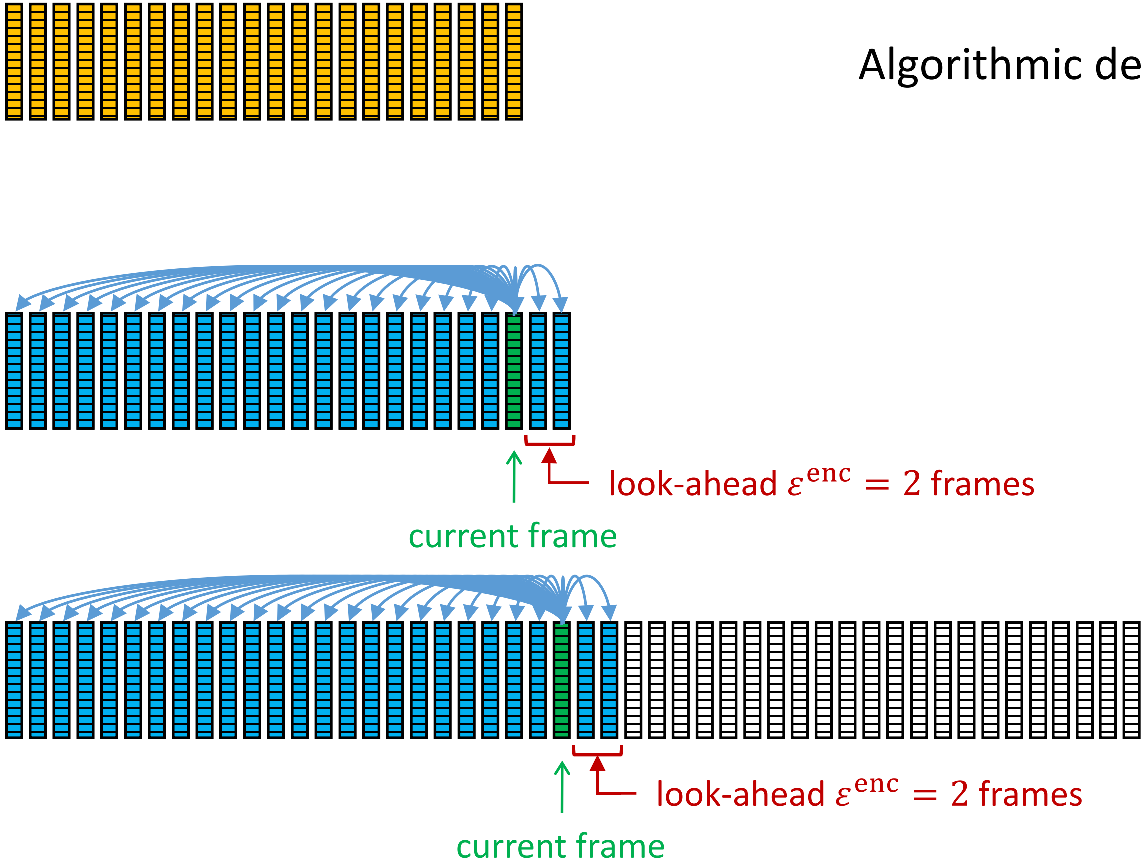
Algorithmic delay: $\#layers \cdot \varepsilon^{\mathrm{enc}} = 4$ frames

Output sequence:

Layer 2

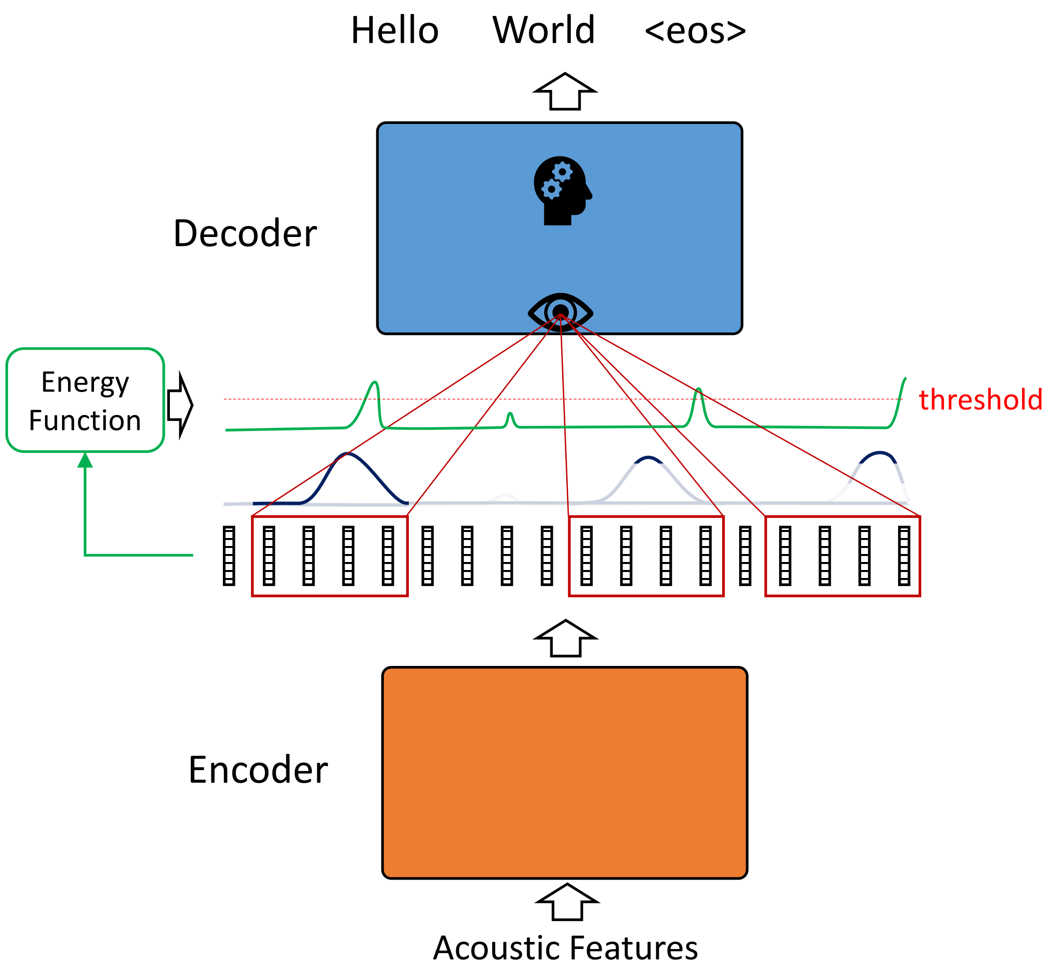
Input sequence:

Layer 1



Streaming Encoder-Decoder Attention (prior work)

Adaptive Chunking based on Selection Probability

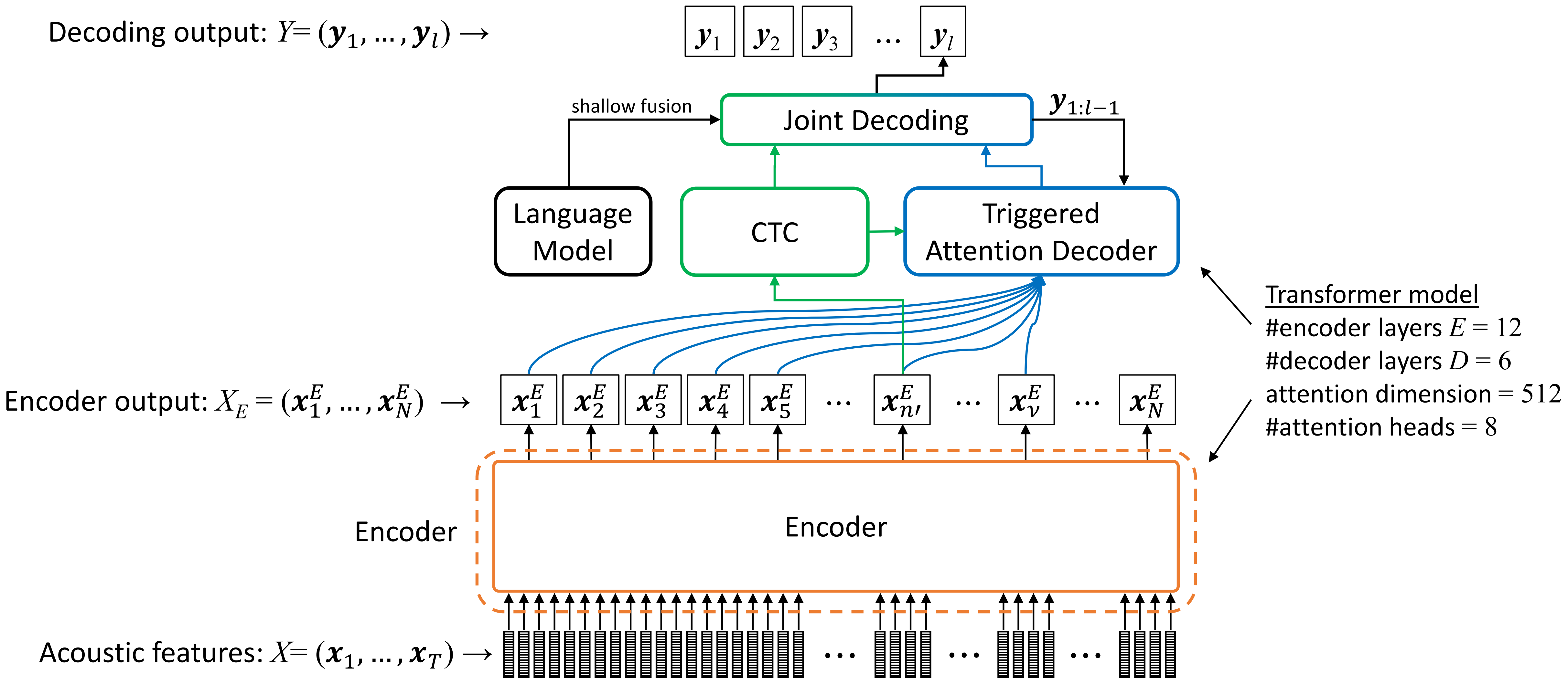


Example:
- Monotonic Chunkwise Attention (MoChA) [1]

Problems:
- Backpropagation with discrete decisions is not possible.
- No frame-synchronous decoding algorithm.
- Detecting word or word-piece positions is a good part of the ASR job that defines insertion and deletion errors.

[1] C. Chiu and C. Raffel, “Monotonic chunkwise attention,” in Proc. ICLR, Apr. 2018.

Triggered Attention (TA) Architecture



N. Moritz, T. Hori, and J. Le Roux, "Triggered attention for end-to-end speech recognition," in Proc. ICASSP, May 2019, pp. 5666–5670.

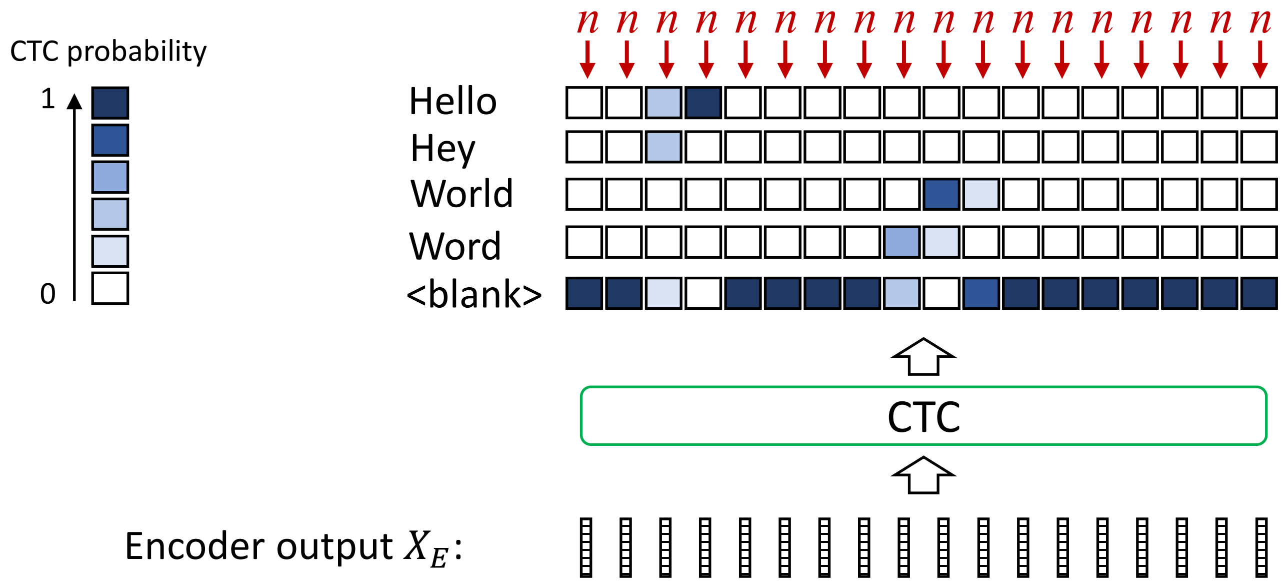
Frame-Synchronous Decoding

Frame-synchronous CTC prefix beam search [1]:

ℓ_1 = (<sos>, Hello) [illegible] World)

ℓ_7 = (<sos>, Hey, Word, World) ℓ_8 = (<sos>, Hello, World)

Set of prefix sequences after pruning: $\Omega = \{\ell_1, \ell_2, \ell_8\}$



$$\log p(\ell | X_{1:n}^E) = \log p_{\mathrm{prfx}}(\ell | X_{1:n}^E) + \alpha \log p_{\mathrm{LM}}(\ell) + \beta |\ell|$$

ℓ: prefix sequence
$X_{1:n}^E$: Encoder state sequence for frame $(1, \dots, n)$
p_{prfx}: CTC prefix probability
p_{LM}: Language model (LM) probability
α: LM weight
β: insertion bonus weight
$|\ell|$: prefix sequence length

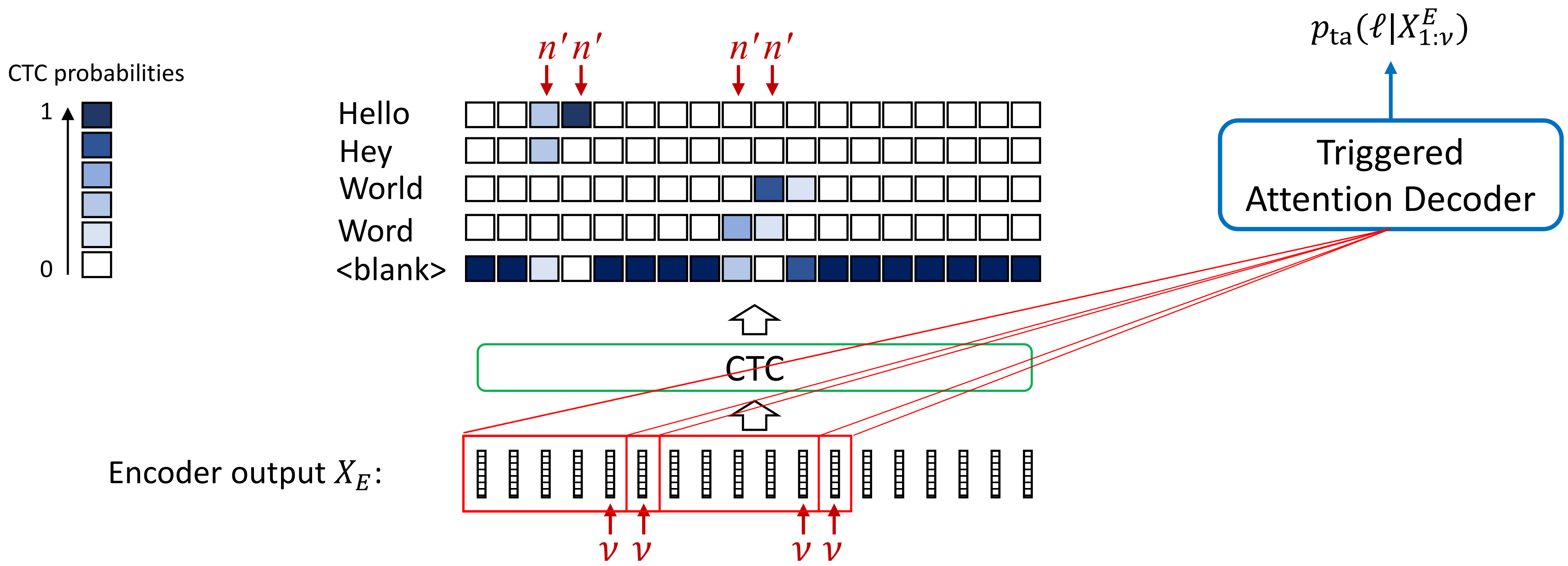
[1] A. L. Maas, A. Y. Hannun, D. Jurafsky, and A. Y. Ng, “Firstpass large vocabulary continuous speech recognition using bidirectional recurrent DNNs,” arXiv preprint arXiv:1408.2873, 2014.

Frame-Synchronous Decoding

Frame-synchronous one-pass TA decoding [1]:

$$\log p_{\text{joint}}(\ell|X^E_{1:n}) = \lambda \log p_{\text{prfx}}(\ell|X^E_{1:n}) + (1-\lambda)\log p_{\text{ta}}(\ell|X^E_{1:\nu}) + \alpha \log p_{\text{LM}}(\ell) + \beta|\ell|$$

p_{ta}: Triggered attention probability
λ: CTC weight
$\nu = \text{n}' + \varepsilon^{\text{dec}}$
n': trigger frame
ε^{dec}: decoder look-ahead



[1] N. Moritz, T. Hori, and J. Le Roux, "Streaming end-to-end speech recognition with joint CTC-attention based models," in Proc. ASRU, Dec. 2019, pp. 936–943.

LibriSpeech Word Error Rates (WERs) [%]

Encoder	Full-sequence CTC-attention decoding [1,2]			
	Clean		Other	
	Dev	Test	Dev	Test
Full-sequence	2.4	2.7	6.0	6.1

Time-restricted encoder	Frame-synchronous CTC prefix beam search				TA: $\varepsilon^{dec} = 18$, delay: $\varepsilon^{dec} \cdot 40\text{ ms} = 720\text{ ms}$			
	Clean		Other		Clean		Other	
$\varepsilon^{\mathrm{enc}}$ / delay*	Dev	Test	Dev	Test	Dev	Test	Dev	Test
0 / 30 ms	3.3	3.7	9.4	9.4	2.9	3.2	8.1	8.0
1 / 510 ms	3.0	3.3	8.4	8.6	2.8	3.0	7.5	7.8
2 / 990 ms	2.9	3.1	8.0	8.2	2.7	2.9	7.3	7.4
3 / 1470 ms	2.8	2.9	7.8	8.1	2.7	2.8	7.1	7.2
Full-sequence	2.5	2.8	6.9	7.0	2.4	2.6	6.1	6.3

1.23 seconds

* Algorithmic encoder delay: $E \cdot \varepsilon^{\mathrm{enc}} \cdot \text{frame}-\text{rate} + \text{CNN}-\text{delay}$

$E = 12$, frame-rate = 40 ms, CNN-delay = 30 ms

[1] S. Watanabe, T. Hori, S. Kim, J. R. Hershey, and T. Hayashi, "Hybrid CTC/attention architecture for end-to-end speech recognition," J. Sel. Topics Signal Processing, vol. 11, no. 8, pp. 1240–1253, 2017.
[2] S. Karita, N. Yalta, S. Watanabe, M. Delcroix, A. Ogawa, and T. Nakatani, "Improving transformer-based end-to-end speech recognition with connectionist temporal classification and language model integration," in Proc. ISCA Interspeech, Sep. 2019, pp. 1408–1412.

Conclusions

- The triggered attention (TA) concept enables frame-synchronous decoding with an encoder-decoder based model for the first time.
- The TA concept enables joint scoring of an CTC and attention-based decoder model in a streaming fashion.
- The proposed system achieves state-of-the-art results for streaming end-to-end ASR on the LibriSpeech corpus.

MITSUBISHI ELECTRIC

Changes for the Better