Interspeech 2025
17-21 August 2025, Rotterdam, The Netherlands

OWSM v4: Improving Open Whisper-Style Speech Models via Data Scaling and Cleaning

Yifan Peng[1], *Muhammad Shakeel*[2], *Yui Sudo*[2], *William Chen*[1], *Jinchuan Tian*[1], *Chyi-Jiunn Lin*[1], *Shinji Watanabe*[1]

[1]Carnegie Mellon University, United States
[2]Honda Research Institute Japan, Japan

pengyf21@gmail.com, shinjiw@ieee.org

Abstract

The Open Whisper-style Speech Models (OWSM) project has developed a series of fully open speech foundation models using academic-scale resources, but their training data remains insufficient. This work enhances OWSM by integrating YODAS, a large-scale web-crawled dataset with a Creative Commons license. However, incorporating YODAS is nontrivial due to its wild nature, which introduces challenges such as incorrect language labels and audio-text misalignments. To address this, we develop a scalable data-cleaning pipeline using public toolkits, yielding a dataset with 166,000 hours of speech across 75 languages. Our new series of OWSM v4 models, trained on this curated dataset alongside existing OWSM data, significantly outperform previous versions on multilingual benchmarks. Our models even match or surpass frontier industrial models like Whisper and MMS in multiple scenarios. We will publicly release the cleaned YODAS data, pre-trained models, and all associated scripts via the ESPnet toolkit.[1]

Index Terms: speech foundation models, data cleaning, open whisper-style speech models, speech recognition

1. Introduction

Speech foundation models (SFMs), typically trained on large amounts of data, have demonstrated state-of-the-art (SOTA) performance in various speech processing tasks [1–4]. A notable example is OpenAI's Whisper [1], which is trained on 680 thousand to 5 million hours of audio data and supports multilingual automatic speech recognition (ASR), any-to-English speech translation (ST), spoken language identification (LID), and voice activity detection (VAD). However, Whisper does not publicly release its training data, code, and logs, leading to concerns about privacy, transparency, and reproducibility. To advance open research, researchers from academic institutions have developed a series of fully open Whisper-style speech models (OWSM) [5] using publicly available data and an open source toolkit, ESPnet [6]. Initial OWSM v1, v2, and v3 [5] establish a reproducible pipeline for Whisper-style training, but their performance is limited.

Recent studies have enhanced the effectiveness and efficiency of SFMs. One approach is improving model architectures. Conformer [7], Branchformer [8], and Zipformer [9] consistently outperform Transformer [10] for speech modeling. Squeezeformer [11], FastConformer [12], and SummaryMixing [13] significantly reduce the training and inference cost. OWSM v3.1 [14] adopts E-Branchformer [15, 16] and achieves significant improvements over OWSM v3. OWSM-CTC [17] proposes a novel non-autoregressive architecture based on hierarchical self-conditioned Connectionist Temporal Classification (CTC) [18], unifying ASR, ST, and LID in a shared encoder-only model. Compared to attention-based encoder-decoder (AED) models, OWSM-CTC improves the inference speed and reduces hallucinations.

Table 1: *ASR error rates (%) on FLEURS. Our OWSM-CTC v4 outperforms v3.1 across all 102 languages and surpasses v3.2 in 100 languages. Here we only show languages where OWSM-CTC v4 achieves error rates below 20%.* ***Bold****: The best result in each row.* *Blue* *: Our v4 model surpasses previous OWSM.*

Lang.	Metric ↓	MMS	OWSM-CTC		
		1B-all	v3.1	v3.2	v4 (ours)
spa	WER	6.60	11.30	9.58	**5.44**
ita	WER	**5.85**	13.38	11.25	5.91
eng	WER	12.26	8.24	7.06	**6.37**
jpn	CER	20.92	7.56	6.51	**6.43**
kor	CER	18.30	20.09	17.72	**6.74**
por	WER	8.97	19.66	16.22	**7.38**
cat	WER	10.75	9.37	8.16	**7.70**
deu	WER	10.45	14.97	13.17	**8.36**
fra	WER	12.53	17.13	14.79	**9.67**
ind	WER	13.19	39.76	33.56	**10.24**
zho	CER	26.46	13.47	12.25	**10.95**
rus	WER	19.79	18.18	15.68	**10.96**
tha	CER	**10.69**	28.29	24.06	12.35
vie	WER	29.96	71.36	65.15	**13.34**
nld	WER	**12.35**	30.46	25.10	14.73
bel	WER	**14.84**	18.99	16.38	15.09
tur	WER	19.55	57.43	48.56	**15.79**
ben	WER	**13.19**	18.63	16.33	15.80
hin	WER	**10.82**	35.63	30.91	16.40
glg	WER	**10.59**	30.06	24.42	17.41
ukr	WER	**18.09**	47.53	40.97	18.39

Another line of research improves training data. Unsupervised data selection is proposed to enhance ASR systems [19, 20]. Data cleaning techniques are widely used when creating ASR datasets [21–24]. Inspired by this, Tian *et al.* filter OWSM v3.1 training data based on ASR error rates and restore punctuation and capitalization using large language models. Compared to OWSM v3.1, the resultant model, OWSM v3.2 [25], achieves comparable ASR results and slightly better ST results, despite being trained on 15% less data. However, Tian *et al.* only consider the original v3.1 data that generally have good quality but do not include new data from other public sources. Hence, the performance gain of data filtering is marginal and inconsistent.

Inspired by the findings that scaling training data improves multilingual ASR systems [1, 2, 26], we propose to enhance OWSM by integrating high-quality data from YODAS [24] using academic-scale resources. YODAS is distinctive from other popular datasets such as MSR-86K [27], LibriHeavy [28], GigaSpeech [22, 29], and MOSEL [30] in the following aspects: (1) YODAS publicly releases audio files in a Creative Commons license instead of links to original sources, simplifying

[1]https://www.wavlab.org/activities/2024/owsm/

10.21437/Interspeech.2025-1062

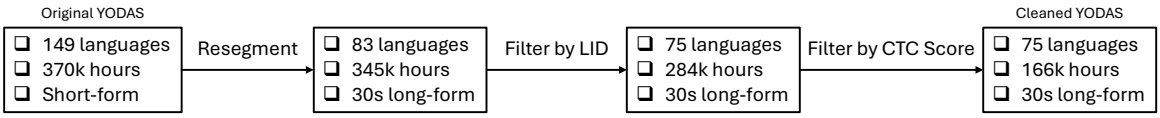

Figure 1: *Our data-cleaning pipeline consists of three steps: (1) realign audio and text using a pre-trained OWSM-CTC model, (2) filter data based on LID results, and (3) filter data based on CTC confidence scores.*

Table 2: *ASR WERs (%) of OWSM v3.1 small fine-tuned on the cleaned YODAS filtered at varying thresholds θ_{CTC}. LF: Long-form web presentations.*

θ_{CTC}	LF	Common Voice										
	eng	deu	eng	fra	ind	ita	nld	por	rus	spa	tur	vie
0.00	5.0	85.7	100+	100+	100+	100+	100+	100+	65.5	100+	100+	100+
0.10	**4.3**	18.7	24.3	24.5	36.4	**18.7**	22.9	37.9	**18.5**	**15.7**	**50.2**	54.6
0.15	4.4	18.2	**24.1**	24.1	35.7	18.8	**22.2**	37.9	19.6	15.8	52.8	59.9
0.20	4.4	18.0	24.2	24.0	34.4	21.2	22.4	38.0	20.6	18.4	52.8	47.3
0.30	4.6	**17.4**	25.0	**22.9**	**31.9**	20.1	24.8	**37.1**	22.4	17.6	51.1	**47.0**

data downloading and providing static sources for redistribution. (2) YODAS establishes a scalable pipeline to crawl data from the web. The current version already includes 370k hours of audio in 149 languages. Future versions can further grow. (3) YODAS covers diverse speaking styles and acoustic environments. It also releases unsegmented long-form audio recordings, which are suitable for Whisper-style training. However, simply adding more data without careful curation can degrade performance due to noisy annotations in the raw data. Hence, data cleaning is crucial for ensuring good quality.

Our contributions are summarized below.

- We propose a scalable data-cleaning pipeline using public LID and ASR models. By applying it to YODAS, we create an ASR dataset with 166k hours of audio in 75 languages.
- We develop a new series of OWSM v4 models using academic-scale resources, comprising three AED models of varying sizes and one CTC model, trained on the cleaned YODAS dataset in conjunction with previous OWSM data (320k hours in total). The new models consistently and significantly outperform previous OWSM versions in multilingual ASR and LID (see Table 1 for example). Furthermore, they achieve competitive results compared to SOTA industrial models on multiple benchmarks.
- To advance academic research, we will publicly release our data-cleaning pipeline, the cleaned YODAS data, training code, pre-trained model weights, and training logs.

2. Proposed Method

2.1. YODAS data cleaning

The raw YODAS data has not undergone a rigorous cleaning process and may contain annotation errors [24]. Common issues include mismatched language labels and misalignment between audio and text. Therefore, data cleaning is essential to ensure accuracy and reliability. Figure 1 illustrates our data-cleaning pipeline consisting of the following three steps. Our scripts will be publicly released, including more implementation details.

2.1.1. Resegmentation

YODAS provides unsegmented long-form recordings, each of them is accompanied by a list of text transcriptions annotated with start and end timestamps. However, these timestamps can be inaccurate. Consequently, our first step is to realign the audio and text using the CTC segmentation algorithm [31]. For this purpose, we employ the publicly available OWSM-CTC v3.2 model[2], which supports only a subset of the languages present in YODAS. Following realignment, the long-form audio recordings are segmented into shorter utterances, each with a maximum duration of 30 seconds. Utterances that consist exclusively of non-speech elements, such as music, are removed. The processed dataset comprises 345k hours of audio across 83 languages. Additionally, after CTC segmentation, each short utterance is assigned a confidence score, which quantifies the alignment quality between the audio and the corresponding text. This confidence score is subsequently utilized to filter low-quality data, as discussed in Section 2.1.3.

Table 3: *Durations (in k hours) for the top 10 languages in the cleaned YODAS filtered at varying thresholds θ_{CTC}.*

θ_{CTC}	Total	eng	spa	rus	por	kor	fra	deu	ita	vie	ind
0.00	283.6	129.0	30.7	25.8	18.2	17.5	15.5	12.2	9.3	7.8	5.2
0.10	166.4	74.6	17.3	15.7	10.8	10.9	8.6	7.1	5.6	4.6	3.3
0.15	118.5	51.5	12.2	11.1	8.1	8.3	6.2	5.0	4.1	3.5	2.5
0.20	84.8	35.7	8.6	7.8	6.1	6.4	4.5	3.6	3.0	2.6	1.9
0.30	43.0	17.0	4.2	3.7	3.4	3.7	2.3	2.0	1.5	1.5	1.1

2.1.2. LID-based filtering

We observe that certain utterances have incorrect language labels. To address this issue, we perform LID on both the audio and text using public models. Specifically, the text-based LID model[3] is sourced from fastText [32, 33], while the spoken LID model is based on ECAPA-TDNN[4], developed by SpeechBrain [34]. We retain only those utterances for which the original language label matches both the predicted language from the text and the predicted language from the audio. Applying this filtering step results in a dataset comprising 284k hours of audio across 75 languages, as shown in Figure 1.

2.1.3. CTC-score-based filtering

The final step removes utterances with low-quality audio-text alignments, as indicated by the CTC score calculated in Section 2.1.1. The CTC confidence score is language-dependent; therefore, we rank the scores of short utterances within each language and select a relative threshold (quantile) θ_{CTC}. For each long-form utterance, if any of its constituent short utterances fall within the lowest θ_{CTC} quantile, the entire utterance will be discarded. Different threshold values yield varying amounts of retained data. To identify a suitable threshold, we fine-tune a pre-trained small-sized OWSM v3.1 (367M) [14] on the cleaned YODAS data filtered at different thresholds. We then evaluate them on Common Voice [35] for short-form ASR and a web presentation corpus for long-form ASR, as shown in Table 2.

When $\theta_{\text{CTC}} = 0.00$, no filtering is applied, and all 284k hours of audio after LID filtering are used for fine-tuning. However, the performance on Common Voice is poor and unstable. The decoding often gets stuck in repetitions of a few tokens, leading to word error rates (WER) exceeding 100%. This obser-

[2]https://huggingface.co/espnet/owsm_ctc_v3.2_ft_1B
[3]https://fasttext.cc/docs/en/language-identification.html
[4]https://huggingface.co/speechbrain/lang-id-voxlingua107-ecapa

Table 4: *Configurations. Models are categorized into three types based on their level of openness, including the availability of pre-trained weights, data details, and training code & logs.*

Model Name	Openness			Model	Data (h)		GPU	# of
	Weights	Data	Logs	Size	ASR	ST	Hrs	Lang.
Open-weight models								
Whisper base [1]	✓	✗	✗	0.07B	555k	125k	unk.	99
Whisper small [1]	✓	✗	✗	0.24B	555k	125k	unk.	99
Whisper medium [1]	✓	✗	✗	0.77B	555k	125k	unk.	99
Whisper large v3 [1]	✓	✗	✗	1.55B	5M		unk.	100
Parakeet-CTC [12]	✓	✗	✗	1.06B	64k	-	unk.	1
Canary [4]	✓	✗	✗	1.02B	82k	66k	6.1k*	4
Open-weight, open-data models								
MMS-fl102 [3]	✓	✓§	✗	0.97B	1.4k¶	-	unk.	102
MMS-all [3]	✓	✓§	✗	0.97B	107k¶	-	unk.	1162
Fully-open models								
AED models								
OWSM v3.1 base [14]	✓	✓	✓	0.10B	140k	40k	2.3k	151
OWSM v3.1 small [14]	✓	✓	✓	0.37B	140k	40k	3.2k	151
OWSM v3.1 medium [14]	✓	✓	✓	1.02B	140k	40k	24.6k	151
OWSM v4 base (ours)	✓	✓	✓	0.10B	290k	30k	1.0k†	151
OWSM v4 small (ours)	✓	✓	✓	0.37B	290k	30k	1.7k†	151
OWSM v4 medium (ours)	✓	✓	✓	1.02B	290k	30k	3.8k†	151
CTC models								
OWSM-CTC v3.1 [17]	✓	✓	✓	1.01B	140k	40k	19.2k	151
OWSM-CTC v3.2‡	✓	✓	✓	1.01B	124k	30k	28.8k	151
OWSM-CTC v4 (ours)	✓	✓	✓	1.01B	290k	30k	4.1k†	151

* Trained on NVIDIA A100 (80GB). Not including encoder pre-training time.
§ Not publicly released, but provides detailed statistics and links to sources.
¶ The 491k hours of unlabeled speech for pre-training are not included here.
† Trained on NVIDIA H100 (96GB). Previous OWSM used A100 (40GB).
‡ Trained on v3.1 data and then fine-tuned on v3.2 [25], a subset of v3.1.

vation confirms the presence of substantial misalignment issues within the raw YODAS data.

Conversely, applying CTC-score-based filtering ($\theta_{\text{CTC}} > 0$) yields significant improvements, demonstrating the effectiveness of data cleaning. Performance trends vary across different test sets. In some cases, increased data removal leads to better performance, while in others, the opposite trend is observed. Although finer-grained filtering could potentially optimize performance for individual languages, we opt for a threshold of $\theta_{\text{CTC}} = 0.10$. This value retains the majority of the data while providing generally good performance across languages. This filtering process results in 166k hours of audio spanning 75 languages, as illustrated in the final panel of Figure 1. The durations of the top 10 languages are presented in Table 3. Similar to the raw YODAS data, the distribution across languages is highly imbalanced. English constitutes the largest share, whereas many other languages continue to be underrepresented. For simplicity, in this work, we keep the original distribution without any resampling.

2.2. OWSM v4 series

To further assess the quality of our cleaned YODAS data, we train a new series of OWSM v4 models using this curated data alongside the previous OWSM v3.2 data [25]. This series includes three AED-based models ranging from 100M to 1B parameters, as well as a CTC-based model with 1B parameters. Table 4 summarizes the model and training configurations. Our v4 models employ the same configurations as the previous v3.1 [14, 17], except that the number of Mel filterbanks is increased from 80 to 128, following Whisper-large-v3. The speech features are subsampled by eight times, resulting in a time shift of 80 ms. The speech encoder is E-Branchformer [15], and the decoder, if exists, is Transformer [10]. We implement models in ESPnet [6] based on PyTorch [36]. FlashAttention-2 [37] is used for better efficiency. We use the AdamW optimizer [38] with a batch size of 320. We train all models for 700k steps, i.e., around three epochs.

Table 5: *LID accuracy (%) on FLERUS and long-form English ASR WER (%) on a web presentation corpus.*

Model	LID Acc. ↑	Long-Form WER ↓
Open-weight models		
Whisper-medium	54.8*	3.8
Whisper-large-v3	58.9*	3.4
Open-weight, open-data models		
MMS-lid-4017	93.3	-
Fully-open models		
AED models		
OWSM v3.1 base	41.9	9.6
OWSM v3.1 small	67.1	6.7
OWSM v3.1 medium	75.6	5.7
OWSM v4 base (ours)	80.1	5.5
OWSM v4 small (ours)	90.0	4.6
OWSM v4 medium (ours)	**95.6**	4.3
+ beam size 5	-	3.6
CTC models		
OWSM-CTC v3.1	87.6	5.2
OWSM-CTC v3.2	91.1	4.8
OWSM-CTC v4 (ours)	93.6	**3.3**

* Whisper supports only a subset of languages in FLEURS.

Table 6: *Multilingual ASR WERs (%) on MLS. The inference speed is based on the total decoding time on an NVIDIA H100.*

Model	eng	spa	fra	deu	nld	ita	por	pol	Ave. ↓	Speed ↑
Open-weight models										
Whisper-base	13.4	14.5	25.2	19.9	30.9	32.9	23.5	25.2	23.2	3.9×
Whisper-small	9.1	9.1	13.6	11.5	18.2	21.3	13.8	12.5	13.6	2.3×
Whisper-medium	10.2	6.1	9.7	8.1	12.2	15.6	8.9	6.8	9.7	1.2×
Whisper-large-v3	5.1	4.1	4.8	5.6	10.2	9.2	7.4	4.4	6.4	1.0×
Open-weight, open-data models										
MMS-fl102	23.6	14.9	22.4	14.7	16.4	18.9	17.1	12.7	17.6	20.9×
MMS-all	10.7	5.8	8.8	8.8	12.8	11.0	16.2	10.5	10.6	21.4×
Fully-open models										
OWSM-AED models										
v3.1 base	12.0	18.5	24.2	18.7	28.6	33.7	44.9	49.7	28.8	3.0×
v3.1 small	8.1	10.8	14.1	12.4	19.7	21.8	26.7	28.5	17.8	2.2×
v3.1 medium	7.1	9.0	12.1	10.8	18.1	20.2	21.6	25.2	15.5	1.2×
v4 base (ours)	11.6	11.6	17.6	15.9	23.1	23.3	18.9	31.5	19.2	3.0×
v4 small (ours)	7.6	7.1	10.2	10.3	15.7	15.7	11.5	16.0	11.8	2.2×
v4 medium (ours)	6.4	5.7	7.8	8.2	13.4	13.1	9.0	11.5	9.4	1.1×
+ beam size 5	5.9	5.5	7.3	7.9	12.9	12.8	8.5	11.0	9.0	0.2×
OWSM-CTC models										
v3.1	7.3	10.3	12.9	11.9	20.4	22.1	23.5	31.6	17.5	26.3×
v3.2	7.0	9.7	11.3	11.4	17.6	20.0	20.5	24.5	15.3	23.7×
v4 (ours)	6.4	5.8	7.8	9.5	15.1	15.5	10.3	15.1	10.7	25.1×

3. Experimental Results

We evaluate our OWSM v4 models on multilingual ASR, LID, and ST benchmarks using greedy decoding unless otherwise specified. While we include results from models developed by well-resourced industry entities such as OpenAI's Whisper and Meta's MMS, our primary comparisons are against baselines from academic institutions, given our constrained resources.

3.1. Results of language identification

Table 5 presents the LID results on FLEURS [40], where our OWSM v4 series outperforms earlier versions. Compared to industrial-scale models, OWSM v4 medium and OWSM-CTC v4 both achieve higher accuracies than Whisper and MMS-lid, with OWSM v4 medium reaching the highest accuracy of 95.6%. These results indicate that our cleaned YODAS data contains high-quality language labels, attributed to the LID filtering stage (see Section 2.1.2).

Table 7: *English ASR WERs (%) on the Hugging Face Open ASR Leaderboard. The Inverse Real Time Factor (RTFx) is measured using an NVIDIA H100 GPU (96GB). Underlined: Our v4 model outperforms previous OWSM.*

Model	Arch.	Size	Ave. WER ↓	RTFx ↑	AMI	Earnings22	Gigaspeech	LS-Clean	LS-Other	SPGISpeech	Web-Presentation	Voxpopuli
Open-weight models												
Whisper-medium-en	AED	0.8B	8.06	289.64	16.66	12.42	11.11	2.89	5.85	3.36	4.13	8.04
Whisper-large-v3	AED	1.6B	7.47	235.61	16.00	**11.39**	**10.10**	2.01	3.92	2.95	3.84	9.51
Canary	AED	1.0B	**6.48**	287.62	13.66	12.19	10.12	**1.47**	**2.96**	**2.06**	3.59	**5.81**
Parakeet-CTC	CTC	1.1B	7.40	**3007.88**	15.66	13.77	10.28	1.86	3.50	4.02	**3.54**	6.55
Open-weight, open-data models												
MMS-fl102	CTC	1.0B	39.90	1066.12	86.80	51.74	42.44	22.13	28.76	26.21	32.35	28.80
MMS-all	CTC	1.0B	22.65	1055.91	42.02	31.19	26.44	12.64	15.98	16.95	17.49	18.50
Fully-open models												
OWSM-CTC v3.1	CTC	1.0B	8.12	853.44	15.66	13.73	11.89	2.36	5.12	2.87	4.97	8.36
OWSM-CTC v3.2	CTC	1.0B	8.24	841.12	16.71	13.50	11.78	2.61	5.32	2.73	5.35	7.95
OWSM-CTC v4 (ours)	CTC	1.0B	7.44	791.18	**13.09**	13.89	10.83	2.56	4.86	2.56	4.40	7.34

Table 8: *BLEU scores (%) for ST on CoVoST-2 [39]. We do not add any new ST data; OWSM-CTC v4 uses the same ST data as v3.2.*

Model	X-En Translation					En-X Translation														
	de	es	fr	ca	**Average**	de	ca	zh	fa	et	mn	tr	ar	sv	lv	sl	ta	ja	id	cy	**Average**
OWSM v3.1	16.7	22.3	22.8	18.8	20.2	26.3	20.4	29.7	10.2	9.6	5.8	7.8	7.2	20.8	8.4	11.0	0.1	**21.1**	17.2	16.3	14.1
OWSM-CTC v3.1	20.7	27.9	27.5	24.2	25.1	26.7	24.0	32.9	9.9	11.4	6.2	7.9	8.3	24.5	10.0	14.2	0.1	20.4	22.6	20.6	16.0
OWSM-CTC v3.2	21.1	28.6	27.6	24.2	25.4	27.8	**25.2**	33.4	**11.0**	**12.0**	**6.7**	8.9	**9.7**	**26.0**	**11.3**	**15.1**	0.1	20.9	24.0	**21.5**	**16.9**
OWSM-CTC v4 (ours)	**22.1**	**31.8**	**29.5**	**25.7**	**27.3**	**28.5**	24.9	**35.3**	10.6	11.6	6.3	**9.2**	9.2	24.9	10.2	14.3	0.0	20.8	**24.6**	19.8	16.7

3.2. Results of multilingual speech recognition

Table 6 presents ASR results on MLS [41]. Again, our OWSM v4 series achieves much lower WERs than previous OWSM of the same size across all eight languages, highlighting the benefit of data scaling and cleaning. Compared to leading industrial models, OWSM v4 medium achieves a lower average WER than Whisper-medium (9.4% vs. 9.7%) with a similar inference speed. OWSM-CTC v4 achieves a much lower WER than MMS-fl102 (10.7% vs. 17.6%) and a similar WER to MMS-all (10.7% vs. 10.6%), while being 20% faster.

We also evaluate OWSM-CTC on FLEURS [40].[5] OWSM-CTC v4 outperforms v3.1 in all 102 languages and surpasses v3.2 in 100 languages. Table 1 shows 21 languages where OWSM-CTC v4 has error rates below 20%. Among them, OWSM-CTC v4 outperforms MMS-all in 13 languages. These findings further validate the effectiveness of our approach.

3.3. Results of English speech recognition

Table 7 presents English ASR WERs on the Hugging Face Open ASR leaderboard [42].[6] Our OWSM-CTC v4 outperforms previous OWSM-CTC on 6 of 8 test sets. The average WER is improved from 8.12% to 7.44%. Our model also significantly surpasses MMS-fl102 and MMS-all with a similar size. Compared to leading industrial models trained on proprietary data, our model outperforms Whisper-medium while achieving performance on par with Whisper-large-v3 and Parakeet-CTC. Regarding inference speed, our OWSM-CTC v4 is several times faster than AED models such as Whisper and Canary, consistent with the findings in [17].[7]

Table 5 shows long-form English ASR results, where our OWSM v4 models significantly outperform previous OWSM v3.1 and v3.2 of the same size and category (AED or CTC). Notably, OWSM v4 base (100M) already surpasses OWSM v3.1 medium (1B). Compared to frontier industrial models, OWSM-CTC v4 achieves the lowest long-form WER of 3.3%, slightly outperforming Whisper-large-v3, which has 50% more parameters and is trained on 15 times more data. These findings highlight the quality of our curated English data from YODAS and demonstrate the benefit of data scaling.

3.4. Results of speech translation

We do not add any new ST data, using exactly the same ST data as v3.2. Here, our goal is to show that our v4 model maintains similar ST performance. Following [17], we evaluate ST performance on CoVoST-2 X-En and En-X [39]. As shown in Table 8, OWSM-CTC v4 achieves higher BLEU scores than previous OWSM in the four X-En test sets and comparable scores to v3.2 in En-X test sets, verifying that using additional ASR data from YODAS does not negatively impact ST performance.

4. Conclusion

We improve fully open speech-to-text foundation models via data scaling and cleaning using academic-scale resources. We reveal that large-scale web-crawled data contains incorrect language labels and audio-text misalignments. To mitigate these issues, we develop a scalable data-cleaning pipeline using public models and toolkits. Applying it to the raw YODAS ASR dataset, we create a higher-quality subset with 166k hours of speech in 75 languages. Furthermore, we train a new series of OWSM v4 models using this curated dataset alongside existing OWSM data. Extensive evaluations show that our models consistently and significantly outperform previous OWSM models on multilingual benchmarks. Our models even match or surpass leading industrial models such as Whisper and MMS on multiple benchmarks. To advance open academic research, we will publicly release our data-cleaning scripts, the curated YODAS dataset, training code, pre-trained models, and training logs.

5. Acknowledgements

We use PSC Bridges2 and NCSA Delta via ACCESS CIS210014, by National Science Foundation grants #2138259, #2138286, #2138307, #2137603, and #2138296.

[5] The raw transcriptions in FLEURS include words in parentheses, some of which are spoken while others are not. As there is no straightforward rule to exclude non-spoken words, we use the "transcription" field from the Hugging Face dataset as the groundtruth. This may result in higher error rates for certain languages, such as Chinese.

[6] ESPnet does not support batched beam search, leading to very slow inference for AED models. Hence, we only decode CTC-based OWSM.

[7] Unlike NeMo or Hugging Face's transformers, ESPnet lacks lower-level optimization for inference. Nevertheless, our model still achieves competitive inference speed.

6. References

[1] A. Radford, J. W. Kim, T. Xu, G. Brockman, *et al.*, "Robust speech recognition via large-scale weak supervision," in *Proc. ICML*, 2023.

[2] Y. Zhang, W. Han, J. Qin, Y. Wang, *et al.*, "Google USM: Scaling automatic speech recognition beyond 100 languages," *arXiv preprint arXiv:2303.01037*, 2023.

[3] V. Pratap, A. Tjandra, B. Shi, P. Tomasello, *et al.*, "Scaling speech technology to 1,000+ languages," *Journal of Machine Learning Research*, vol. 25, no. 97, pp. 1–52, 2024.

[4] K. C. Puvvada, P. Żelasko, H. Huang, O. Hrinchuk, *et al.*, "Less is more: Accurate speech recognition & translation without web-scale data," in *Proc. Interspeech*, 2024.

[5] Y. Peng, J. Tian, B. Yan, D. Berrebbi, *et al.*, "Reproducing Whisper-Style Training Using an Open-Source Toolkit and Publicly Available Data," in *Proc. ASRU*, 2023.

[6] S. Watanabe, T. Hori, S. Karita, T. Hayashi, *et al.*, "ESPnet: End-to-End Speech Processing Toolkit," in *Proc. Interspeech*, 2018.

[7] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, *et al.*, "Conformer: Convolution-augmented Transformer for Speech Recognition," in *Proc. Interspeech*, 2020.

[8] Y. Peng, S. Dalmia, I. Lane, and S. Watanabe, "Branchformer: Parallel MLP-attention architectures to capture local and global context for speech recognition and understanding," in *Proc. ICML*, 2022.

[9] Z. Yao, L. Guo, X. Yang, W. Kang, *et al.*, "Zipformer: A faster and better encoder for automatic speech recognition," in *Proc. ICLR*, 2023.

[10] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, *et al.*, "Attention is all you need," in *Proc. NeurIPS*, 2017.

[11] S. Kim, A. Gholami, A. E. Shaw, N. Lee, *et al.*, "Squeezeformer: An efficient transformer for automatic speech recognition," in *Proc. NeurIPS*, 2022.

[12] D. Rekesh, N. R. Koluguri, S. Kriman, S. Majumdar, *et al.*, "Fast conformer with linearly scalable attention for efficient speech recognition," in *Proc. ASRU*, 2023.

[13] T. Parcollet, R. van Dalen, S. Zhang, and S. Bhattacharya, "SummaryMixing: A linear-complexity alternative to self-attention for speech recognition and understanding," in *Proc. Interspeech*, 2024.

[14] Y. Peng, J. Tian, W. Chen, S. Arora, *et al.*, "OWSM v3.1: Better and faster open whisper-style speech models based on E-Branchformer," in *Proc. Interspeech*, 2024.

[15] K. Kim, F. Wu, Y. Peng, J. Pan, *et al.*, "E-Branchformer: Branchformer with enhanced merging for speech recognition," in *Proc. SLT*, 2023.

[16] Y. Peng, K. Kim, F. Wu, B. Yan, *et al.*, "A Comparative Study on E-Branchformer vs Conformer in Speech Recognition, Translation, and Understanding Tasks," in *Proc. Interspeech*, 2023.

[17] Y. Peng, Y. Sudo, M. Shakeel, and S. Watanabe, "OWSM-CTC: An open encoder-only speech foundation model for speech recognition, translation, and language identification," in *Proc. ACL*, 2024.

[18] A. Graves, S. Fernández, F. J. Gomez, and J. Schmidhuber, "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks," in *Proc. ICML*, 2006.

[19] Z. Lu, Y. Wang, Y. Zhang, W. Han, *et al.*, "Unsupervised data selection via discrete speech representation for ASR," in *Proc. Interspeech*, 2022.

[20] C. Park, R. Ahmad, and T. Hain, "Unsupervised data selection for speech recognition with contrastive loss ratios," in *Proc. ICASSP*, 2022.

[21] P. K. O'Neill, V. Lavrukhin, S. Majumdar, V. Noroozi, *et al.*, "Spgispeech: 5, 000 hours of transcribed financial audio for fully formatted end-to-end speech recognition," in *Proc. Interspeech*, 2021.

[22] G. Chen, S. Chai, G. Wang, J. Du, *et al.*, "Gigaspeech: An evolving, multi-domain ASR corpus with 10, 000 hours of transcribed audio," in *Proc. Interspeech*, 2021.

[23] D. Galvez, G. Diamos, J. Ciro, J. F. Cerón, *et al.*, "The people's speech: A large-scale diverse english speech recognition dataset for commercial usage," *CoRR*, vol. abs/2111.09344, 2021.

[24] X. Li, S. Takamichi, T. Saeki, W. Chen, *et al.*, "YODAS: Youtube-Oriented Dataset for Audio and Speech," in *Proc. ASRU*, 2023.

[25] J. Tian, Y. Peng, W. Chen, K. Choi, *et al.*, "On the effects of heterogeneous data sources on speech-to-text foundation models," in *Proc. Interspeech*, 2024.

[26] W. Chan, D. Park, C. Lee, Y. Zhang, *et al.*, "Speechstew: Simply mix all available speech recognition data to train one large neural network," *arXiv preprint arXiv:2104.02133*, 2021.

[27] S. Li, Y. You, X. Wang, Z. Tian, *et al.*, "MSR-86K: An Evolving, Multilingual Corpus with 86,300 Hours of Transcribed Audio for Speech Recognition Research," in *Proc. Interspeech*, 2024.

[28] W. Kang, X. Yang, Z. Yao, F. Kuang, *et al.*, "Libriheavy: A 50,000 hours asr corpus with punctuation casing and context," in *Proc. ICASSP*, 2024.

[29] Y. Yang, Z. Song, J. Zhuo, M. Cui, *et al.*, "Gigaspeech 2: An evolving, large-scale and multi-domain asr corpus for low-resource languages with automated crawling, transcription and refinement," *arXiv preprint arXiv:2406.11546*, 2024.

[30] M. Gaido, S. Papi, L. Bentivogli, A. Brutti, *et al.*, "MOSEL: 950,000 Hours of Speech Data for Open-Source Speech Foundation Model Training on EU Languages," in *Proc. EMNLP*, 2024.

[31] L. Kürzinger, D. Winkelbauer, L. Li, T. Watzel, and G. Rigoll, "CTC-Segmentation of Large Corpora for German End-to-End Speech Recognition," in *Speech and Computer*, 2020, pp. 267–278.

[32] A. Joulin, E. Grave, P. Bojanowski, and T. Mikolov, "Bag of tricks for efficient text classification," *arXiv preprint arXiv:1607.01759*, 2016.

[33] A. Joulin, E. Grave, P. Bojanowski, M. Douze, *et al.*, "Fasttext.zip: Compressing text classification models," *arXiv preprint arXiv:1612.03651*, 2016.

[34] M. Ravanelli, T. Parcollet, P. Plantinga, A. Rouhe, *et al.*, *Speechbrain: A general-purpose speech toolkit*, arXiv:2106.04624, 2021.

[35] R. Ardila *et al.*, "Common voice: A massively-multilingual speech corpus," *arXiv:1912.06670*, 2019.

[36] A. Paszke *et al.*, "Pytorch: An imperative style, high-performance deep learning library," in *Proc. NeurIPS*, 2019.

[37] T. Dao, "FlashAttention-2: Faster Attention with Better Parallelism and Work Partitioning," in *Proc. ICLR*, 2024.

[38] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *Proc. ICLR*, 2019.

[39] C. Wang *et al.*, "CoVoST 2 and Massively Multilingual Speech Translation," in *Proc. Interspeech*, 2021.

[40] A. Conneau *et al.*, "FLEURS: Few-Shot Learning Evaluation of Universal Representations of Speech," in *Proc. SLT*, 2022.

[41] V. Pratap, Q. Xu, A. Sriram, G. Synnaeve, and R. Collobert, "MLS: A large-scale multilingual dataset for speech research," *arXiv:2012.03411*, 2020.

[42] V. Srivastav, S. Majumdar, N. Koluguri, A. Moumen, S. Gandhi, *et al.*, *Open automatic speech recognition leaderboard*, `https://huggingface.co/spaces/hf-audio/open_asr_leaderboard`, 2023.