CR-CTC: Consistency regularization on CTC for improved speech recognition

Zengwei Yao, 20250330

Background

- End-to-end approaches for automatic speech recognition (ASR)
  - Connectionist Temporal Classification (**CTC**)
  - **Transducer** (also known as RNN-T)
  - Combining CTC and attention-based encoder-decoder (AED), referred to as **CTC/AED**
- Among these, **CTC** is the simplest and most computationally efficient
- However, it significantly **lags behind transducer and CTC/AED in recognition performance**, which limits its applicability.

Method

Consistency-Regularized CTC (CR-CTC)

- **Different augmented views**
  a) Time warping before duplicating
  b) Duplicate -> two copies
  c) Random frequency masking and time masking on two copies (**using larger amount of time masking**)
- **Consistency regularization loss**
  - Bidirectional D_{KL} on each pair of distributions at frame t
  - $\mathcal{L}_{\text{CR}}(\mathbf{z}^{(a)}, \mathbf{z}^{(b)}) = \frac{1}{2}\sum_{1}^{T} D_{KL}(\text{sg}(z_t^{(b)})||z_t^{(a)}) + D_{KL}(\text{sg}(z_t^{(a)})||z_t^{(b)})$
- **Overall loss:**
  - $\mathcal{L} = \frac{1}{2}\left(\mathcal{L}_{\text{CTC}}(\mathbf{x}^{(a)}, \mathbf{y}) + \mathcal{L}_{\text{CTC}}(\mathbf{x}^{(b)}, \mathbf{y})\right) + \alpha\mathcal{L}_{\text{CR}}(\mathbf{z}^{(a)}, \mathbf{z}^{(b)})$

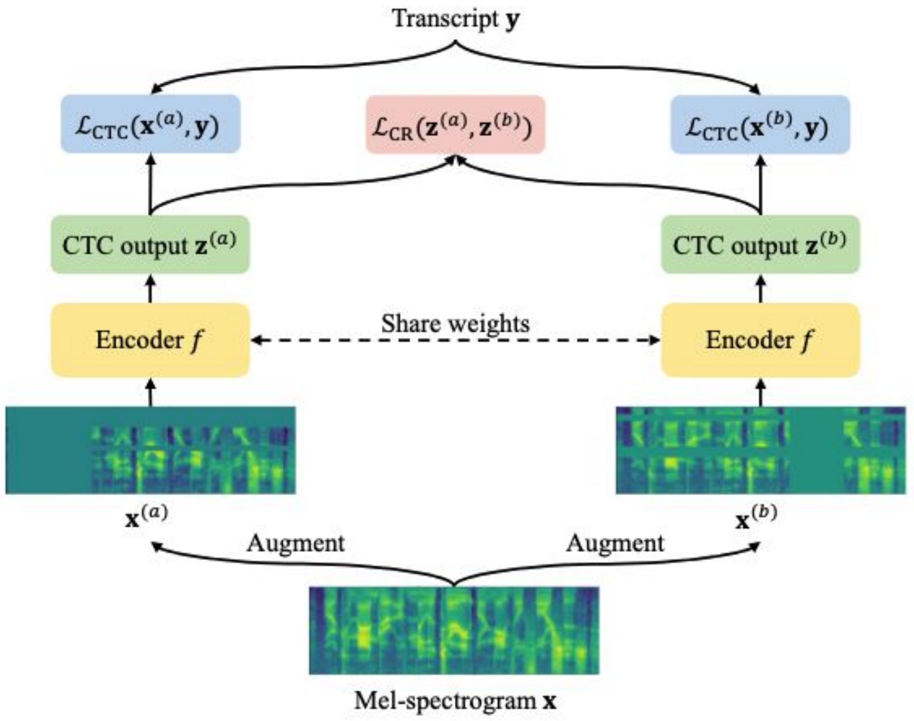


Figure 1: Overall architecture of *CR-CTC*.

Method

Explanations (1/3)

- **Self-distillation**
  - Using dropout and stochastic depth: **implicitly training randomly sampled sub-models** -> ultimately combined into an ensemble during inference
  - CR-CTC performs **self-distillation between pairs of randomly sampled sub-models**, with each sub-model receiving supervision signals in the form of per-frame predictions from the other
  - **Using different augmented views** (with larger amount of time masking) exposes these sub-models to varied aspects of the input data -> **enhancing their prediction diversity** -> richer knowledge transfer

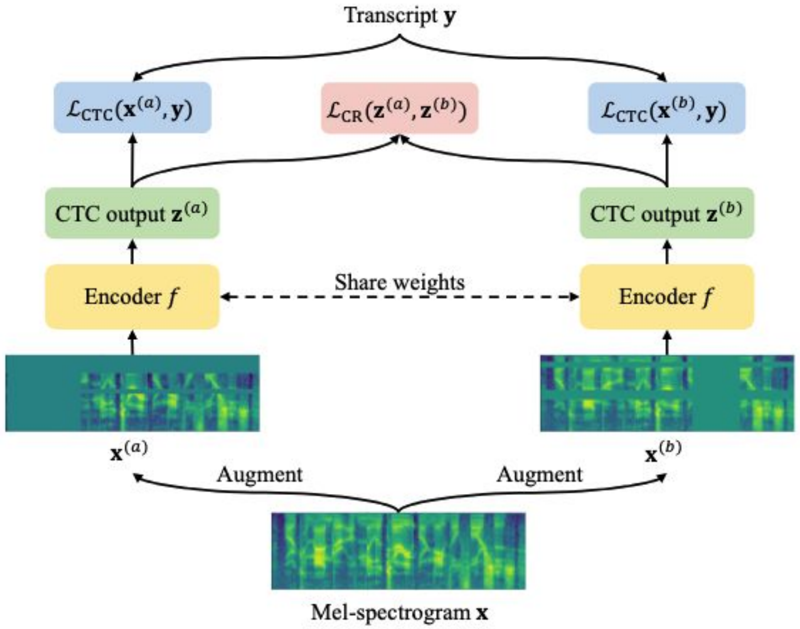


Figure 1: Overall architecture of *CR-CTC*.

Method

Explanations (1/3)

- **Self-distillation**
  - No larger time masking, no different augmented views -> worse results
  - Hard-label CE-based $\mathcal{L}_{\mathrm{CR}}$ only distills the best alignment, while the D_{KL}-based $\mathcal{L}_{\mathrm{CR}}$ distills the full CTC distribution
  - Remove sg in $\mathcal{L}_{\mathrm{CR}}$ -> the model might have a tendency towards a degenerated solution that is insensitive to the pattern of input masking and model dropout.

Table 4: Ablation studies for self-distillation in *CR-CTC* on LibriSpeech dataset using Zipformer-M encoder and greedy search decoding.

Method	WER (%) *test-clean*	WER (%) *test-other*
CTC baseline	2.51	6.02
***CR-CTC* (final)**	**2.12**	**4.62**
No larger time masking	2.19	4.98
No larger time masking, no different augmented views	2.27	5.11
Use hard-label CE-based $\mathcal{L}_{\mathrm{CR}}$	2.14	4.84
Remove *sg* in $\mathcal{L}_{\mathrm{CR}}$	2.24	4.97

Method

Explanations (2/3)

- **Masked prediction**
  - CR-CTC requires frames within the time-masked regions in each branch to predict the corresponding token distributions
  - **Similar to masked-based self-supervised models,** this behavior encourages the model to capture acoustic information on the unmasked context and exploit its implicit language modeling capability
  - **Different augmented views** -> reduces the occurrence of positions masked by both branches -> improve the quality of the provided target distributions for these masked positions
  - **Larger amount of time masking** -> enhance contextual representation learning through the masked prediction behavior

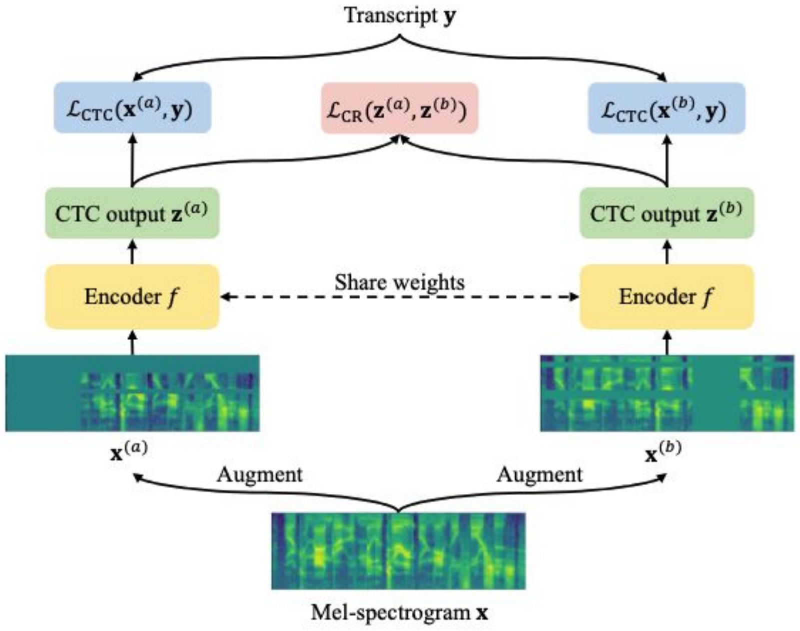


Figure 1: Overall architecture of *CR-CTC*.

Method

Explanations (2/3)

- **Masked prediction**
  - No larger time masking, no different augmented views -> worse results
  - Larger amount of frequency masking -> slightly worse result
  - Larger amount of time masking in CTC baseline -> worse result
  - Excluding self-masked frames leads to a larger WER degradation than excluding self-unmasked frames

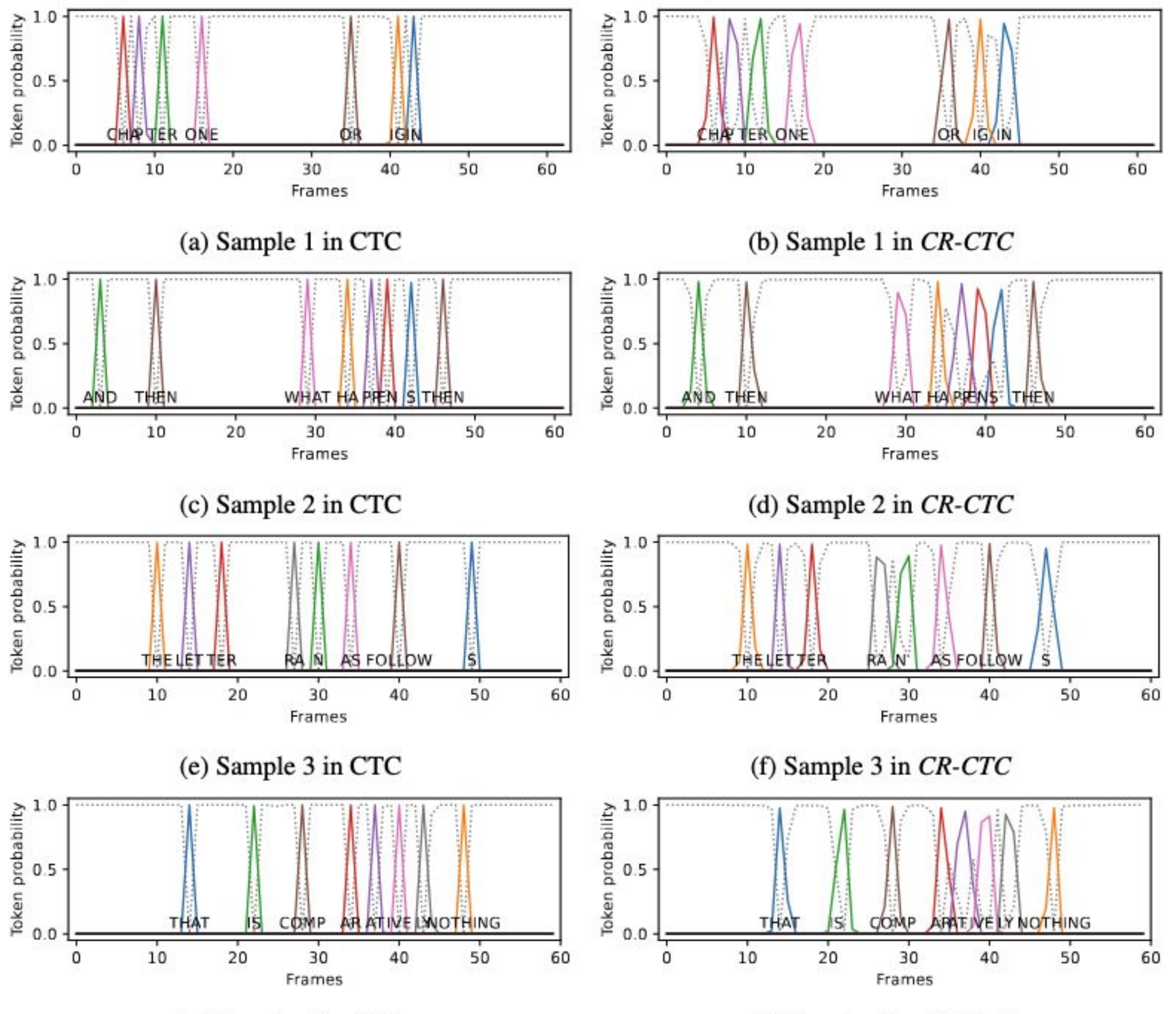
Table 5: Ablation studies for masked prediction in *CR-CTC* on LibriSpeech dataset using Zipformer-M encoder and greedy search decoding.

Method	WER (%) *test-clean*	WER (%) *test-other*
CTC baseline	2.51	6.02
Use larger time masking	2.68	6.28
CR-CTC **(final)**	**2.12**	**4.62**
No larger time masking	2.19	4.98
No larger time masking, no different augmented views	2.27	5.11
No larger time masking, use larger frequency masking	2.26	4.98
Exclude self-masked frames in $\mathcal{L}_{\mathrm{CR}}$	2.32	5.26
Exclude self-unmasked frames in $\mathcal{L}_{\mathrm{CR}}$	2.32	5.02

Method

Explanations (3/3)

- **Peak suppression**
  - CTC tends to learn extremely peaky distributions, suggesting potential **overfitting**
  - CR-CTC guides the model to learn the average of their prediction -> **smoother distributions -> reduces overconfidence -> better generalization**
  - Smooth-regularized CTC (**SR-CTC**)
    - Apply a smooth kernel $K = (0.25, 0.5, 0.25)$
    - $\mathbf{z}^{(s)} = smooth(\mathbf{z}, K)$
    - $\mathcal{L}_{\mathrm{SR}}(\mathbf{z}, \mathbf{z}^{(s)}) = \sum_1^T D_{KL}(\mathrm{sg}(z_t^{(s)}) || z_t)$
    - $\mathcal{L} = \mathcal{L}_{\mathrm{CTC}}(\mathbf{x}, \mathbf{y}) + \beta \mathcal{L}_{\mathrm{SR}}(\mathbf{z}, \mathbf{z}^{(s)})$



(a) Sample 1 in CTC
(b) Sample 1 in *CR-CTC*
(c) Sample 2 in CTC
(d) Sample 2 in *CR-CTC*
(e) Sample 3 in CTC
(f) Sample 3 in *CR-CTC*
(g) Sample 4 in CTC
(h) Sample 4 in *CR-CTC*

Figure 2: Visualization of token emitting probabilities for vanilla CTC (left) and our *CR-CTC* (right) on four randomly selected samples from LibriSpeech test set. The gray dashed lines indicate the blank token. Compared to vanilla CTC, the token distributions in *CR-CTC* are smoother with lower emitting probabilities and more repeating non-blank tokens.

Method

Explanations (3/3)

- **Peak suppression**
  - Compared to the CTC baseline, CR-CTC learns smoother distributions and significantly improves the recognition performance
  - SR-CTC also surpasses the CTC baseline while exhibiting a notably larger average duration of non-blank tokens.

Table 6: Ablation studies for peak suppression in *CR-CTC* on LibriSpeech dataset using Zipformer-M encoder and greedy search decoding. We include the averaged duration of all non-blank tokens, as well as the averaged emitting probabilities of the blank token and all non-blank tokens on the best alignments.

Method	Non-blank duration (frames)	Emit probability (%)		WER (%)	
		blank	non-blank	*test-clean*	*test-other*
CTC baseline	1.04	99.64	98.50	2.51	6.02
SR-CTC	4.25	95.44	90.04	2.32	5.22
CR-CTC	1.28	94.19	89.42	**2.12**	**4.62**

Experiment

Compared to using auxiliary head for jointly training

- w/ AED head
- w/ pruned transducer head

Table 7: Comparison between *CR-CTC* and methods using an auxiliary head for jointly training on LibriSpeech dataset using Zipformer-M encoder and greedy search decoding.

Method	Params (M)	WER (%)	
		test-clean	*test-other*
CTC baseline	64.3	2.51	6.02
CTC w/ AED head	90.0	2.46	5.57
CTC w/ pruned transducer head	65.8	2.42	5.4
CR-CTC	64.3	**2.12**	**4.62**

Experiment

- **LibriSpeech dataset (1000h), no external language model**

Table 1: WER(%) performance of our method on LibriSpeech dataset compared to the best results reported in the literature without using an external language model.

Model	Params (M)	WER (%) test-clean	WER (%) test-other
CTC/AED, E-Branchformer-B (Kim et al., 2023)	41.1	2.49	5.61
CTC/AED, Branchformer (Peng et al., 2022)	116.2	2.4	5.5
CTC/AED, E-Branchformer-L (Kim et al., 2023)	148.9	2.14	4.55
Transducer, ContextNet-S (Han et al., 2020)	10.8	2.9	7.0
Transducer, ContextNet-M (Han et al., 2020)	31.4	2.4	5.4
Transducer, ContextNet-L (Han et al., 2020)	112.7	2.1	4.6
Transducer, Conformer-S (Gulati et al., 2020)	10.3	2.7	6.3
Transducer, Conformer-M (Gulati et al., 2020)	30.7	2.3	5.0
Transducer, Conformer-L (Gulati et al., 2020)	118.8	2.1	4.3
Transducer, MH-SSM 32L (Fathullah et al., 2023)	140.3	2.01	4.61
Transducer, Stateformer 25L (Fathullah et al., 2023)	139.8	1.91	4.36
CTC/AED, Zipformer-S (Yao et al., 2024)	46.3	2.46	6.04
CTC/AED, Zipformer-M (Yao et al., 2024)	90.0	2.22	4.97
CTC/AED, Zipformer-L (Yao et al., 2024)	174.3	2.09	4.59
Pruned transducer, Zipformer-S (Yao et al., 2024)	23.3	2.42	5.73
Pruned transducer, Zipformer-M (Yao et al., 2024)	65.6	2.21	4.79
Pruned transducer, Zipformer-L (Yao et al., 2024)	148.4	2.00	4.38
CTC, Zipformer-S	22.1	2.85	6.89
CTC, Zipformer-M	64.3	2.52	6.02
CTC, Zipformer-L	147.0	2.5	5.72
CR-CTC, Zipformer-S (ours)	22.1	2.52	5.85
CR-CTC, Zipformer-M (ours)	64.3	2.1	4.61
CR-CTC, Zipformer-L (ours)	147.0	2.02	4.35
CR-CTC/AED, Zipformer-L (ours)	174.3	1.96	4.08
Pruned transducer w/ *CR-CTC*, Zipformer-L (ours)	148.8	**1.88**	**3.95**

Experiment

- **Aishell-1 dataset (170h), no external language model**

Table 2: WER(%) performance of our method on Aishell-1 dataset compared to the best results reported in the literature without using an external language model.

Model	Params (M)	WER (%) *dev*	WER (%) *test*
CTC/AED, Conformer in ESPnet (Watanabe et al., 2018)	46.2	4.5	4.9
CTC/AED, Conformer in WeNet (Yao et al., 2021)	46.3	–	4.61
CTC/AED, E-Branchformer in ESPnet (Watanabe et al., 2018)	37.9	4.2	4.5
CTC/AED, Branchformer (Peng et al., 2022)	45.4	4.19	4.43
Pruned transducer, Zipformer-S (Yao et al., 2024)	30.2	4.4	4.67
Pruned transducer, Zipformer-M (Yao et al., 2024)	73.4	4.13	4.4
CTC, Zipformer-S	23.1	4.89	5.26
CTC, Zipformer-M	66.2	4.47	4.8
CTC/AED, Zipformer-S	39.3	4.47	4.8
CTC/AED, Zipformer-M	83.2	4.0	4.32
CR-CTC, Zipformer-S (ours)	23.1	3.9	4.12
CR-CTC, Zipformer-M (ours)	66.2	**3.72**	**4.02**

Experiment

- **GigaSpeech dataset (10000h), no external language model**

Table 3: WER(%) performance of our method on GigaSpeeech dataset compared to the best results reported in the literature without using an external language model.

Model	Params (M)	WER (%) *dev*	WER (%) *test*
CTC/AED, Transformer (Chen et al., 2021a)	87	12.30	12.30
CTC/AED, Conformer in Wenet (Zhang et al., 2022)	113.2	10.7	10.6
CTC/AED, Conformer in ESPnet (Chen et al., 2021a)	113.2	10.9	10.8
CTC/AED, E-Branchformer in ESPnet (Watanabe et al., 2018)	148.9	10.6	10.5
CTC, Zipformer-S	22.1	12.08	11.95
CTC, Zipformer-M	64.3	11.23	11.27
CTC, Zipformer-L	147.0	11.16	11.16
CTC, Zipformer-XL	286.6	10.8	10.87
CTC/AED, Zipformer-S	46.3	11.4	11.39
CTC/AED, Zipformer-M	90.0	10.57	10.61
CTC/AED, Zipformer-L	174.3	10.26	10.38
CTC/AED, Zipformer-XL	315.5	10.22	10.33
Pruned transducer, Zipformer-S	23.3	10.98	10.94
Pruned transducer, Zipformer-M	65.6	10.37	10.42
Pruned transducer, Zipformer-L	148.4	10.23	10.28
Pruned transducer, Zipformer-XL	288.2	10.09	10.2
CR-CTC, Zipformer-S (ours)	22.1	11.68	11.58
CR-CTC, Zipformer-M (ours)	64.3	10.62	10.72
CR-CTC, Zipformer-L (ours)	147.0	10.31	10.41
CR-CTC, Zipformer-XL (ours)	286.6	10.15	10.28
CR-CTC/AED, Zipformer-XL (ours)	315.5	**9.92**	10.07
Pruned transducer w/ *CR-CTC*, Zipformer-XL (ours)	286.6	9.95	**10.03**

Thanks!
Q & A

- Paper: https://arxiv.org/pdf/2410.05101
- Code: https://github.com/k2-fsa/icefall/pull/1766