OWSM v3.1:
Better and Faster Open Whisper-Style Speech Models based on E-Branchformer

Yifan Peng[1], Jinchuan Tian[1], William Chen[1], Siddhant Arora[1], Brian Yan[1], Yui Sudo[2], Muhammad Shakeel[2], Kwanghee Choi[1], Jiatong Shi[1], Xuankai Chang[1], Jee-weon Jung[1], Shinji Watanabe[1]

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

Introduction

- OpenAI Whisper (Radford et al., 2023) has strong performance but does not release its complete training pipeline
- Open Whisper-style Speech Model (OWSM) (Peng et al., 2023) is an initial step towards reproducing Whisper using publicly available data and open-source toolkits
- This work proposes OWSM v3.1, the latest version of OWSM, which improves the performance and efficiency of previous versions with the same amount of training data
  - OWSM v3.1 uses E-Branchformer as encoder
  - Three sizes: base (101M), small (367M), medium (1.02B)
- OWSM v3.1 outperforms previous OWSM in the following test sets while being 16% to 25% faster:
  - 8 of 9 English ASR
  - 10 of 11 multilingual ASR
  - 13 of 19 ST
  - 3 of 4 SLUE-PERB
- OWSM v3.1 shows emergent abilities in zero-shot contextual biasing

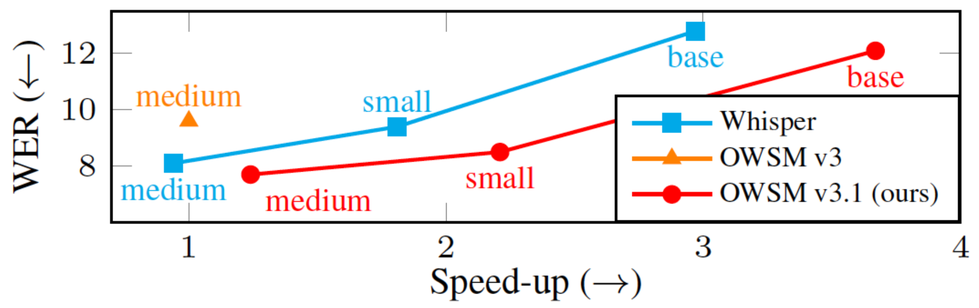
Experimental Setups

- OWSM v3.1 architecture
  - Encoder: E-Branchformer (OWSM v3 uses vanilla Transformer)
  - Decoder: Transformer
- Training data amount is the same as OWSM v3 (no new data)

Model	Model size	Training data (h)	GPU hours
Whisper base	74M	680k	Unknown
Whisper small	244M		
Whisper medium	769M		
OWSM v3	889M	180k	30.7k
OWSM v3.1 base	101M		2.3k
OWSM v3.1 small	367M		3.2k
OWSM v3.1 medium	1020M		24.6k
OWSM v3.1 low-restriction	367M	70k	3.2k

Results: Speech Recognition (WER)

English ASR



Multilingual ASR

Test set	Language	Metric	Whisper data	Whisper base	Whisper small	Whisper medium	OWSM v3 data	OWSM v3 medium	OWSM v3.1 (ours) base	OWSM v3.1 (ours) small	OWSM v3.1 (ours) medium
MLS [24]	Spanish	WER	11.1K	14.5	9.1	**6.1**	2.0K	11.7	18.5	10.8	9.0
	French		9.8K	25.2	13.6	**9.7**	2.5K	14.1	24.2	14.1	12.1
	German		13.3K	19.9	11.5	**8.1**	3.7K	11.9	18.7	12.4	10.8
	Dutch		2.1K	30.9	18.2	**12.2**	1.7K	17.7	28.6	19.7	18.1
	Italian		2.6K	32.9	21.3	**15.6**	0.7K	24.5	33.7	21.8	20.2
	Portuguese		8.6K	23.5	13.8	**8.9**	0.3K	28.2	44.9	26.7	21.6
	Polish		4.3K	25.2	12.5	**6.8**	0.3K	37.0	49.7	28.5	25.2
AISHELL-1 [32]	Chinese	CER	23.4K	39.1	25.1	15.7	16.0K	7.1	12.2	7.5	**6.4**
KsponSpeech clean [22]	Korean		8.0K	27.0	24.0	17.6	1.0K	20.5	23.8	17.2	**16.7**
KsponSpeech other [22]				22.9	15.4	**12.8**		22.6	26.1	18.9	18.9
ReazonSpeech [33]	Japanese		7.1K	54.1	32.5	25.3	18.9K	11.3	11.2	8.5	**7.9**
Average WER/CER (↓)			-	28.7	17.9	**12.6**	-	18.8	26.5	16.9	15.2

Long-form ASR

Whisper base	Whisper small	Whisper medium	OWSM v3 medium	OWSM v3.1 (ours) base	OWSM v3.1 (ours) small	OWSM v3.1 (ours) medium
5.3	4.4	**3.8**	9.2	9.6	6.7	5.7

Results: Spoken Language Understanding

The pre-trained speech encoder is frozen, and a randomly initialized shallow decoder is trained on task-specific SLU data from SLUE-PERB

Task	Metric	OWSM v3	OWSM v3.1 (ours)
Sentiment Analysis	F1 score	**60.1**	56.2
Named Entity Recognition	F1 score	54.8	**65.8**
Named Entity Localization	frame-F1	40.5	**50.4**
Dialogue Act Classification	F1 score	56.5	**64.8**

Results: Speech Translation (BLEU)

X-to-English

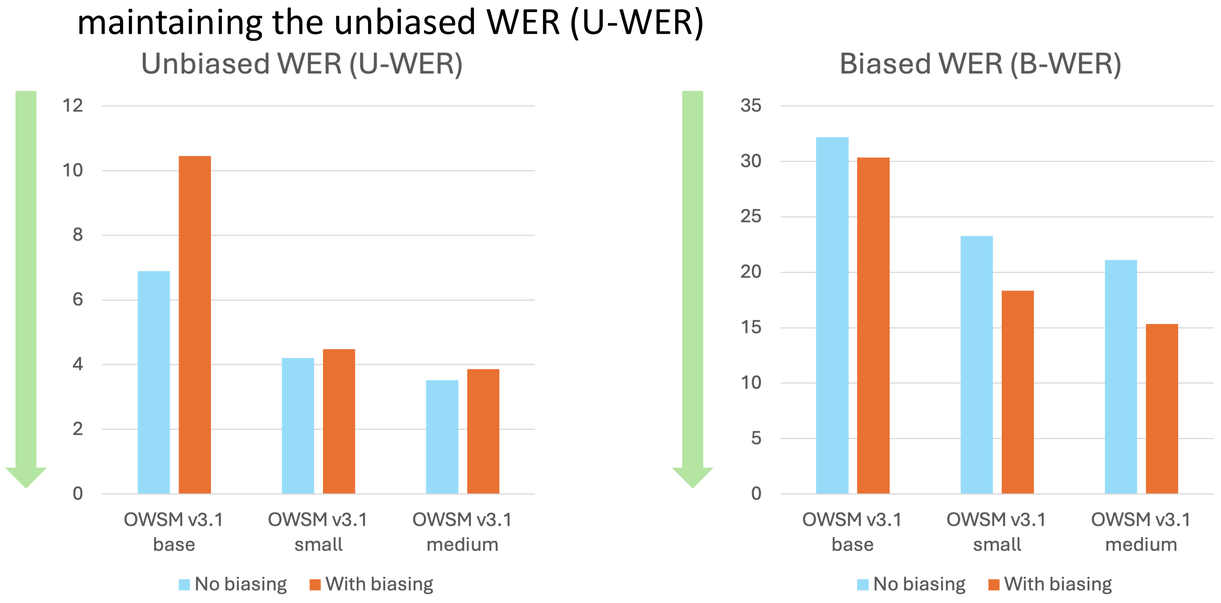
Source	Whisper data	Whisper base	Whisper small	Whisper medium	OWSM v3 data	OWSM v3 medium	OWSM v3.1 (ours) base	OWSM v3.1 (ours) small	OWSM v3.1 (ours) medium
German	4.3K	11.4	25.0	**33.6**	0.2K	16.2	7.3	15.1	17.1
Spanish	6.7K	19.2	32.8	**39.7**	0.1K	20.5	10.0	19.3	22.3
French	4.5K	13.1	26.4	**34.4**	0.3K	21.7	11.1	20.3	22.7
Catalan	0.2K	9.7	21.7	**29.2**	0.1K	16.8	9.0	16.2	18.4
Ave. BLEU (↑)		13.4	26.5	**34.2**	-	18.8	9.4	17.7	20.1
Speed-up (↑)		2.14x	1.80x	0.98x	-	1.00x	**3.23x**	2.26x	1.16x

English-to-X

Target	Training Data (h)	OWSM v3 medium	OWSM v3.1 (ours) base	OWSM v3.1 (ours) small	OWSM v3.1 (ours) medium
German	14.0K	**25.4**	14.6	22.8	**25.4**
Catalan	0.4K	**20.0**	7.7	15.9	19.6
Chinese	13.7K	**33.4**	14.5	26.7	32.1
Persian	0.8K	9.5	3.0	7.7	**10.1**
Estonian	0.4K	**7.8**	1.8	5.8	7.7
Mongolian	0.4K	3.1	1.0	3.3	**4.6**
Turkish	0.9K	6.1	1.2	4.8	**6.5**
Arabic	0.9K	6.6	1.6	5.1	**7.2**
Swedish	0.4K	19.9	8.1	16.6	**20.3**
Latvian	0.4K	6.3	1.3	4.4	**6.4**
Slovenian	0.4K	8.6	0.7	5.7	**9.0**
Tamil	0.4K	0.0	0.0	0.0	0.0
Japanese	1.0K	17.3	8.7	16.4	**19.6**
Indonesian	0.4K	14.5	5.1	12.4	**16.1**
Welsh	0.4K	**15.9**	4.5	11.6	15.3
Ave. BLEU (↑)		13.0	4.9	10.6	**13.3**
Speed-up (↑)		1.00x	**3.00x**	2.43x	1.25x

Emergent Abilities in Zero-Shot Contextual Biasing ASR

- Similarly to Whisper, OWSM optionally takes a text prompt as input, which is a condition for generation
  - During training, the previous sentence is used with probability 0.5
  - During inference, the user can provide a prompt to bias the output
- We evaluate OWSM v3.1 on LibriSpeech biasing test sets (Le et al., 2021)
  - Contextual biasing aims to reduce the biased WER (B-WER) while maintaining the unbiased WER (U-WER)



The phenomenon that the smaller OWSM performs poorly in zero-shot biasing ASR while larger ones perform well reveals that speech foundation models also have the emergent ability

References

1. Radford, Alec, et al., "Robust speech recognition via large-scale weak supervision," in Proc. ICML, 2023.
2. Peng, Yifan, et al., "Reproducing Whisper-Style Training Using An Open-Source Toolkit and Publicly Available Data," in Proc. ASRU, 2023.
3. Le, Duc, et al., "Contextualized streaming end-to-end speech recognition with trie-based deep biasing and shallow fusion," in Proc. Interspeech, 2021.