# CIF-RNNT: STREAMING ASR VIA ACOUSTIC WORD EMBEDDINGS WITH CONTINUOUS INTEGRATE-AND-FIRE AND RNN-TRANSDUCERS

Wen Shen TEO, Yasuhiro MINAMI (UEC)

## Introduction

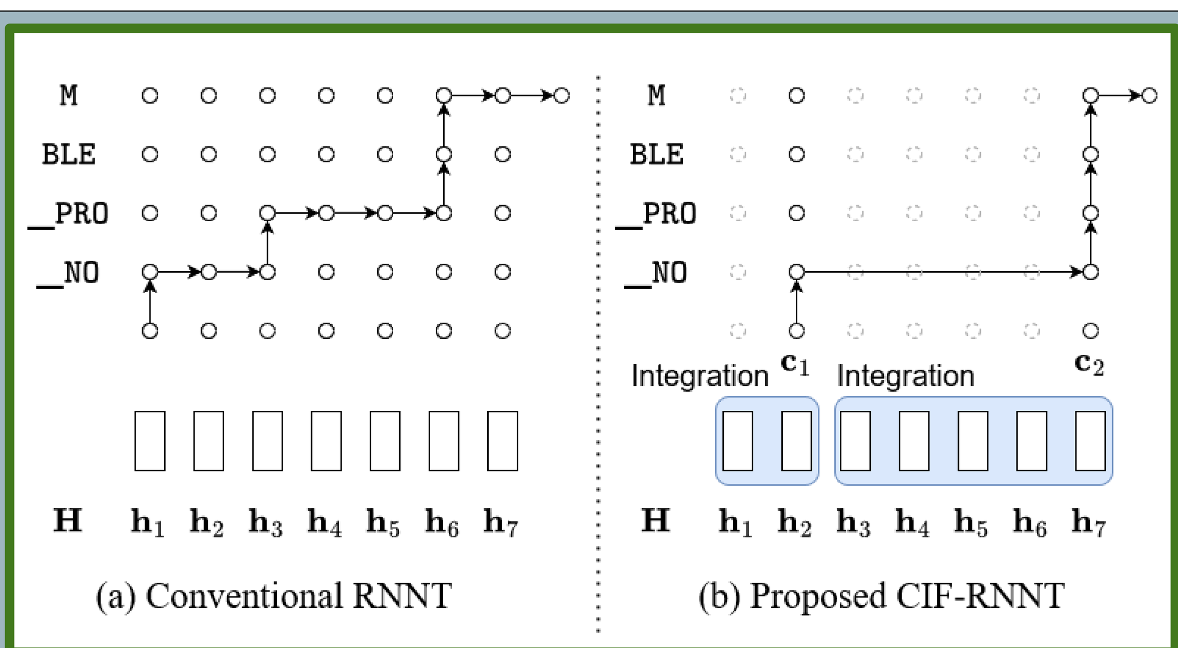
### Conventional CIF

But!!

Bridge between speech and text ➡ Speech and text must share same vocabulary!

### Idea

- CIF : Compress encoder outputs into acoustic word embeddings (AWEs)
  +
- RNN-T : Spell the word from the AWE using a limited output token set




(a) Conventional RNNT (b) Proposed CIF-RNNT

## Continuous Integrate-and-Fire [1]

Encoder outputs $\mathbf{H} = (\mathbf{h}_0, \dots, \mathbf{h}_t, \dots, \mathbf{h}_{T-1})$ ➡ CIF ➡ Compressed AWEs $\mathbf{C} = (\mathbf{c}_0, \dots, \mathbf{c}_m, \dots, \mathbf{c}_{M-1})$

Weight Prediction ($\omega$) ➡ Weight Scaling ➡ Quantity Loss ➡ Integration ($\varphi$) ➡ Training Objective

$$\alpha'_t = \begin{cases} \frac{M^*}{\hat{M}}\alpha_t, & \text{if training} \\ \alpha_t, & \text{otherwise} \end{cases}$$

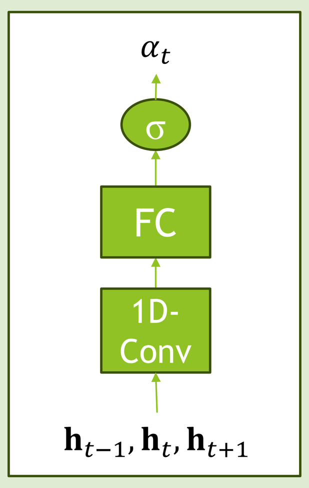
$M^*$: target length

$\hat{M} = \sum_{t=0}^{T-1} \alpha_t$ : predicted length

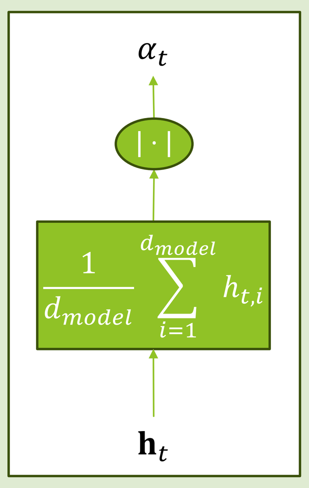
$$\mathcal{L}_{qua} = | M^* - \hat{M} |$$

- Optimized with Cross Entropy (CE) loss against target output tokens $\mathbf{Y}^*$.
  ➡ $M^* \overset{!}{=} |\mathbf{Y}^*|$

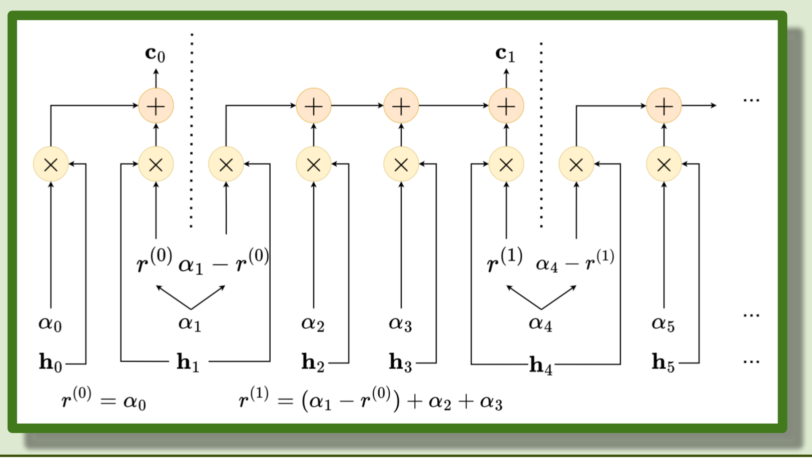
Conventional: **ConvFc**



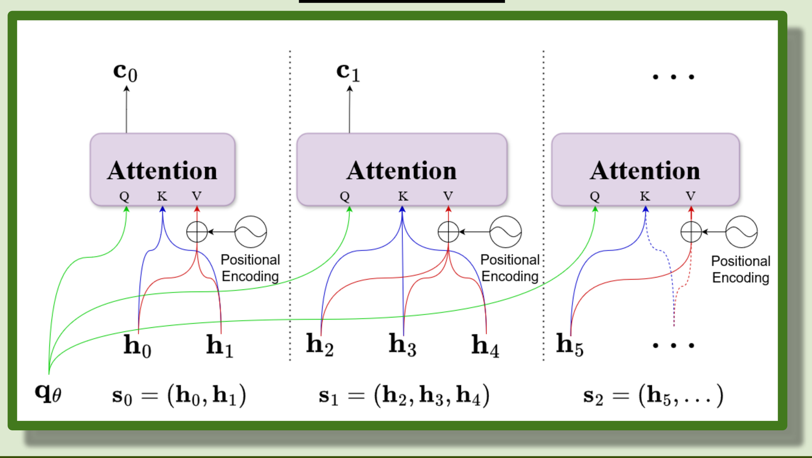

Proposed: **MeanAbs**



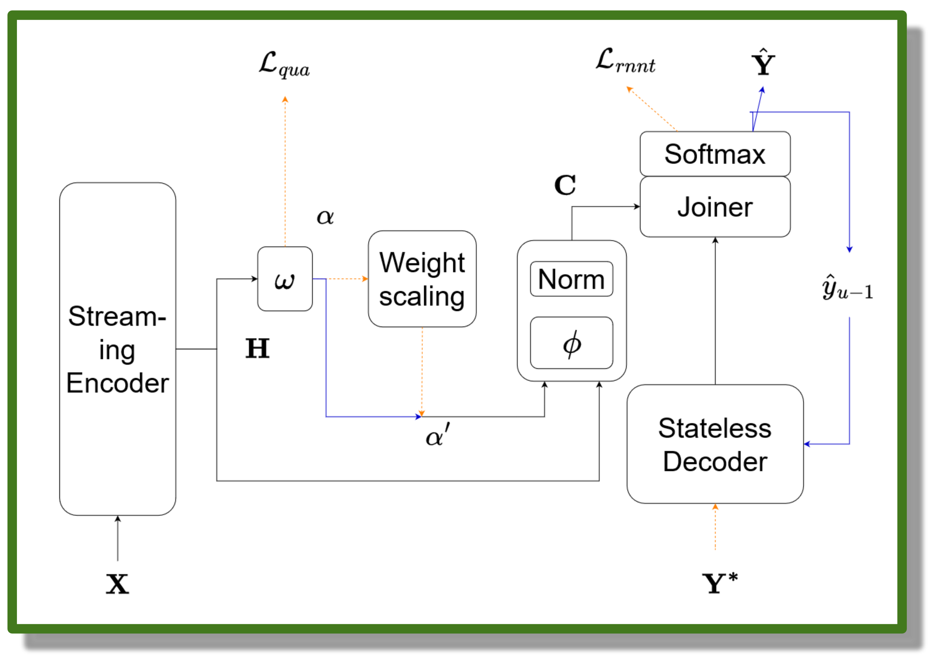

Conventional: **Cascade**




Proposed: **Ragged Attention (R.Attn)**




## CIF-RNNT




$$\mathcal{L} = \mathcal{L}_{rnnt} + \mathcal{L}_{qua}$$

| Modifications | Advantages |
|---|---|
| Detach CE loss from CIF | Detach $M$ from $|\mathbf{Y}|$ |
| | No limit to representable words |
| Insert CIF into RNN-T | Less decoding steps |

## Experiment Setup

### Datasets

**LibriSpeech**
- Training data: 'clean-100' and 'clean-360'
- Output token set: 500 BPE tokens
- $M^*$: Number of words

**CSJ**
- Training data: All
- Output token set: 3266 characters
- $M^*$: Number of morphemes
- Transcript: Fluent

### Setup

Framework: Next Generational Kaldi (NGK)
Encoder: pruned transducer stateless7 streaming [2] (removed final downsampling layer)
Decoder: stateless decoder [3] (extended context size to 4)
Pad: 30 frames at the end of utterances

### Training

| | |
|---|---|
| Epochs | 40 |
| RNNT loss | Pruned RNNT loss [4] |
| Prune range | (LibriSpeech) 16 |
| | (CSJ) 8 |
| Augmentation | Spec- and noise augmentation |

### Decoding

| | |
|---|---|
| Beam size | 4 |
| Chunk size | 640ms |
| Search method | 'batched beam search' |
| Max token per AWE | (LibriSpeech) 9 |
| | (CSJ) 5 |

## Error Rates & RTF

| Design | | LibriSpeech | | | CSJ (fluent) | | | |
|---|---|---|---|---|---|---|---|---|
| $\omega$ | $\varphi$ | clean | other | RTF | eval1 | eval2 | eval3 | RTF |
| CIF-less | | 4.27 | 13.06 | 0.0119 | 3.99 | 3.39 | 3.61 | 0.0064 |
| ConvFc | Cascade | 5.28 | 14.74 | 0.0045 | 4.98 | 4.22 | 4.30 | 0.0026 |
| MeanAbs | Cascade | 5.03 | 14.33 | 0.0042 | 4.96 | 4.25 | 4.35 | 0.0025 |
| ConvFc | R.Attn | 4.62 | 13.38 | 0.0042 | 4.45 | 3.75 | 4.29 | 0.0032 |
| MeanAbs | R.Attn | 4.41 | 13.09 | 0.0046 | 4.50 | 3.74 | 4.02 | 0.0028 |

Note:
1. Among combinations of averaged checkpoints from previous epochs up to epoch 40, models with the best (W/C)ERs on the validation set were selected to evaluate the test sets.
2. "CIF-less" refers to the model trained with the original "pruned transducer stateless7 streaming" recipe.

## Ø-WER

### Definition

- Measures capability to locate word boundaries
- Splits output of greedy path per Ø for evaluation

(Relaxed)
- To isolate identified but wrongly located Øs
- Splits word with "_" if lone Ø is encountered

| | |
|---|---|
| Ref | _THERE Ø _IS Ø _NO Ø _OPENING Ø _FOR Ø _YOU Ø |
| Hyp | _THERE _IS Ø Ø _NO Ø _OPEN Ø ING Ø _FOR _YOU Ø |
| Strict Ø-WER (100%) | ( _THERE Ø → _THERE _IS Ø) ( _IS Ø → Ø) _NO Ø ( _OPENING Ø → _OPEN Ø) (* → ING Ø) ( _FOR Ø → _FOR _YOU Ø) ( _YOU Ø → *) |
| Relaxed Ø-WER (66%) | _THERE Ø _IS Ø _NO Ø ( _OPENING Ø → _OPEN Ø) (* → ING Ø) ( _FOR Ø → _FOR _YOU Ø) ( _YOU Ø → *) |

### Results

- Only Librispeech
  ➡ Because Japanese lacks whitespace

| Design | | clean | | other | |
|---|---|---|---|---|---|
| $\omega$ | $\varphi$ | Strict | Relax | Strict | Relax |
| ConvFc | Cascade | 24.13 | 12.55 | 37.33 | 24.21 |
| MeanAbs | Cascade | 23.26 | 13.02 | 35.56 | 23.32 |
| ConvFc | R.Attn | 21.99 | 8.22 | 32.22 | 17.99 |
| MeanAbs | R.Attn | 30.20 | 7.92 | 38.55 | 17.19 |

## Conclusion

**CIF-RNNT: Incorporated CIF into RNN-Ts**

✓ Streamingly compressed acoustic information into meaningful word units (AWEs)
✓ Sped up decoding by reducing decoding operations.
✓ Minimized accuracy degradation with novel CIF mechanisms.

## References


[1] Linhao Dong and Bo Xu, "CIF: Continuous Integrate-and-Fire for End-to-End Speech Recognition", ICASSP 2020.
[2] Daniel Povey, et. al., https://github.com/k2-fsa/icefall/egs/librispeech/ASR/pruned_transducer_stateless7_streaming/zipformer.py
[3] Mohammadreza Ghodsi, et. al., "RNN Transducers with Stateless Prediction Network", ICASSP 2020.
[4] Fangjun Kuang, et. al., "Pruned RNN-T for Fast Memory-Efficient ASR Training", Interspeech 2022.



teouuenshen@gmail.com
teowenshen
Teo Wen Shen


Read my paper here!

Let's connect on LinkedIn!