OWSM-CTC: An Open Encoder-Only Speech Foundation Model for Speech Recognition, Translation, and Language Identification

Yifan Peng[1], Yui Sudo[2], Muhammad Shakeel[2], Shinji Watanabe[1]

[1]Carnegie Mellon University [2]Honda Research Institute Japan

Background

- Speech foundation models perform well in various speech processing tasks with a unified architecture
- They typically use an autoregressive architecture
  - Encoder-decoder
  - Decoder-only
- Limitations of autoregressive models compared to non-autoregressive models
  - Slower inference speed
  - More hallucination

Training Setups

- **Data:** 180k hours of public ASR & ST covering 151 languages (same as previous OWSM)
- **Architecture:** 27-layer E-Branchformer encoder with 1B parameters
- **Cost:** 19k hours on 64 NVIDIA A100 (40GB) GPUs
- **Toolkit:** ESPnet based on PyTorch

Results: LID

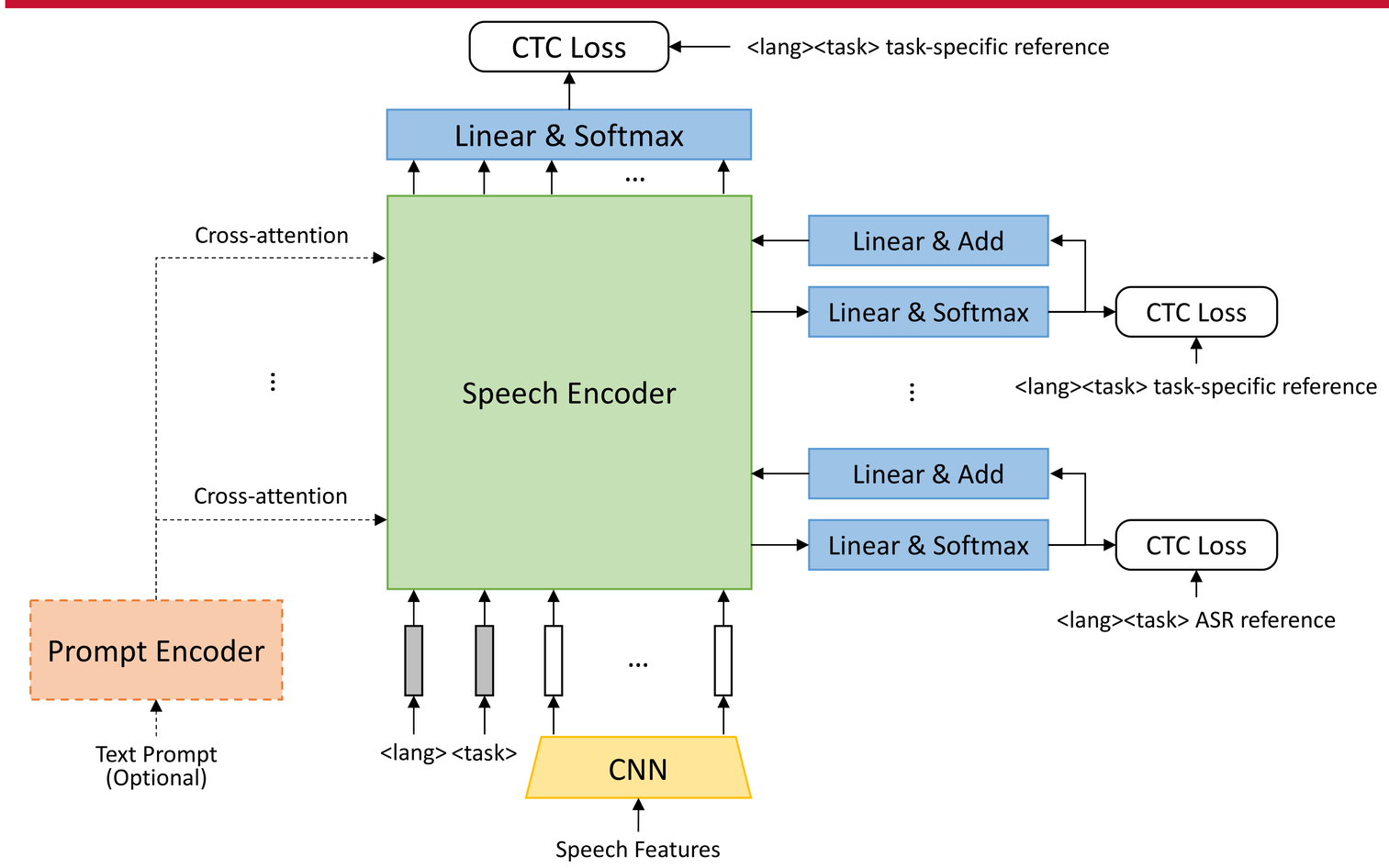
- OWSM-CTC significantly outperforms previous encoder-decoder models on FLEURS

	Accuracy % (↑)
Whisper (encoder-decoder) (Radford et al., 2023)	
base	47.6
small	53.1
medium	54.8
OWSM v3 (encoder-decoder) (Peng et al., 2023e)	
medium	81.4
OWSM v3.1 (encoder-decoder) (Peng et al., 2024)	
base	41.9
medium	75.6
OWSM-CTC (ours)	
medium	**87.6**

OWSM-CTC

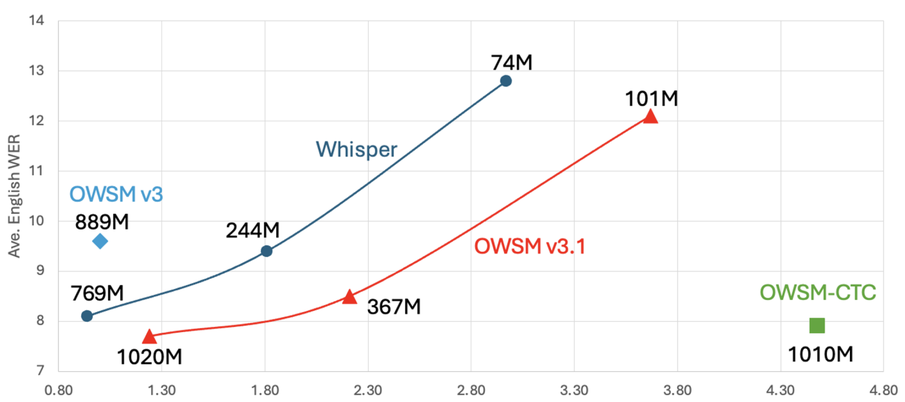
- Open Whisper-style Speech Models (OWSM) reproduce Whisper (Radford et al., 2023) using publicly available data and open-source toolkits (Peng et al., 2023, 2024)
  - Previous OWSM models use an encoder-decoder architecture
- OWSM-CTC is a CTC-based non-autoregressive model that mimics the functionalities of Whisper and OWSM
  - Multilingual automatic speech recognition (ASR)
  - Any-to-any speech translation (ST)
  - Spoken language identification (LID)
- Advantages of OWSM-CTC compared to encoder-decoder OWSM
  - Comparable or superior performance on various benchmarks
  - 3x to 4x inference speed-up on segmented speech; 20x speed-up on long speech
  - Less hallucination

Model Architecture

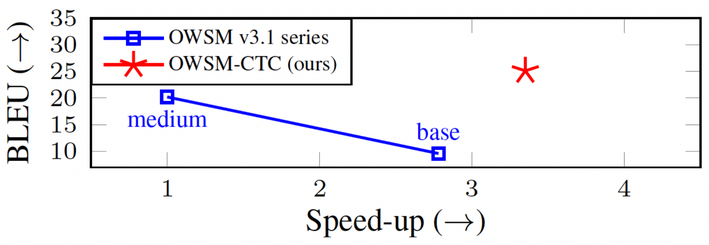


Results: ASR & ST

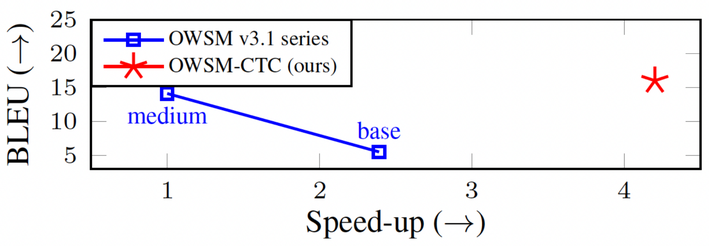
- OWSM-CTC achieves competitive or better performance while being 3x to 4x faster at inference time
  - English ASR: Word Error Rate averaged over 9 test sets



- ST: BLEU scores averaged over various directions on CoVoST-2



(b) X-to-En speech translation



(c) En-to-X speech translation

Results: Long-Form ASR

- Audios ranging from 6 to 27 minutes
- First, an unsegmented long recording is split into overlapped chunks of 30s
  - Context length: the duration of overlapped regions
- Then, CTC greedy decoding is performed on batched chunks
- CTC-based non-autoregressive decoding is faster and more robust for long-form ASR

	Context Length	WER % (↓)	Speed-up (↑)
Whisper (encoder-decoder) (Radford et al., 2023)			
base	-	5.3	1.40x
small	-	4.4	1.62x
medium	-	**3.8**	0.86x
OWSM v3.1 (encoder-decoder) (Peng et al., 2024)			
base	-	9.6	1.40x
medium	-	5.7	1.00x
OWSM-CTC (ours)			
medium	2s	5.4	**22.40x**
	4s	5.2	19.35x
	6s	5.2	16.07x
	8s	5.2	12.09x

Results: Robustness

- A quantitative measure of "hallucination"
  - Autoregressive decoding might fall into repetitions of a few characters or words
  - A hypothesis is considered as a failure if it contains any character-level θ-gram ($\theta = 1, \ldots, \theta_{\max}$) that consecutively occurs for at least δ times
- There are 286k ST samples in total
- OWSM-CTC is much more robust to this type of errors
- OWSM-CTC hallucinates less for noise input

$\theta_{\max}$	δ	Model	#Failures (↓)
10	5	OWSM v3.1	2448
		OWSM v3.1 (beam 5)	630
		OWSM-CTC (ours)	3

Input length	5s	10s	20s
Whisper (encoder-decoder) (Radford et al., 2023)			
large-v3	Fjell	Fusilet	Rekordverk
OWSM v3.1 (encoder-decoder) (Peng et al., 2024)			
medium	thank you	thank you	(Applause)
OWSM-CTC (ours)			
medium	.	(	( )

References

- Radford, Alec, et al. "Robust speech recognition via large-scale weak supervision." in *Proc. ICML*, 2023.
- Peng, Yifan, et al. "Reproducing whisper-style training using an open-source toolkit and publicly available data." in *Proc. ASRU*, 2023.
- Peng, Yifan, et al. "OWSM v3.1: Better and Faster Open Whisper-Style Speech Models based on E-Branchformer." in *Proc. INTERSPEECH*, 2024.