# CTC Alignments Improve Autoregressive Translation

**Brian Yan*[1]** **Siddharth Dalmia*[1]** **Yosuke Higuchi[2]**
**Graham Neubig[1]** **Florian Metze[1]** **Alan W Black[1]** **Shinji Watanabe[1,3]**
[1]Language Technologies Institute, Carnegie Mellon University, USA
[2]Department of Communications and Computer Engineering, Waseda University, Japan
[3]Human Language Technology Center of Excellence, Johns Hopkins University, USA
{byan, sdalmia}@cs.cmu.edu

Carnegie Mellon University
Language Technologies Institute

CMU-LTI WAV Lab

# Does CTC make sense for translation?*

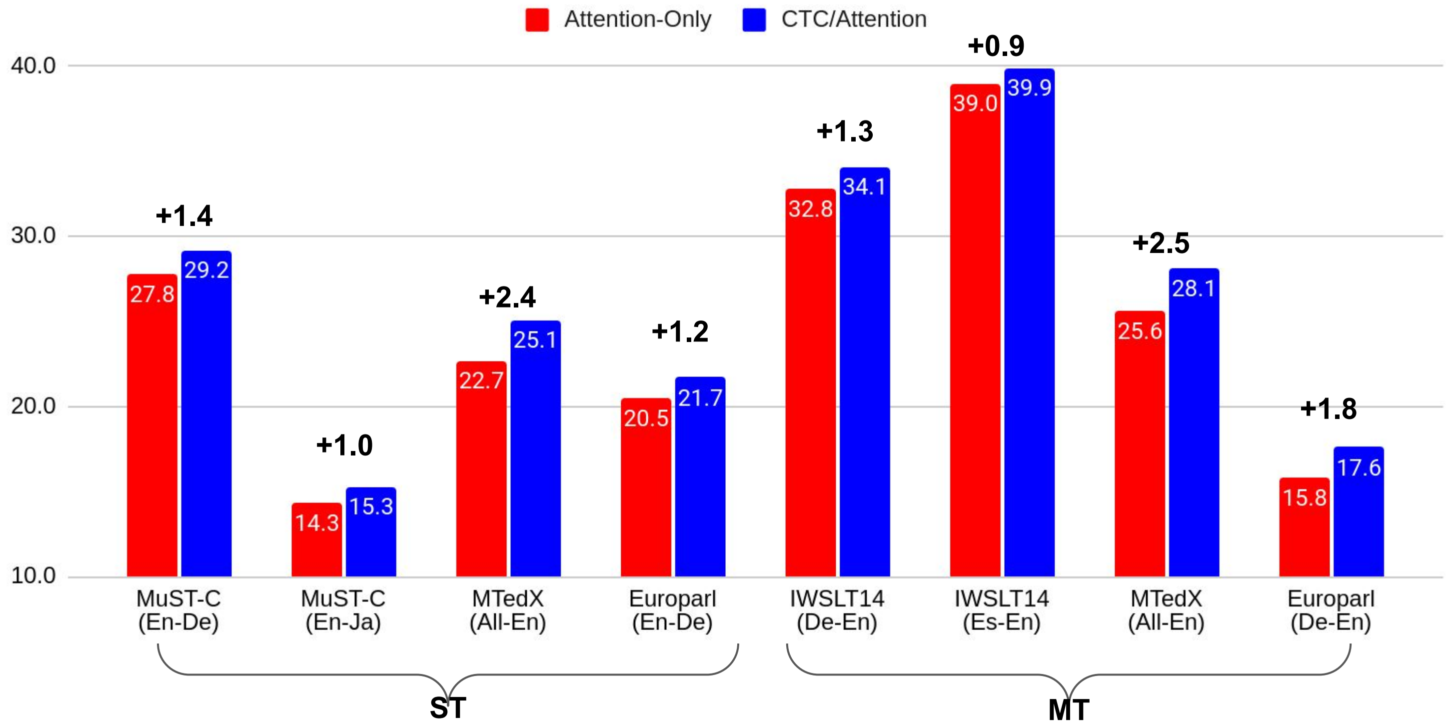


- Part 1: CTC vs. Attentional Encoder-Decoder
- Part 2: Joint CTC/Attention

* Considering translation quality only

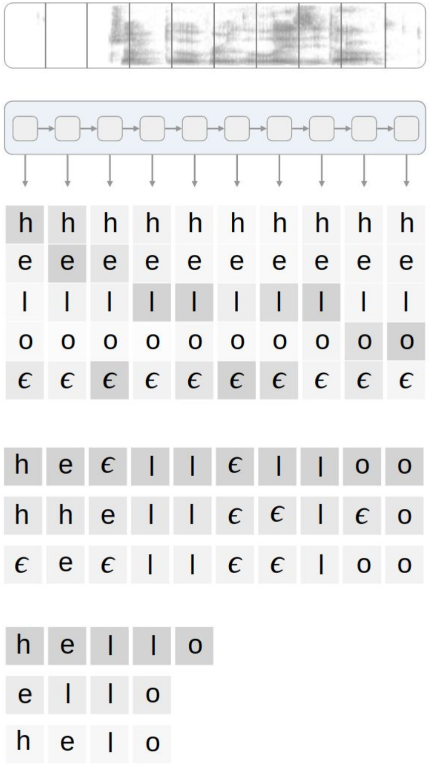
# Results (Preview)

- Joint CTC/attention outperforms pure-attention by an average of **+1.6 BLEU**




* “Attention” refers to autoregressive encoder-decoder models with cross-attention mechanisms, optimized via cross-entropy

# Connectionist Temporal Classification (CTC)




Input (speech signal) $X$

Frame-level posteriors
(aka alignment posteriors) $P(z_t|X, \cancel{z_{1:t-1}})$

Alignment sequence likelihoods $\prod_{t=1}^{T} P(z_t|X, \cancel{z_{1:t-1}})$

Label sequence likelihoods $\sum_{Z \in \mathcal{Z}} \prod_{t=1}^{T} P(z_t|X, \cancel{z_{1:t-1}})$

Image citation:
Hannun, "Sequence Modeling with CTC", Distill, 2017.

# Properties of CTC

CMU-LTI WAV Lab

| CTC |
| --- |
| $P_{\text{CTC}}(Y\|X) \triangleq \sum_{Z \in \mathcal{Z}} \prod_{t=1}^{T} P(z_t\|X, \cancel{z_{1:t-1}})$ |
| **Hard Alignment**<br>Criterion only allows monotonic alignments of inputs to outputs |
| **Conditional Independence**<br>Assumes that there are no dependencies between each output unit given the input |
| **Input-Synchronous Emission**<br>Each input representation emits exactly one blank or non-blank output token |

# CTC vs. Attention



| CTC | ATTENTION |
|---|---|
| $P_{\text{CTC}}(Y\|X) \triangleq \sum_{Z \in \mathcal{Z}} \prod_{t=1}^{T} P(z_t\|X, \cancel{z_{1:t-1}})$ | $P_{\text{Attn}}(Y\|X) \triangleq \prod_{l=1}^{L} P(y_l\|y_{1:l-1}, X)$ |
| **Hard Alignment**<br>Criterion only allows monotonic alignments of inputs to outputs | **Soft Alignment** |
| **Conditional Independence**<br>Assumes that there are no dependencies between each output unit given the input | **Conditional Dependence** |
| **Input-Synchronous Emission**<br>Each input representation emits exactly one blank or non-blank output token | **Autoregressive Generation** |

Weaker for translation

(Gu+ 2021, Huang+ 2022)

# CTC vs. Attention



| CTC |
|---|
| $P_{\text{CTC}}(Y \mid X) \triangleq \sum_{Z \in \mathcal{Z}} \prod_{t=1}^{T} P(z_t \mid X, \cancel{z_{1:t-1}})$ |
| **Hard Alignment** Criterion only allows monotonic alignments of inputs to outputs |
| **Conditional Independence** Assumes that there are no dependencies between each output unit given the input |
| **Input-Synchronous Emission** Each input representation emits exactly one blank or non-blank output token |

| ATTENTION |
|---|
| $P_{\text{Attn}}(Y \mid X) \triangleq \prod_{l=1}^{L} P(y_l \mid y_{1:l-1}, X)$ |
| **Soft Alignment** Flexible attention-based input-to-output mappings may overfit to irregular patterns |
| **Conditional Dependence** Locally normalized models with output dependency exhibit label/exposure biases |
| **Autoregressive Generation** Need to detect end-points and compare hypotheses of different length in beam search |

Attention is not perfect

(Murray+ 2018, Hannun+ 2019, Watanabe+ 2018)

# CTC and Attention are complementary



- Joint CTC/attention is excellent for ASR (Kim+ 2017, Watanabe+ 2018)

- Joint CTC/attention should also benefit MT/ST due to positive interactions:
  - Hard alignment + soft alignment
  - Conditional independence + conditional dependence
  - Input synchronous emission + autoregressive generation

# CTC/Attention for MT/ST (1)

CMU-LTI WAV Lab

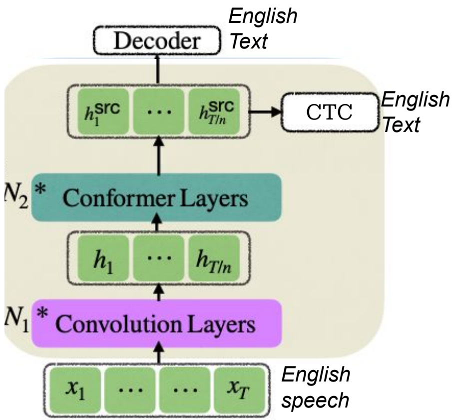
- Hard alignment + soft alignment
  - Conjecture: hard alignment objective produces stable encoder representations allowing the decoder to **more easily learn soft alignment patterns during training**
- **Can CTC encoders perform input-to-output mappings for translation?**
  - Outputs may be longer than inputs (Libovicky+ 2018, Dalmia+ 2022)
  - Input-to-output re-ordering (Chuang+ 2021)

→ Let's incorporate these requirements into our encoder architecture

# Joint Training with Hierarchical CTC

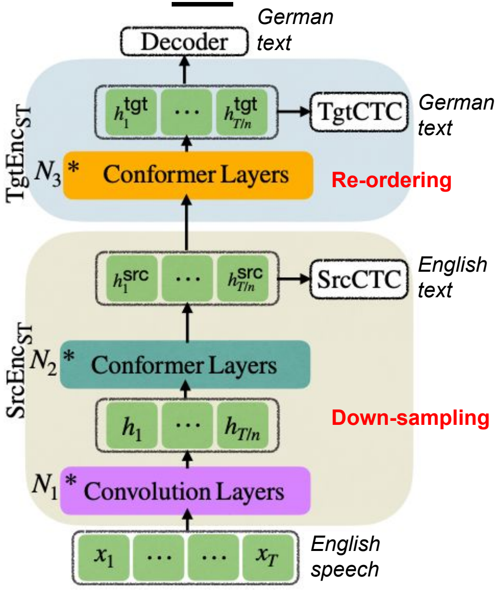
- For ASR, CTC and attention simply share a monolithic encoder
- For MT/ST, we decompose the encoder into 2 stages: **1) length-adjustment 2) re-ordering**

## ASR

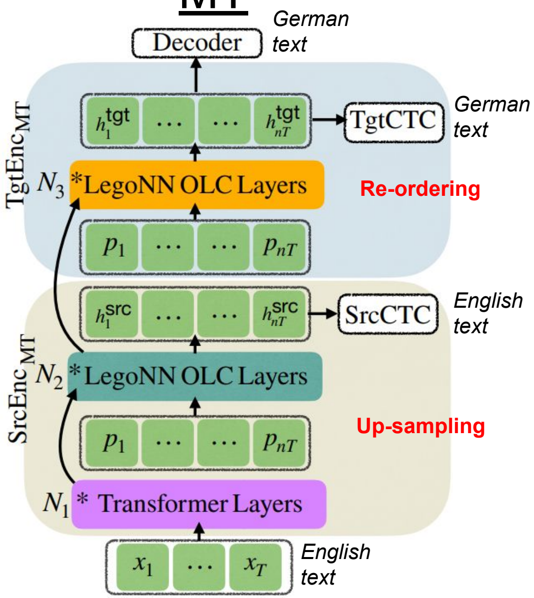



$$\mathcal{L} = \lambda_1 \mathcal{L}_{\text{CTC}} + \lambda_2 \mathcal{L}_{\text{ATTN}}$$

## ST




## MT




$$\mathcal{L} = \mathcal{L}_{\text{SRCCTC}} + \lambda_1 \mathcal{L}_{\text{TGTCTC}} + \lambda_2 \mathcal{L}_{\text{ATTN}}$$

# Joint Training with Hierarchical CTC

CMU-LTI WAV Lab

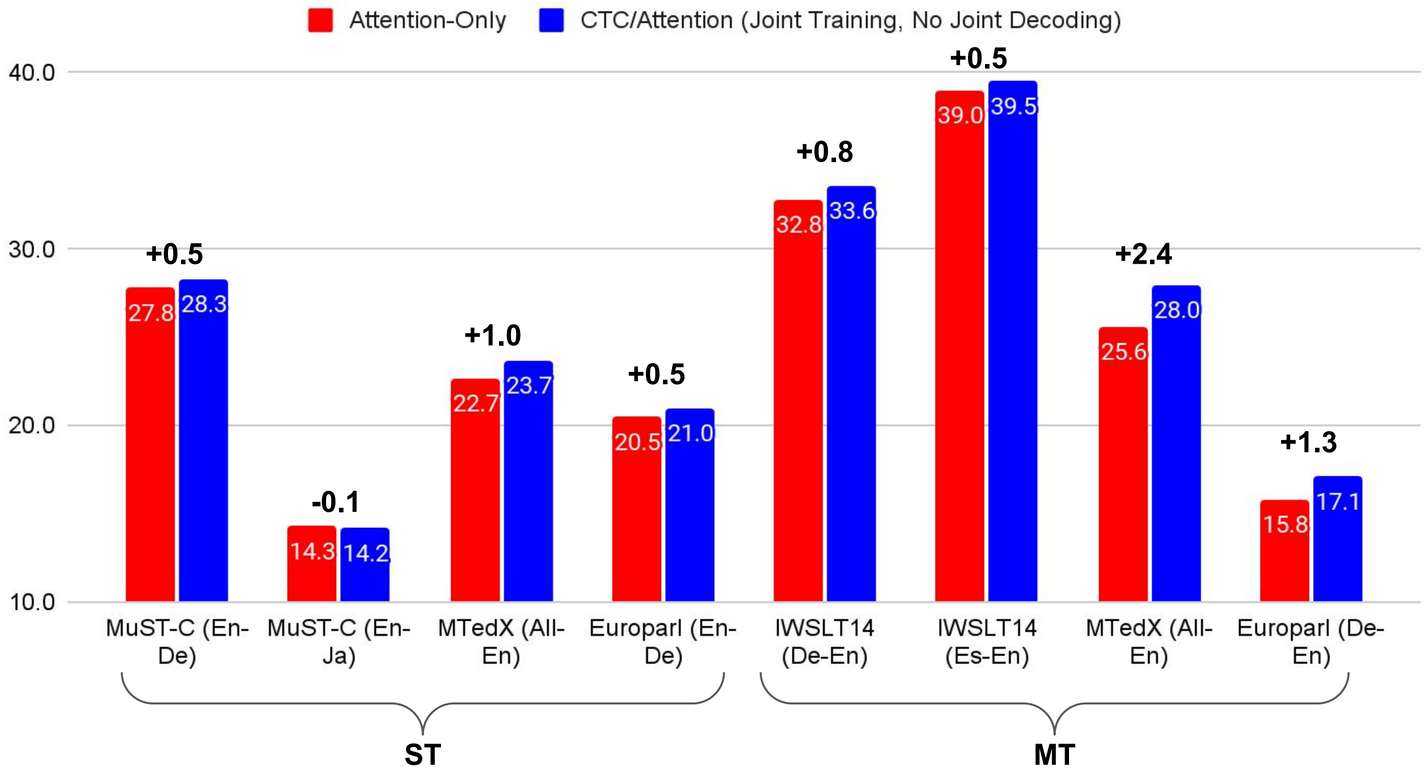
- Ablation shows that separating length-adjustment and re-ordering is beneficial

| | | MT (De-En) | ST (En-De) |
|---|---|---|---|
| SrcCTC | TgtCTC | IWSLT14 | MuST-C-v2 |
| ✗ | ✗ | 32.1 | 27.7 |
| ✓ | ✗ | 34.1 | 27.8 |
| ✗ | ✓ | 33.3 | 28.1 |
| ✓ | ✓ | **34.8** | **28.3** |

# Attention w/ CTC Joint Training vs. Pure-Attention

CMU-LTI WAV Lab

- Joint training yields an average of **+0.9** BLEU improvement

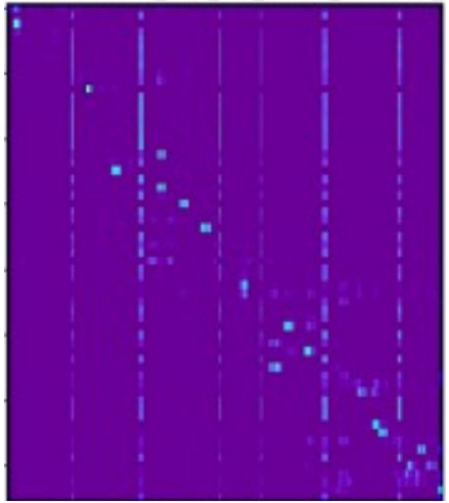


# Reduced Soft Alignment Burden

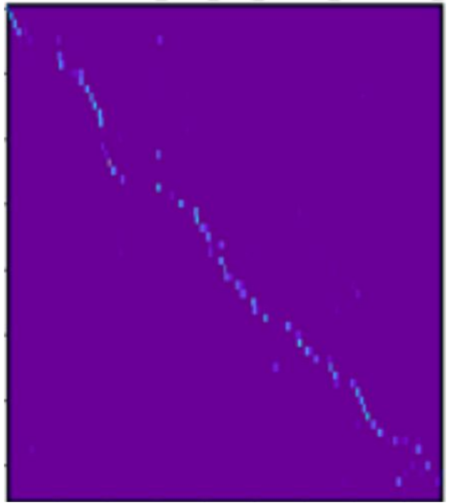


- Joint training results in more regular, diagonal source-attention patterns

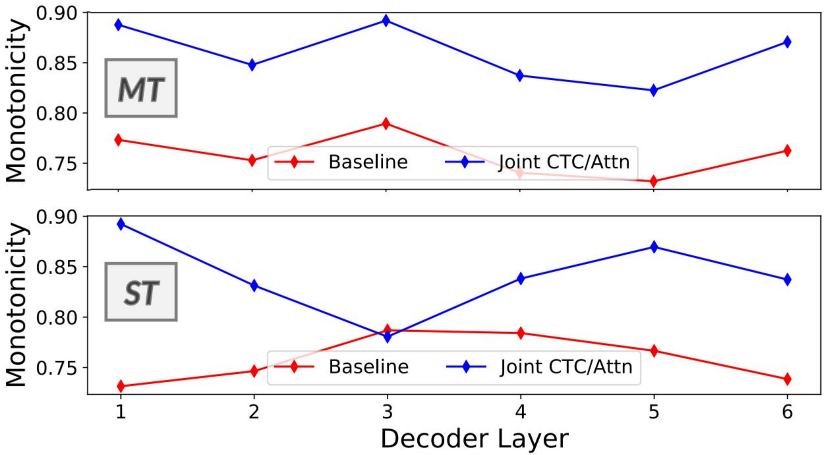
Pure-Attention



CTC/Attention



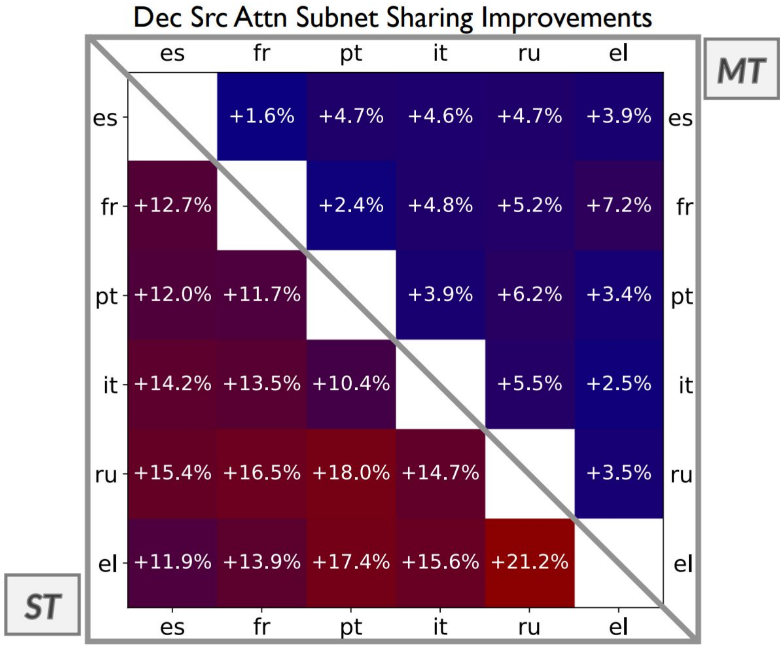
Layer-wise Monotonicity



# Increased Multilingual Parameter Sharing



- For X→En models, decoder source-attention parameter sharing between languages was higher in CTC/attention vs. pure-attention




Language specific subnets extracted via Lottery Ticket Sparse Fine-Tuning (Ansell+ 2022)

# CTC/Attention for MT/ST (2)

CMU-LTI WAV Lab

- Conditional independence + conditional dependence
  - Conjecture: use of conditionally independent likelihoods in joint scoring **eases the exposure/label biases** from conditionally dependent likelihoods **during decoding**
- **Does CTC translation quality lag too far behind attention to be useful?**
  - In our study, pure-CTC models are up to 28% worse than pure-attention models

→ Let’s examine joint decoding of CTC/attention

# Joint Decoding with Output-Sync Beam Search



- Attention plays a primary role while CTC plays a secondary role (Watanabe+ 2018)

**Algorithm 1** *Output*-Synchronous Step Function: attentional decoder proposes candidates to expand hypotheses which are all of $l$-length at step $l$.

**procedure** $\textsc{OutputStep}(\mathrm{prtHs}, X, l, p, \max L)$
  $\mathrm{newPrtHs} = \{\}$; $\mathrm{endHs} = \{\}$
  **for** $y_{1:l-1} \in \mathrm{prtHs}$ **do**
    $\boxed{\mathrm{attnCnds}} = \text{top-k}(P_{\mathrm{Attn}}(y_l | X, y_{1:l-1}), \mathrm{k} = p)$
    **for** $c \in \mathrm{attnCnds}$ **do**
      $y_{1:l} = y_{1:l-1} \oplus \mathrm{c}$
      $\alpha_{\mathrm{CTC}} = \mathrm{CTCScore}(y_{1:l}, \boxed{X_{1:T}})$
      $\alpha_{\mathrm{Attn}} = \mathrm{AttnScore}(y_{1:l}, X_{1:T})$
      $\beta = \mathrm{LengthPen}(y_{1:l})$
      $P_{\mathrm{Beam}}(y_{1:l}|X) = \alpha_{\mathrm{CTC}} + \alpha_{\mathrm{Attn}} + \beta$
      **if** $\boxed{(c \text{ is } \langle\mathrm{eos}\rangle) \text{ or } (l \text{ is } \max L)}$ **then**
        $\mathrm{endHs}[y_{1:l}] = P_{\mathrm{Beam}}(\cdot)$
      **else**
        $\mathrm{newPrtHs}[y_{1:l}] = P_{\mathrm{Beam}}(\cdot)$
      **end if**
    **end for**
  **end for**
  **return** newPrtHs, endHs
**end procedure**

Hypothesis Expansion — Choose top-$p$ candidates per hypothesis (e.g. $p$=1.5*$b$) **from attention posteriors**

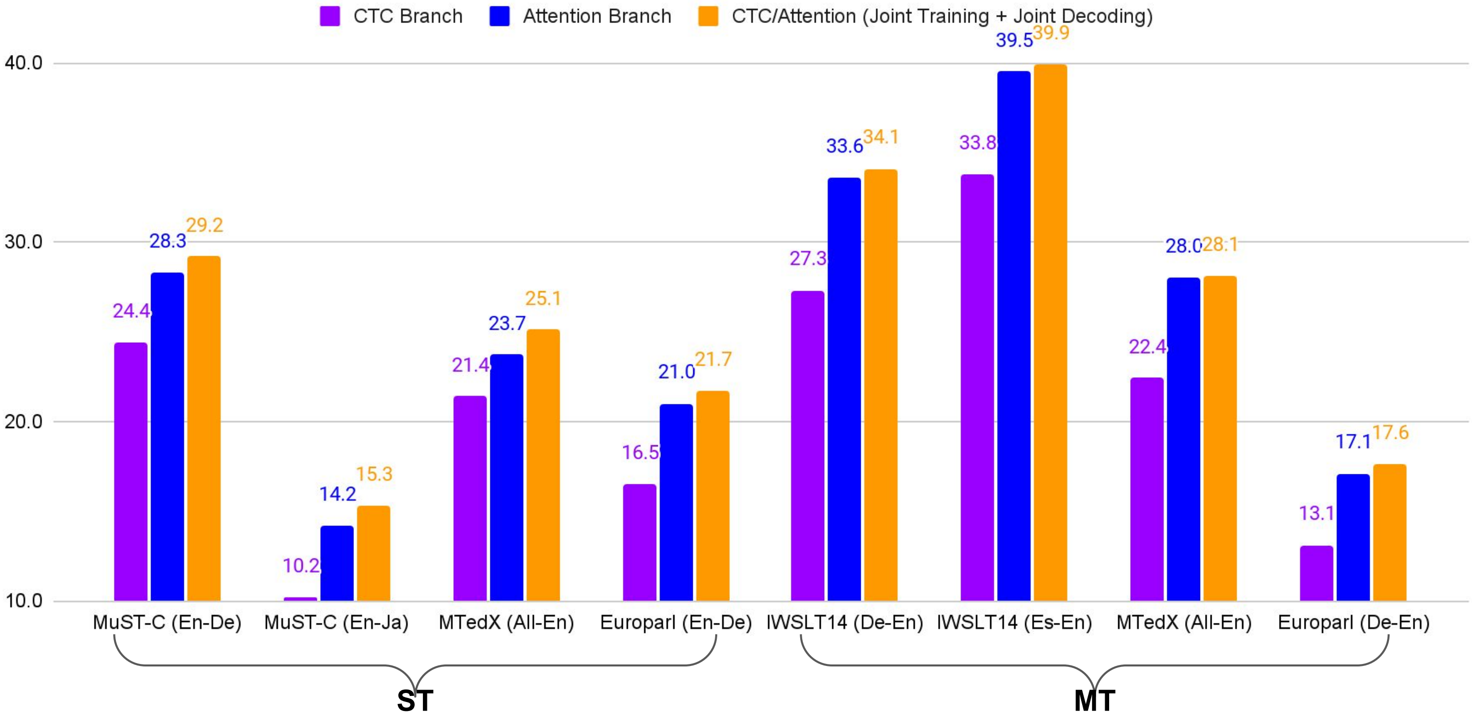
Joint Scoring — Interpolate label sequence likelihoods **from attention and CTC**

End Detection — End loop based on conditions **from attention** (CTC's role in end-detection is implicit)

# Joint Training + Decoding vs. Only Joint Training

CMU-LTI WAV Lab

- Joint decoding yields an average of **+0.7** BLEU improvement over the attention branch
- For jointly trained models, attention branch outperforms CTC branch by an avg. of **+4.5** BLEU



# CTC/Attention for MT/ST (3)

CMU-LTI WAV Lab

- Input synchronous emission + autoregressive generation
  - Conjecture: input-synchronous emission determines output length based on input length **counteracting the autoregressive end-detection problem during decoding**
- **Is the alignment information from CTC translation models reasonable?**
  - Even in ASR, some alignment “drift” can occur (Kurzinger+ 2020)
  - If alignments are highly noisy, CTC’s end-detection property may not be useful

→ We address this via sanity checks

# Joint Decoding with Input-Sync Beam Search

- CTC plays a primary role while attention plays a secondary role (inverse of output-sync)

**Algorithm 2** *Input*-Synchronous Step Function: CTC proposes candidates to expand hypotheses which are all produced from $t$ input units at step $t$.

**procedure** INPUTSTEP(prtHs, $X, t, p, T$)
  newPrtHs = {}; endHs = {}
  $\boxed{\text{CTCCnds}} = \text{top-k}(P_{\text{CTC}}(z_t|X), \text{k} = p)$
  **for** $y \in$ prtHs **do**
    **for** $c \in$ CTCCnds **do**
      **if** ($c$ is $\varnothing$) or ($c$ is $y[-1]$) **then**
        $\tilde{y} = y$
      **else**
        $\tilde{y} = y \oplus \text{c}$
      **end if**
      $\alpha_{\text{CTC}} = \text{CTCScore}(\tilde{y}, \boxed{X_{1:t}})$
      $\alpha_{\text{Attn}} = \text{AttnScore}(\tilde{y}, X_{1:T})$
      $\beta = \text{LengthPen}(\tilde{y})$
      $P_{\text{Beam}}(\tilde{y}|X) = \alpha_{\text{CTC}} + \alpha_{\text{ATTN}} + \beta$
      **if** $\boxed{t \text{ is } T}$ **then**
        $\text{endHs}[\tilde{y}] = P_{\text{Beam}}(\cdot)$
      **else**
        $\text{newPrtHs}[\tilde{y}] = P_{\text{Beam}}(\cdot)$
      **end if**
    **end for**
  **end for**
  **return** newPrtHs, endHs
**end procedure**

Hypothesis Expansion → Choose top-$p$ candidates (e.g. $p$=1.5*$b$) **from CTC posterior**

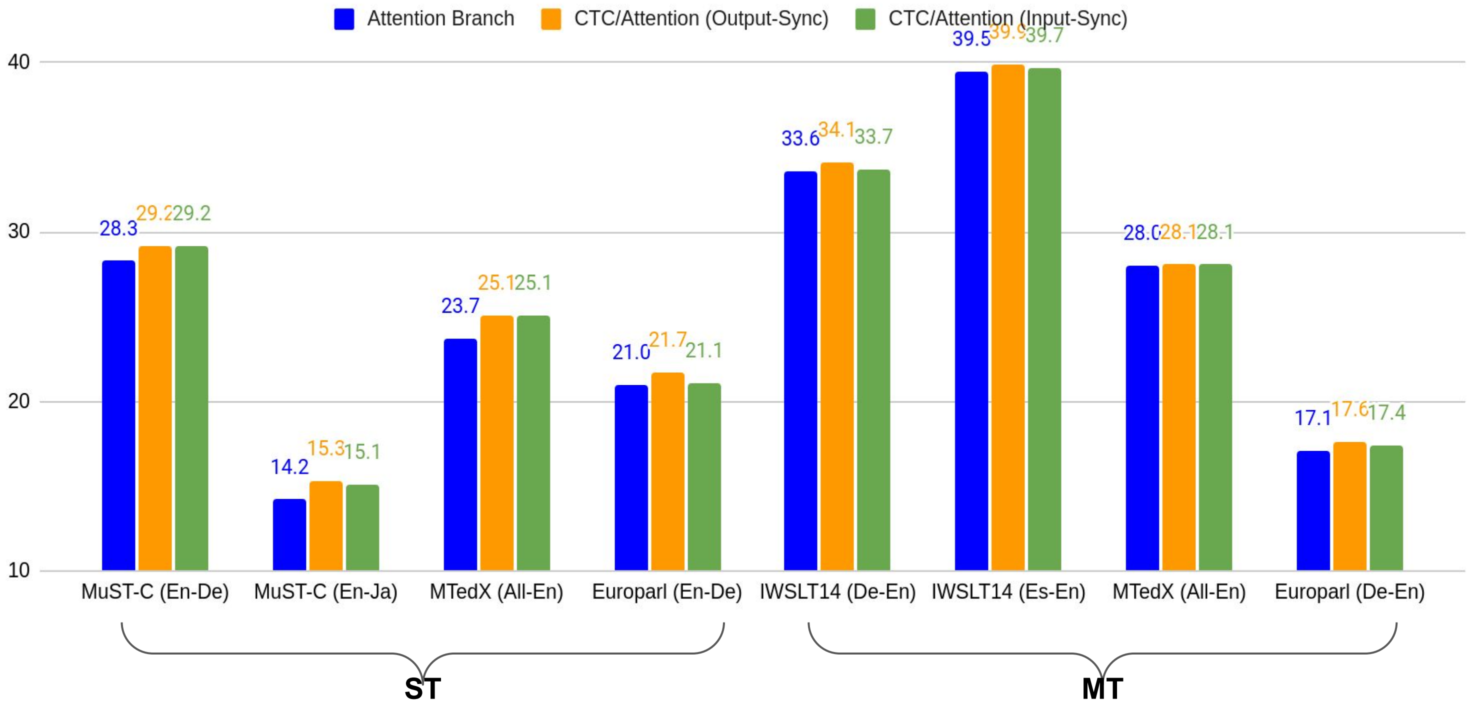
Joint Scoring → Interpolate label sequence likelihoods **from attention and CTC**

End Detection → End loop based on conditions **from CTC**

# Input-Sync vs. Output-Sync Joint Decoding

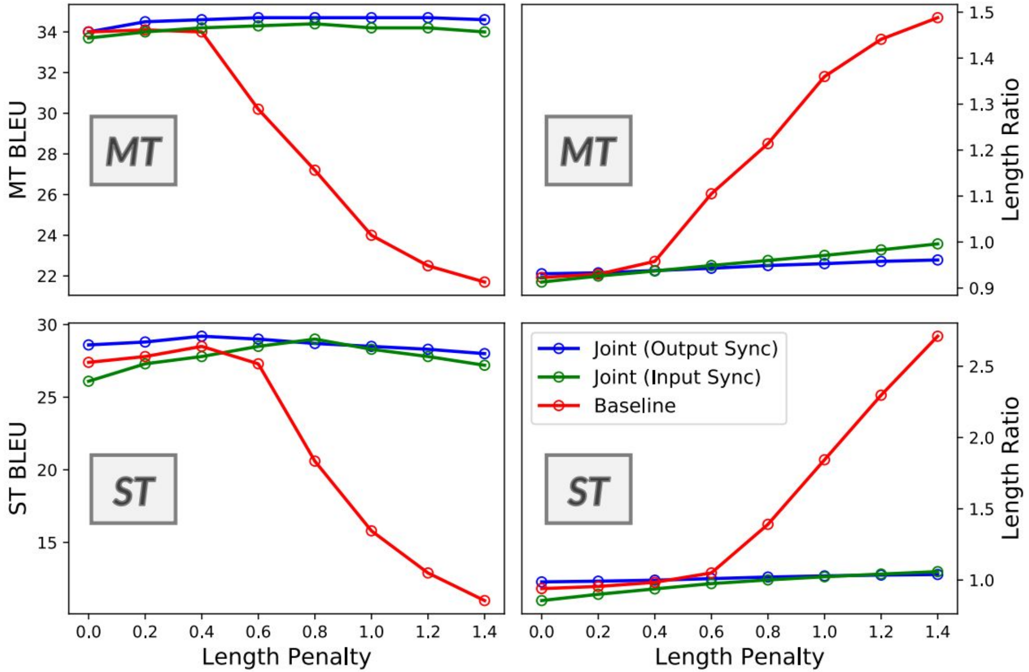
CMU-LTI WAV Lab

- Input-sync joint CTC/attention outperforms the attention branch by **+0.5** BLEU
- Input-sync is only **-0.2** BLEU worse than output-sync



# Robust End-Detection

- Both variants of joint CTC/attention decoding have low length penalty elasticity
- Pure-attention models are highly sensitive to length penalty → easily overtuned



# Summary

CMU-LTI WAV Lab

- Joint CTC/attention is effectively applied to MT/ST with only minor changes from ASR
- Both joint training and joint decoding yield performance gains
- Why is joint CTC/attention better than pure-attention?
  - Simplifies the soft alignment task
  - Positive ensembling effect
  - Robust end-detection

# Thank You!