FINDINGS OF THE 2023 ML-SUPERB CHALLENGE: PRE-TRAINING AND EVALUATION OVER MORE LANGUAGES AND BEYOND

Jiatong Shi[1], William Chen[1], Dan Berrebbi[1], Hsiu-Hsuan Wang[2], Wei-Ping Huang[2], En-Pei Hu[2], Ho-Lam Chuang[2], Xuankai Chang[1], Yuxun Tang[3], Shang-Wen Li[4], Abdelrahman Mohamed[5], Hung-yi Lee[2], Shinji Watanabe[1]

[1]Carnegie Mellon University, [2]National Taiwan University
[3]Renmin University of China, [4]Meta AI, [5]Rembrand

ABSTRACT

The 2023 Multilingual Speech Universal Performance Benchmark (ML-SUPERB) Challenge expands upon the acclaimed SUPERB framework, emphasizing self-supervised models in multilingual speech recognition and language identification. The challenge comprises a research track focused on applying ML-SUPERB to specific multilingual subjects, a Challenge Track for model submissions, and a New Language Track where language resource researchers can contribute and evaluate their low-resource language data in the context of the latest progress in multilingual speech recognition. The challenge garnered 12 model submissions and 54 language corpora, resulting in a comprehensive benchmark encompassing 154 languages. The findings indicate that merely scaling models is not the definitive solution for multilingual speech tasks, and a variety of speech/voice types present significant challenges in multilingual speech processing.

Index Terms— Multilingual speech recognition, self-supervised learning, ML-SUPERB

1. INTRODUCTION

Self-supervised learning (SSL) has gained significant popularity in the speech community due to its effectiveness in capturing essential speech features, such as phonemes and acoustic units, through training on large amounts of unlabeled speech data [1]. These SSL models have shown remarkable improvements in various downstream tasks, including speech recognition, speaker identification, and emotion recognition [2]. In recent years, researchers have proposed diverse SSL models with different training objectives, operating under various data conditions, model architectures, and modalities [3,4].

The Speech Universal PERformance Benchmark (SUPERB), established in 2021, has emerged as a popular benchmark for evaluating speech SSL representations [2]. The primary goal of this benchmark is to compare speech SSLs across various speech processing tasks, encompassing aspects such as content, speaker, semantics, and paralinguistics. However, one limitation of SUPERB is its exclusive focus on English speech in its downstream tasks. In contrast, there is a growing interest in applying SSL models to multilingual scenarios, including training multilingual SSL models [5–7] or utilizing SSL models in a cross-lingual manner [8–11]. To facilitate research in these areas, a new benchmark called multilingual SUPERB (ML-SUPERB) has been proposed [12].

ML-SUPERB has been designed to encompass a wide range of languages, including both high-resource languages like English and endangered languages such as Totonac or Mixtec [13, 14].

The benchmark primarily focuses on evaluating SSL models for automatic speech recognition (ASR) and language identification (LID). To cater to different use cases for SSL models, ML-SUPERB includes two tracks with four different tasks: the monolingual track (monolingual ASR) and the multilingual track (multilingual ASR, LID, joint multilingual ASR/LID). Similar to SUPERB, ML-SUPERB utilizes frozen SSL models as feature extractors and employs a lightweight downstream model that can be fine-tuned for different tracks to achieve high training efficiency. The released public benchmark of ML-SUPERB covers 143 languages, making it highly inclusive and representative of diverse linguistic contexts.

Following the release of ML-SUPERB, the ML-SUPERB Challenge was inaugurated. In addition to a research track encompassing a variety of topics, we implemented two tracks for competitors: the Challenge Track and the freshly minted New Language Track. The Challenge Track emphasizes challenge performance, analogous to the SUPERB SLT2022 challenge. Conversely, the New Language Track instigates a novel design to the challenge by inviting participants to contribute their language resources. Through the creation of the New Language Track, ML-SUPERB continues to evolve by integrating new languages into its framework.

The challenge drew considerable interest and participation, with the Challenge Track receiving 12 model submissions and the New Language Track gaining an additional 54 valuable language resources. In incorporating these new languages, ML-SUPERB now extends its reach to an impressive total of **154** languages. It's noteworthy that all contributions were exclusively from academic institutions, demonstrating that the realm of multilingual SSL research isn't confined to large corporations, and indeed, academia can exert substantial influence. We also saw unique entries like WavLabLM, which were created from the ground up, independent of any pre-existing SSLs. The challenge's key findings include: (1) Scaling large models is not the only viable strategy for tackling multilingual speech tasks. (2) Diverse speech and voice types pose significant difficulties when applying multilingual speech representation to low-resource languages.

2. BACKGROUND

2.1. SUPERB and its Challenges

Since the public release of SUPERB, researchers have widely adopted the benchmark, showcasing its increasing prominence. The speech SSL toolkit, S3PRL, designed to facilitate researchers' participation in SUPERB, has gained significant attention from re-

searchers in self-supervised speech representation[1]. Reflecting the growing interest, the SUPERB team organized a dedicated SUPERB session during the 2nd Workshop on Self-supervised Learning for Audio and Speech Processing at AAAI 2022[2]. Additionally, some of the organizers conducted tutorials on SSL methodologies in speech and benchmarking with SUPERB at ICASSP 2022 and NAACL 2022[3], further emphasizing the relevance and impact of SUPERB in the research community.

More recently, the SUPERB team successfully organized the SUPERB challenge at SLT 2023 [15]. The challenge received 12 speech SSL submissions, highlighting the continued interest and advancement in the field. Recognizing the importance of addressing efficiency concerns in speech SSL, the challenge introduced an additional examination of memory and computation estimation, supplementing the original SUPERB framework that primarily focused on the model's performance across different tasks. This expansion reflects the community's emphasis on optimizing the efficiency of speech SSL models alongside their task-specific capabilities.

2.2. Multilingual Speech Self-supervised Representation

Multilingual speech representation learning has received significant attention from both academia and industry.

Before the advent of end-to-end ASR systems, numerous studies delved into multilingual representation for Hidden Markov Model (HMM)-based architectures [16–19]. With the evolution towards end-to-end ASR, researchers began investigating the use of large-scale multilingual end-to-end ASR models to learn generalized representations across multiple languages [20–22]. These studies underscored the importance of incorporating a broader range of languages and larger datasets to enhance performance, especially in low-resource ASR scenarios.

To further leverage the abundance of unlabeled data available in the wild, speech self-supervised models have been introduced in multilingual representation learning, offering the potential for even greater data utilization and performance gains. Kawakami et al., in 2020, leveraged various multilingual corpora with contrastive predictive coding (CPC), along with some English corpora, to jointly learn representations, resulting in significant performance enhancements across 22 languages [23]. More recently, Meta teams delved into multilingual representation learning using Transformer architecture, employing wav2vec 2.0, which led to notable advancements in the XLSR series of works [5–7]. These efforts exemplify the exploration and progress in leveraging self-supervised learning approaches for multilingual speech representation learning.

In recent works addressing low-resource or multilingual ASR, researchers have extensively explored the use of the XLSR model as a backbone, consistently achieving improvements over spectral features and other monolingual SSL approaches [8, 9, 11, 24–27]. However, there have been limited efforts to explore the integration of multilingual attributes into different self-supervised models, despite the success of WavLM [28] in outperforming wav2vec2 models across various speech processing tasks in SUPERB. This challenge, therefore, encourages the community to explore and develop improved architectures for multilingual speech processing.

In addition to ML-SUPERB, several other benchmarking initiatives have honed in on multilingual speech representation. LeBenchmark, for example, explores multilingual SSL in French speech processing [29], while IndicSUPERB concentrates on a range of In-

dian languages [30]. Furthermore, XTREME-S covers an expansive array of tasks by directly amalgamating existing multilingual corpora without confining itself to unsupervised conditions [31]. These benchmarking efforts provide invaluable resources and platforms for furthering research in multilingual speech representation.

Compared to other works, ML-SUPERB distinguishes itself by aiming to offer an efficient evaluation framework that covers an extensive range of languages and accounts for varying scenarios. To be precise, the public release of ML-SUPERB incorporates 143 languages, which, to the best of our knowledge, represents the benchmark with the most extensive language coverage. In terms of design, each language is equally sampled for ten minutes and one hour, thus curtailing the computational effort required for a complete evaluation cycle. With the introduction of the New Language Track, we anticipate an evolving, "open" ML-SUPERB that continues to integrate new languages.

3. TRACKS IN ML-SUPERB CHALLENGE

In this challenge, participants are invited to engage in three different tracks. In addition to the research track, which welcomes regular research papers, the main challenge comprises two distinct tracks: the Challenge Track and the New Language Track. These tracks offer participants the opportunity to showcase their expertise and innovation in specific areas related to the challenge objectives.

3.1. Challenge Track

The Challenge Track serves as the primary focus of this challenge, aiming to explore new multilingual self-supervised models. Participants have the opportunity to compete on two leaderboards: a public leaderboard and a hidden leaderboard.

The public leaderboard utilizes the ML-SUPERB benchmark data, which was released in [12]. On the other hand, the hidden leaderboard incorporates data submitted from the New Language Track. The evaluation methodology of ML-SUPERB follows a similar design to that of SUPERB. It utilizes a frozen upstream SSL model and a fixed downstream model architecture. Specifically, a weighted sum of layer-wise SSL features is employed, and a connectionist-temporal classification (CTC)-based transformer network is used as the downstream model.[4]

Both leaderboards consist of four tasks: monolingual ASR, multilingual ASR, LID, and multilingual ASR+LID. Each task has two configurations, with a 10-minute training set and a 1-hour training set, respectively. Additionally, the public leaderboard includes few-shot learning cases, where only 5 utterances from 20 languages are used in the training set. However, the hidden set does not consider this specific case.

3.2. New Language Track

In addition to the Challenge Track, the ML-SUPERB Challenge presents a unique New Language Track. This track is specifically tailored for researchers focused on language resources, particularly those keen on evaluating their low-resource language data employing cutting-edge ASR techniques. The principal aim of this track is to encourage researchers to contribute their unique language data to ML-SUPERB, consequently broadening the spectrum of multilingual research to encapsulate a wider variety of global languages. Participants engaging in the New Language Track are obligated

---

[1] https://github.com/s3prl/s3prl
[2] https://aaai-sas-2022.github.io/
[3] https://sites.google.com/view/tutorial-ssl-speech

[4] Implementation details are at https://github.com/espnet/espnet/tree/master/egs2/ml_superb/asr1.
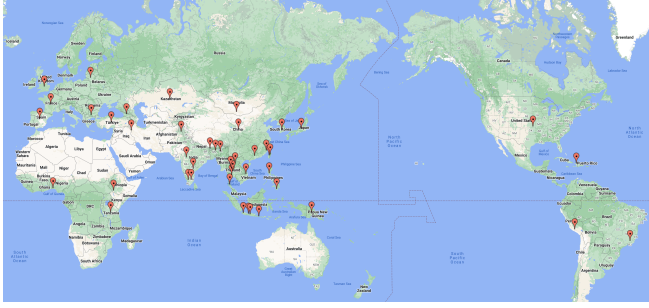
Fig. 1. Geographical distribution of the New Language track submissions. The 45 languages are marked on a map with their rough locations of speaking.

to offer a comprehensive description of their submitted data and to execute experiments utilizing established models from the ML-SUPERB benchmark.

An important aspect of the New Language Track is that the submitted data serves as an open evaluation set for other participants who have submitted their own SSL models to the challenge. This establishes collaboration and allows participants to evaluate their models on a diverse set of languages, contributing to a more comprehensive and inclusive evaluation of multilingual speech recognition.

4. SUBMISSIONS

4.1. New Language Track Submissions

For the New Language track, the challenge received a total of 8 submissions. In combination with the base evaluation hidden set prepared by the organizers, these submissions resulted in the creation of 54 new additional ML-SUPERB-style corpora. Figure 1 illustrates the distribution of these hidden languages, which are primarily concentrated in East Asia [32], Southeast Asia [33], and South Asia, while also exhibiting a reasonable distribution across other regions of the world (ÌròyìnSpeech [34], Quechua Speech [35], etc.). A number of submissions are created with existing corpora [36, 37, 37–39], while there are also effects in releasing newly published low-resource languages to the challenge.

In addition to the submissions received for the New Language Track, the organizers have prepared additional data for evaluation purposes, known as the base-hidden set. The base-hidden set consists of two main concentrations: multilingual conversational speech and multilingual singing voice. These additions create more challenging scenarios for multilingual understanding and evaluation. The conversational speech samples are drawn from various sources, including Babel [40], Fisher [41], Switchboard [42], KsponSpeech [43], and AccentedFrench[5]. On the other hand, the singing voice samples are taken from Muskits recipes [44], including Opencpop [45], PopCS [46], M4Singer [47], CSD [48], as well as a combination of the Oniku and Ofuton corpora[6]. The inclusion of these additional datasets in the base-hidden set enhances the complexity and diversity of the evaluation, providing more realistic and challenging scenarios for evaluating multilingual speech understanding systems.

The statistics of the hidden set are detailed in Table 1. Out of the 54 newly incorporated ML-SUPERB-style corpora, a total of

---

[5]https://www.openslr.org/57/

[6]Gotanmiya Kurumi Singing Voice Database: http://onikuru.info/db-download/ and Ofuton P Singing Voice Database: https://sites.google.com/view/oftn-utagoedb

Table 1. Benchmark statistics on the hidden leaderboard from the New Language track.

Dataset	Hours	Normal Langs (45)
10-minute	9.52	~10min × 54 (lang, data)
1-hour	57.13	~1h × 54 (lang, data)
Dev	9.31	~10min × 54 (lang, data)
Test	9.38	~10min × 54 (lang, data)

Table 2. Selected models from the public ML-SUPERB and challenge submissions from participants. Different colors represent different pre-trained languages: purple stands for monolingual SSL, blue stands for SSL trained in a few languages from the same region, and yellow stands for multilingual SSLs.

Model	Params (M)	Pre-Training	
		# Hours	# Langs
wav2vec2-base [3]	95	1k	1
wav2vec2-base-23 [49]	95	100k	23
XLSR-128 [5]	317	400k	128
HuBERT-base [4]	95	1k	1
HuBERT-large [4]	317	60k	1
mHuBERT-base [50]	95	14k	3
MMS-300m	317	491k	1,406
MMS-1b	965	491k	1,406
CV-HuBERT-base	95	13k	92
CV-HuBERT-base (40ms)	96	13k	92
CV-HuBERT-base (80ms)	96	13k	92
CV-HuBERT-MR-base	287	13k	92
EFFUSE (W2V2+XLSR)	634	400k	128
EFFUSE (HuBERT+XLSR)	634	400k	128
NWHC1	317	400k	128
NWHC2	317	400k	128
WavLabLM-base	95	40k	136
WavLabLM-large-EK	317	40k	136
WavLabLM-large-MK	317	40k	136
WavLabLM-large-MS	317	40k	136

18 have been deemed fitting for integration into the future public benchmark of ML-SUPERB. These 18 corpora will serve to broaden and enrich the publicly available benchmark. The remaining corpora have been earmarked for internal evaluation within the ML-SUPERB framework. With this newly established schema, we anticipate that ML-SUPERB will continue to evolve by perpetually adding new languages, fostering a better ecosystem for multilingual speech research worldwide.

4.2. Challenge Submissions

For the ML-SUPERB challenge, we received 12 model submissions, shown in Table 2. For readers' reference, the organizers also present the results of six example models in the original public benchmark.[7] The followings are their brief descriptions, which are categorized by their pre-training methods:

HuBERT with multiple resolutions: The set of models takes the insight from [51] by utilizing HuBERT with multiple resolutions [52]. In its pre-training stage, the participants utilized Common-Voice 11.0 dataset [53] and extracted the K-means units with the English HuBERT released in [54]. Similar to [51], the participants trained three HuBERT-based on 20ms, 40ms, and 80ms resolution

---

[7]The six models are selected based on their relative performances over the existing public benchmark reported in [12].

Table 3. {10-minute / 1-hour} set ML-SUPERB public benchmark (143 languages).

SSL	Monolingual ASR CER/PER	Multilingual ASR Normal CER	Multilingual ASR Few-shot CER	LID Normal ACC	Multilingual ASR + LID Normal ACC	Multilingual ASR + LID Normal CER	Multilingual ASR + LID Few-shot CER	SUPERB$_s$
FBANK	72.1 / 63.7	62.4 / 59.3	58.3 / 57.4	11.1 / 9.3	35.9 / 43.5	62.0 / 58.6	58.9 / 58.1	0 / 0
wav2vec2-base [3]	44.2 / 35.9	43.0 / 35.5	45.7 / 44.3	54.4 / 80.8	66.9 / 83.6	40.6 / 32.1	44.2 / 42.6	590.4 / 707.8
wav2vec2-base-23 [49]	49.2 / 35.1	37.7 / 32.0	43.4 / 42.2	58.7 / 71.9	45.1 / 66.3	37.2 / 30.9	44.3 / 43.0	563.0 / 676.5
XLSR-128 [5]	39.7 / 30.6	29.2 / 22.0	40.9 / 39.3	66.9 / 87.9	55.6 / 85.6	28.4 / 22.9	42.1 / 42.4	734.9 / 854.2
HuBERT-base [4]	42.8 / 35.3	39.8 / 31.4	44.5 / 42.7	61.2 / 86.1	71.5 / 86.0	39.2 / 30.9	43.8 / 41.8	650.8 / 757.0
HuBERT-large [4]	38.2 / 32.2	44.4 / 37.7	48.2 / 43.5	46.5 / 64.1	55.4 / 77.7	45.6 / 35.1	49.3 / 42.2	541.8 / 659.8
mHuBERT-base [50]	41.0 / 33.0	40.5 / 33.4	45.6 / 43.6	52.4 / 72.5	46.6 / 70.9	36.8 / 29.7	44.2 / 43.1	580.3 / 692.1
MMS-300m	33.8 / 30.5	28.7 / 24.0	36.5 / 36.5	62.3 / 84.3	71.9 / 74.3	31.5 / 30.0	30.9 / 29.2	826.7 / 841.3
MMS-1b	**33.3 / 25.7**	**21.3 / 18.1**	**30.2 / 30.8**	**84.8** / 86.1	73.3 / 74.8	**26.0** / 25.5	**25.4 / 24.8**	**983.5 / 943.2**
CV-HuBERT-base	41.9 / 32.9	35.4 / 27.5	44.0 / 40.8	71.2 / 84.0	76.6 / 87.3	35.1 / 28.2	43.6/ 41.1	726.3 / 796.7
CV-HuBERT-base (40ms)	71.6 / 62.6	60.5 / 52.0	57.5 / 53.0	65.6 / 83.0	65.7 / 83.3	59.6 / 52.3	57.7 / 53.4	179.6 / 380.4
CV-HuBERT-base (80ms)	76.4 / 67.6	72.7 / 70.7	66.1 / 64.1	33.2 / 57.2	17.2 / 39.4	72.3 / 70.4	64.2 / 64.4	130.4 / 16.2
CV-HuBERT-MR-base	47.8 / 38.3	37.0 / 28.3	43.2 / 40.8	64.1 / 86.0	74.8 / 84.5	36.2 / 30.6	42.5 / 41.0	659.0 / 755.0
EFFUSE (W2V2+XLSR)	38.5 / 28.9	31.0 / 24.4	40.9 / 40.2	23.0 / 13.5	69.7 / 87.8	31.4 / 24.4	41.8 / 39.3	610.2 / 624.6
EFFUSE (HuBERT+XLSR)	37.9 / 29.5	31.8 / 23.5	42.5 / 38.3	59.4 / 79.0	72.7 / 89.9	31.2 / 23.3	41.0 / 37.6	736.1 / 849.8
NWHC1	39.5 / 30.5	28.9 / 21.5	41.4 / 38.6	67.1 / 87.4	77.1 / **90.6**	28.8 / **21.5**	40.3 / 38.2	781.5 / 878.6
NWHC2	39.5 / 30.5	29.3 / 21.6	42.0 / 39.3	64.4 / **88.1**	**77.4 / 90.6**	28.4 / 21.8	41.5 / 38.8	767.5 / 875.3
WavLabLM-base	45.6 / 37.6	45.3 / 39.6	45.7 / 44.5	40.7 / 56.5	51.8 / 67.8	44.3 / 38.2	44.7 / 43.9	488.0 / 561.8
WavLabLM-large-EK	40.7 / 33.7	41.0 . 33,5	44.1 / 41.9	61.2 / 83.4	60.0 / 79.9	40.0 / 33.1	42.6 / 41.3	640.0 / 741.7
WavLabLM-large-MK	40.5 / 32.3	38.8 / 32.8	44.4 / 42.8	67.6 / 79.0	69.0 / 79.6	38.6 / 32.8	44.2 / 42.4	686.4 / 732.7
WavLabLM-large-MS	40.5 / 32.8	37.8 / 31.9	43.8 / 42.8	71.7 / 81.1	70.8 / 80.0	37.0 / 32.2	43.4 / 41.2	715.0 / 743.3

Table 4. {10-minute / 1-hour} set ML-SUPERB hidden benchmark (54 languages).

SSL	Monolingual ASR CER/PER	Multilingual ASR CER	LID ACC	Multilingual ASR + LID ACC	Multilingual ASR + LID CER	SUPERB$_s$
FBANK	76.4 / 70.7	71.9 / 68.4	21.8 / 14.7	29.5 / 37.1	70.4 / 65.8	0 / 0
wav2vec2-base [3]	62.7 / 54.2	55.0 / 47.0	43.0 / 60.2	47.0 / 42.2	54.3 / 42.6	585.3 / 620.3
wav2vec2-base-23 [49]	63.7 / 56.1	55.4 / 49.9	44.1 / 57.0	48.8 / 59.9	55.0 / 48.5	581.8 / 654.2
XLSR-128 [5]	57.6 / 49.2	47.9 / 39.0	48.4 / **70.2**	51.2 / **70.7**	47.0 / **37.7**	789.4 / 962.0
HuBERT-base [4]	61.1 / 53.8	54.9 / 47.2	47.5 / 65.1	46.8 / 63.1	53.6 / 45.8	637.2 / 760.5
HuBERT-large [4]	62.7 / 52.3	54.3 / 47.3	44.4 / 57.9	43.2 / 58.9	53.3 / 45.0	589.1 / 730.9
mHuBERT-base [50]	59.8 / 53.3	53.2 / 46.1	45.2 / 65.3	44.0 / 61.6	52.7 / 45.5	644.0 / 771.2
MMS-300m	60.1 / 51.0	47.3 / 42.2	48.9 / 45.3	55.4/ 66.2	46.2 / 40.8	788.6 / 774.7
MMS-1b	**55.4 / 46.8**	**42.0 / 37.4**	**59.4** / 65.4	**60.0** / 60.9	**40.9** / 39.5	**1000.0** / 933.9
CV-HuBERT-base	59.1 / 52.3	52.0 / 43.7	48.9 / 68.9	51.0 / 68.9	50.9 / 42.6	723.3 / 857.3
CV-HuBERT-base (40ms)	71.0 / 63.2	64.9 / 52.0	47.3 / 66.0	40.4 / 57.4	64.5 / 59.0	362.0 / 547.6
CV-HuBERT-base (80ms)	72.5 / 67.9	76.4 / 69.8	28.3 / 50.9	28.3 / 40.5	71.0 / 69.1	44.6 / 179.0
CV-HuBERT-MR-base	62.6 / 54.8	54.3 / 44.5	47.3 / 55.9	47.3 / 66.4	53.0 / 39.3	627.7 / 804.8
EFFUSE (W2V2+XLSR)	61.3 / 50.6	49.6 / 41.9	45.0 / 57.3	49.5 / 63.7	48.8 / 40.6	694.0 / 826.9
EFFUSE (HuBERT+XLSR)	57.5 / 49.1	50.7 / 40.3	48.2 / 55.9	48.9 / 66.4	49.2 / 39.3	747.1 / 865.0
NWHC1	56.7 / 47.7	48.2 / 38.9	47.6 / 69.4	51.5 / 69.4	47.0 / 38.5	793.5 / **966.5**
NWHC2	56.6 / 47.8	47.9 / 38.9	47.6 / 68.9	51.5 / 68.9	47.1 / 38.4	796.8 / 961.8
WavLabLM-base	55.9 / 47.2	56.6 / 47.6	45.4 / 58.2	45.7 / 58.1	54.7 / 46.8	661.8 /772.1
WavLabLM-large-EK	57.2 / 50.3	62.9 / 54.2	46.9 / 55.7	45.2 / 45.4	60.3 / 60.3	577.9 / 567.9
WavLabLM-large-MK	56.9 / 50.3	64.1 / 55.6	49.3 / 62.0	49.3 / 61.0	61.6 / 53.7	598.6 / 672.4
WavLabLM-large-MS	59.3 / 50.7	61.6 / 55.5	49.7 / 61.6	52.3 / 59.2	58.9 / 53.9	617.4 / 659.7

by controlling the convolutional feature extractor. To form a Hu-BERT with Multiple Resolution (HuBERT-MR), the participants further combined the three pre-trained HuBERT by simply concatenating them with upsampling. Following [51], the upsampling is simply repeating without additional introduction of learnable parameters. Four submissions are received from the participants, including three CommonVoice-HuBERT (i.e., CV-HuBERT)-base models trained on 20ms, 40ms, and 80ms resolutions and a CV-HuBERT-MR of the combination of all three HuBERT with different resolutions.

Ensemble modeling: The ensemble modeling is straightforward by stacking the representation of two SSL models. The implementa-

tion concept is related to the previous investigation in SSL model fusion [9,55]. The method does not introduce additional parameters as previous works [9,55], but instead has the requirement of the same representation resolution across different SSL models. The EFFUSE team submitted two models, including EFFUSE (wav2vec2+XLSR) and EFFUSE (HuBERT+XLSR) [56].

Parameter-level modification: The NWHC team's submissions introduce an inventive strategy aimed at preserving a higher volume of content information derived from SSL models [57]. Their hierarchical representation analysis, executed during ASR training, revealed a continual decrease in content information. This was particularly apparent in the diminished ASR performance within the concluding layers. To enhance the performance of downstream tasks involving the SSL, they propose a modification to the Massively Multilingual Speech (MMS)-300m model. Specifically, they replace the last few layers with intermediate layers: for NWHC1, the final three layers are supplanted by the 17th-19th layers, while for NWHC2, the last four layers are supplanted by the 17th-20th layers. Crucially, this modification is implemented directly at the level of network parameters, rather than at the resultant representation.

WavLM-style pre-training: WavLabLM are submitted by the WAVLab team [52], by adopting WavLM-style pre-training [28] into multilingual scenarios. A noisy speech simulation protocol is applied to the pre-training by mixing utterances and noises. The pre-training data combined several open-source corpora, reaching around 40k hours of speech over 136 languages. Four models are submitted for the challenge, including a base model with less parameter size and three large models with different training targets and strategies. All four models are trained with the k-means clusters from their previous HuBERT-large model [54]. WavLabLM-large-EK utilizes the English data to extract k-means targets, while the other two models utilize multilingual data to extract targets. WavLabLM-large-MS further adopts a multi-stage training strategy by upsampling low-resource languages within the dataset. Note that WavLabLMs are distinctive submissions that produce pre-training SSL models from the ground up, without relying on pre-existing SSLs. These models were developed using academic computing resources, and their source codes have been made publicly available to researchers. This makes it easier for individuals who do not have access to advanced computing resources to conduct pre-training research on SSL.

In addition to the submissions, the organizers also include the evaluation of pre-trained wav2vec2.0 presented in the Massively Multilingual Speech (MMS) project from Meta [58]. Therefore, in total, the evaluation of this ML-SUPERB challenge adds up to 14 new models to the benchmark.

5. CHALLENGE RESULTS SUMMARY

The public leaderboard results are presented in Table 3, while the hidden leaderboard results are shown in Table 4.

Overall: Across both the 10-minute and 1-hour public benchmarks (Table 3), MMS-1b delivered the best SUPERB score, significantly outpacing its competitors. Given the wide coverage of languages, it is unsurprising to observe the strong generalization capability of the MMS-based model in multilingual speech tasks, particularly in few-shot tasks. In the hidden benchmarks (Table 4), MMS-1b excels in the 10-minute benchmark, but was overtaken by the XLSR-128 and NWHC models in the 1-hour benchmark, especially in the LID and multilingual ASR+LID tasks. Remarkably, NWHC2 achieves the best performance in the 1-hour hidden leaderboard, outperforming XLSR-128 and NWHC1.

HuBERT with multiple resolutions: In all scenarios, CV-HuBERT-base utilizing the default 20ms resolution units outperformed the original HuBERT-base model, emphasizing the importance of multilingual pre-training. Contrary to the observation in [51], HuBERT with multiple resolutions did not invariably enhance performance, but often undermined the performance of the CV-HuBERT-base. This suggests that techniques used in monolingual contexts may not always translate effectively to multilingual scenarios due to linguistic variation and phonetic distribution differences.

Ensemble modeling: Despite being a prevalent strategy in numerous challenges, naive ensemble modeling by concatenating SSL features does not always improve multilingual task performance. Specifically, in this challenge, ensemble modeling's performance seems largely limited by weaker SSL models. Consequently, both ensemble models (i.e., EFFUSE models) did not significantly outperform XLSR-128, even though the XLSR-128 representation is included in the ensemble frameworks.

Parameter-level modification: Despite being based on modified versions of MMS-300m, the NWHC systems surpassed MMS-300m across all benchmarks, even achieving the best performance in the 1-hour hidden benchmark. These impressive results underscore the significance of layer-wise analysis of self-supervised models and raise questions about how to fully leverage SSL models.

WavLM-style pre-training: While WavLM has been a top performer in the SUPERB benchmark [2], one may wonder whether similar pre-training approaches could excel in multilingual scenarios. The analysis of WavLabLMs revealed that the impact of denoising modeling in ML-SUPERB may not be as beneficial as in SUPERB. While the model series outperformed the original HuBERT-based model in the public benchmark, their performance deteriorated in the hidden benchmarks, especially in multilingual tasks.

Public benchmarks vs. hidden benchmarks: As per [12], the majority of the public benchmark comprises read speech, while the hidden benchmark includes a substantial set of conversational speech and singing voices. These varied voice styles present significant challenges for the ML-SUPERB tasks. The performance in the hidden sets is considerably worse than that in the public sets. Despite generally similar rankings, some SSL model performances vary. For instance, MMS-1b, despite being the top performer in all other leaderboards, did not achieve the best performance in the 1-hour hidden benchmark. As in the real world, researchers focusing on low-resource languages often have limited access to clean-read speech. Therefore, studying multilingual representation across different voice types could become a major research direction.

Multilingual vs. monolingual: In the inaugural ML-SUPERB release, most models were either monolingual-based or focused on a small set of languages, with limited exploration of multilingual SSL models. For the 2023 ML-SUPERB challenge, all model submissions underwent pre-training in multilingual data spanning over 50 languages. The leaderboard results clearly indicate that multilingual SSL typically outperforms those trained with a limited language scope, suggesting that focusing on multilingual representation for multilingual tasks is a promising future research direction.

Performance vs. efficiency: Taking a cue from the SUPERB challenge at SLT2022 [15], we've also considered efficiency as a critical factor when evaluating self-supervised models in the ML-SUPERB challenge. We've gauged the theoretical multiply-accumulate operation (MACs) based on the profiling toolkit utilized in the prior SUPERB challenge[8]. The tradeoff between model computational com-

---

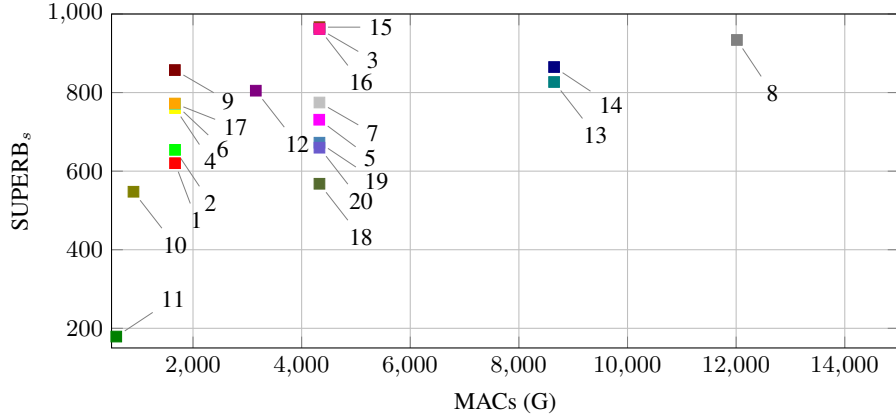[8] https://github.com/B06901052/DeepSpeed/tree/superb-challenge

Fig. 2. MACs v.s. SUPERB score in ML-SUPERB 1-hour hidden benchmark.

plexity (measured by MACs) and ML-SUPERB performance (indicated by the SUPERB score) is illustrated in Fig 2.

In the previous SUPERB challenge [15], we observed a scaling rule where increasing model size typically led to improved performances across various speech processing tasks. However, it is evident that the scaling rule isn't always effective for multilingual tasks. Specifically, smaller models can sometimes deliver performance equal to or better than their larger counterparts. For instance, CV-HuBERT-base outperforms MMS-300m and HuBERT-large, even though it utilizes a base architecture. Similarly, NWHC2 slightly outperforms MMS-1b in the 1-hour hidden benchmark, despite its smaller size and computational burden. In more comparable scenarios, WavLabLM-base outperforms all three WavLabLM-large models, with model size being the only differentiating factor. These observations hint that large-scale SSL pre-training might not be the only viable path for multilingual SSL representation learning.

6. CONCLUSION

The ML-SUPERB challenge of 2023 has provided a platform for exploring and developing multilingual speech SSL models in multilingual ASR and LID. The challenge attracted wide-ranging participation, yielding valuable insights into the state-of-the-art methods and potential future directions for this emerging field.

Findings from the challenge underscore that while large model scaling can be effective, it is not the exclusive solution to advancing multilingual speech tasks. It was observed that smaller models could potentially deliver comparable or even superior performance, highlighting the potential for efficient and effective model development. The challenge also uncovered the major difficulties in tackling varying speech and voice types, especially in low-resource languages, pointing to a crucial research direction for future multilingual representation learning endeavors.

The results of this year's ML-SUPERB challenge also reinforce the notion that multilingual SSL models usually outperform those trained with limited language coverage. Therefore, the path towards multilingual representation for multilingual tasks stands out as a promising direction to further explore.

As the field continues to evolve, the insights gathered from this challenge will serve as valuable stepping stones, informing and directing future research efforts in multilingual SSL model development. We expect the ML-SUPERB challenge to continue to play a pivotal role in shaping this fascinating and crucial area of research.

7. ACKNOWLEDGEMENTS

8. REFERENCES

[1] Abdelrahman Mohamed, Hung-yi Lee, Lasse Borgholt, Jakob D Havtorn, Joakim Edin, Christian Igel, Katrin Kirchhoff, Shang-Wen Li, Karen Livescu, et al., "Self-supervised speech representation learning: A review," *JSTSP*, 2022.

[2] Shu wen Yang, Po-Han Chi, Yung-Sung Chuang, Cheng-I Jeff Lai, Kushal Lakhotia, Yist Y. Lin, Andy T. Liu, Jiatong Shi, et al., "SUPERB: Speech Processing Universal PERformance Benchmark," in *Proc. Interspeech*, 2021, pp. 1194–1198.

[3] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Proc. NeurIPS*, vol. 33, pp. 12449–12460, 2020.

[4] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *TASLP*, vol. 29, pp. 3451–3460, 2021.

[5] Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, et al., "XLS-R: Self-supervised cross-lingual speech representation learning at scale," *arXiv preprint arXiv:2111.09296*, 2021.

[6] Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli, "Unsupervised cross-lingual representation learning for speech recognition," *Proc. Interspeech*, 2020.

[7] Paul-Ambroise Duquenne, Hongyu Gong, Ning Dong, Jingfei Du, Ann Lee, Vedanuj Goswani, Changhan Wang, Juan Pino, Benoît Sagot, and Holger Schwenk, "Speechmatrix: A large-scale mined corpus of multilingual speech-to-speech translations," *arXiv preprint arXiv:2211.04508*, 2022.

[8] Jing Zhao and Wei-Qiang Zhang, "Improving automatic speech recognition performance for low-resource languages with self-supervised models," *JSTSP*, vol. 16, no. 6, pp. 1227–1241, 2022.

[9] Dan Berrebbi, Jiatong Shi, Brian Yan, Osbel López-Francisco, Jonathan Amith, and Shinji Watanabe, "Combining Spectral and Self-Supervised Features for Low Resource Speech Recognition and Translation," in *Proc. Interspeech*, 2022, pp. 3533–3537.

[10] Anne Wu, Changhan Wang, Juan Pino, and Jiatao Gu, "Self-supervised representations improve end-to-end speech translation," *Proc. Interspeech*, pp. 1491–1495, 2020.

[11] Xinjian Li, Florian Metze, David R Mortensen, Alan W Black, and Shinji Watanabe, "ASR2K: Speech recognition for around 2000 languages without audio," *Proc. Interspeech*, pp. 4885–4889, 2022.

[12] Jiatong Shi, Dan Berrebbi, William Chen, Ho-Lam Chung, En-Pei Hu, Wei Ping Huang, Xuankai Chang, Shang-Wen Li, Abdelrahman Mohamed, Hung-yi Lee, et al., "ML-SUPERB: Multilingual speech universal performance benchmark," *Proc. Interspeech*, 2023.

[13] Jiatong Shi, Jonathan D Amith, Xuankai Chang, Siddharth Dalmia, Brian Yan, and Shinji Watanabe, "Highland puebla nahuatl speech translation corpus for endangered language documentation," in *Proc. AmericaNLP*, 2021, pp. 53–63.

[14] Jiatong Shi, Jonathan D Amith, Rey Castillo García, Esteban Guadalupe Sierra, Kevin Duh, and Shinji Watanabe, "Leveraging end-to-end asr for endangered language documentation: An empirical study on Yolóxochitl Mixtec," in *Proc. EACL*, 2021, pp. 1134–1145.

[15] Tzu-hsun Feng, Annie Dong, Ching-Feng Yeh, Shu-wen Yang, Tzu-Quan Lin, Jiatong Shi, Kai-Wei Chang, Zili Huang, Haibin Wu, et al., "SUPERB@ SLT 2022: Challenge on generalization and efficiency of self-supervised speech representation learning," in *Proc. SLT*, 2023, pp. 1096–1103.

[16] Karel Veselý, Martin Karafiát, František Grézl, Miloš Janda, and Ekaterina Egorova, "The language-independent bottleneck features," in *Proc. SLT*, 2012, pp. 336–341.

[17] Ngoc Thang Vu, Florian Metze, and Tanja Schultz, "Multilingual bottle-neck features and its application for under-resourced languages," in *Spoken language technologies for under-resourced languages*, 2012.

[18] Jia Cui, Brian Kingsbury, Bhuvana Ramabhadran, Abhinav Sethy, Kartik Audhkhasi, Xiaodong Cui, Ellen Kislal, Lidia Mangu, Markus Nussbaum-Thom, Michael Picheny, et al., "Multilingual representations for low resource speech recognition and keyword search," in *Proc. ASRU*, 2015, pp. 259–266.

[19] Tom Sercu, George Saon, Jia Cui, Xiaodong Cui, Bhuvana Ramabhadran, Brian Kingsbury, and Abhinav Sethy, "Network architectures for multilingual speech representation learning," in *Proc. ICASSP*, 2017, pp. 5295–5299.

[20] Wenxin Hou, Yue Dong, Bairong Zhuang, Longfei Yang, Jiatong Shi, and Takahiro Shinozaki, "Large-Scale End-to-End Multilingual Speech Recognition and Language Identification with Multi-Task Learning," in *Proc. Interspeech*, 2020, pp. 1037–1041.

[21] Vineel Pratap, Anuroop Sriram, Paden Tomasello, Awni Hannun, Vitaliy Liptchinsky, Gabriel Synnaeve, and Ronan Collobert, "Massively Multilingual ASR: 50 Languages, 1 Model, 1 Billion Parameters," in *Proc. Interspeech*, 2020, pp. 4751–4755.

[22] Bo Li, Ruoming Pang, Tara N Sainath, Anmol Gulati, Yu Zhang, James Qin, Parisa Haghani, et al., "Scaling end-to-end models for large-scale multilingual asr," in *Proc. ASRU*, 2021, pp. 1011–1018.

[23] Kazuya Kawakami, Luyu Wang, Chris Dyer, Phil Blunsom, and Aaron van den Oord, "Learning robust and multilingual speech representations," in *Findings of EMNLP*, 2020, pp. 1182–1192.

[24] William Chen, Brian Yan, Jiatong Shi, Yifan Peng, Soumi Maiti, and Shinji Watanabe, "Improving massively multilingual ASR with auxiliary CTC objectives," *Proc. Interspeech*, 2023.

[25] Jinyi Yang, Amir Hussein, Matthew Wiesner, and Sanjeev Khudanpur, "Jhu iwslt 2022 dialect speech translation system description," in *Proc. IWSLT*, 2022, pp. 319–326.

[26] Siddhant Arora, Siddharth Dalmia, Pavel Denisov, Xuankai Chang, Yushi Ueda, Yifan Peng, Yuekai Zhang, Sujay Kumar, Karthik Ganesan, Brian Yan, et al., "ESPnet-SLU: Advancing spoken language understanding through espnet," in *Proc. ICASSP*, 2022, pp. 7167–7171.

[27] Andros Tjandra, Diptanu Gon Choudhury, Frank Zhang, Kritika Singh, Alexis Conneau, Alexei Baevski, Assaf Sela, Yatharth Saraf, and Michael Auli, "Improved language identification through cross-lingual self-supervised learning," in *Proc. ICASSP*, 2022, pp. 6877–6881.

[28] Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, et al., "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," *JSTSP*, vol. 16, no. 6, pp. 1505–1518, 2022.

[29] Solène Evain, Ha Nguyen, Hang Le, Marcely Zanon Boito, Salima Mdhaffar, Sina Alisamir, Ziyi Tong, Natalia Tomashenko, Marco Dinarelli, Titouan Parcollet, Alexandre Allauzen, et al., " LeBenchmark: A Reproducible Framework for Assessing Self-Supervised Representation Learning from Speech," in *Proc. Interspeech*, 2021, pp. 1439–1443.

[30] Tahir Javed, Kaushal Santosh Bhogale, Abhigyan Raman, Anoop Kunchukuttan, Pratyush Kumar, and Mitesh M Khapra, "Indicsuperb: A speech processing universal performance benchmark for indian languages," *Proc. AAAI*, 2022.

[31] Alexis Conneau, Ankur Bapna, Yu Zhang, Min Ma, Patrick von Platen, Anton Lozhkov, Colin Cherry, Ye Jia, Clara Rivera, et al., "XTREME-S: Evaluating Cross-lingual Speech Representations," in *Proc. Interspeech*, 2022, pp. 3248–3252.

[32] Yi-Hui Chou, Kalvin Chang, Meng-Ju Wu, Winston Ou, Alice Wen-Hsin Bi, Carol Yang, Bryan Y. Chen, Rong-Wei Pai, Po-Yen Yeh, Jo-Peng Chiang, Iu-Tshiann Phoann, Winnie Chang, Chenxuan Cui, Noel Chen, and Jiatong Shi, "Evaluating self-supervised speech models on a taiwanese hokkien corpus," in *Proc. ASRU*, 2023.

[33] Sakriani Sakti and Benita Angela Titalim, "Leveraging the multilingual indonesian ethnic languages dataset in self-supervised model for low-resource asr task," in *Proc. ASRU*, 2023.

[34] Tolulope Ogunremi, Kola Tubosun, Anuoluwapo Aremu, Iroro Orife, and David Ifeoluwa Adelani, "Ìròyìnspeech: A multi-purpose yorùbá speech corpus," *arXiv preprint arXiv:2307.16071*, 2023.

[35] Chih-Chen Chen, William Chen, Rodolfo Zevallos, and John Ortega, "Evaluating self-supervised speech representations for indigenous American languages," *arXiv preprint arXiv:2310.03639*, 2023.

[36] Artit Suwanbandit, Burin Naowarat, Orathai Sangpetch, and Ekapol Chuangsuwanich, "Thai dialect corpus and transfer-based curriculum learning investigation for dialect automatic speech recognition," in *Proc. Interspeech*, 2023.

[37] Sakriani Sakti and Satoshi Nakamura, "Towards language preservation: Design and collection of graphemically balanced and parallel speech corpora of indonesian ethnic languages," in *Proc. O-COCOSDA/CASLRE*, 2013, pp. 1–5.

[38] Samuel Cahyawijaya, Holy Lovenia, Alham Fikri Aji, Genta Winata, Bryan Wilie, Fajri Koto, Rahmad Mahendra, Christian Wibisono, Ade Romadhony, Karissa Vincentio, et al., "NusaCrowd: Open source initiative for Indonesian NLP resources," in *Findings of ACL*, 2023, pp. 13745–13818.

[39] Ronald Cardenas, Rodolfo Zevallos, Reynaldo Baquerizo, and Luis Camacho, "Siminchik: A speech corpus for preservation of southern quechua," *ISI-NLP 2*, p. 21, 2018.

[40] Peter Roach, Simon Arnfield, William Barry, Julia Baltova, Marian Boldea, Adrian Fourcin, Wiktor Gonet, et al., "BABEL: An eastern european multi-language database," in *Proc. ICSLP*, 1996, vol. 3, pp. 1892–1893.

[41] Christopher Cieri, David Miller, and Kevin Walker, "The Fisher corpus: A resource for the next generations of speech-to-text.," in *LREC*, 2004, vol. 4, pp. 69–71.

[42] John J Godfrey, Edward C Holliman, and Jane McDaniel, "SWITCHBOARD: Telephone speech corpus for research and development," in *Proc. ICASSP*, 1992, vol. 1, pp. 517–520.

[43] Jeong-Uk Bang, Seung Yun, Seung-Hi Kim, Mu-Yeol Choi, Min-Kyu Lee, et al., "Ksponspeech: Korean spontaneous speech corpus for automatic speech recognition," *Applied Sciences*, vol. 10, no. 19, pp. 6936, 2020.

[44] Jiatong Shi, Shuai Guo, Tao Qian, Nan Huo, Tomoki Hayashi, Yuning Wu, Frank Xu, Xuankai Chang, et al., "Muskits: an end-to-end music processing toolkit for singing voice synthesis," in *Proc. Interspeech*, 2022, pp. 4277–4281.

[45] Yu Wang, Xinsheng Wang, Pengcheng Zhu, Jie Wu, Hanzhao Li, Heyang Xue, Yongmao Zhang, Lei Xie, and Mengxiao Bi, "Opencpop: A High-Quality Open Source Chinese Popular Song Corpus for Singing Voice Synthesis," in *Proc. Interspeech*, 2022, pp. 4242–4246.

[46] Rongjie Huang, Feiyang Chen, Yi Ren, Jinglin Liu, Chenye Cui, and Zhou Zhao, "Multi-singer: Fast multi-singer singing voice vocoder with a large-scale corpus," in *Proc. ACMMM*, 2021, pp. 3945–3954.

[47] Lichao Zhang, Ruiqi Li, Shoutong Wang, Liqun Deng, Jinglin Liu, Yi Ren, Jinzheng He, Rongjie Huang, Jieming Zhu, Xiao Chen, et al., "M4singer: A multi-style, multi-singer and musical score provided mandarin singing corpus," *Proc. NIPS*, vol. 35, pp. 6914–6926, 2022.

[48] Soonbeom Choi, Wonil Kim, Saebyul Park, Sangeon Yong, and Juhan Nam, "Children's song dataset for singing voice research," in *Proc. ISMIR*, 2020.

[49] Changhan Wang, Morgane Riviere, Ann Lee, Anne Wu, Chaitanya Talnikar, Daniel Haziza, et al., "Voxpopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation," in *Proc. ACL*, 2021, pp. 993–1003.

[50] Ann Lee, Hongyu Gong, Paul-Ambroise Duquenne, Holger Schwenk, Peng-Jen Chen, Changhan Wang, Sravya Popuri, et al., "Textless speech-to-speech translation on real data," in *Proc. NAACL*, 2022, pp. 860–872.

[51] Jiatong Shi, Yun Tang, Hirofumi Inaguma, Hongyu Gong, Juan Pino, and Shinji Watanabe, "Exploration on hubert with multiple resolutions," *Proc. Interspeech*, 2023.

[52] William Chen, Jiatong Shi, Brian Yan, Dan Berrebbi, Wangyou Zhang, Yifan Peng, Xuankai Chang, Soumi Maiti, and Shinji Watanabe, "Joint prediction and denoising for large-scale multilingual self-supervised learning," *arXiv preprint arXiv:2309.15317*, 2023.

[53] Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber, "Common voice: A massively-multilingual speech corpus," in *Proc. LREC*, 2020, pp. 4218–4222.

[54] William Chen, Xuankai Chang, Yifan Peng, Zhaoheng Ni, Soumi Maiti, and Shinji Watanabe, "Reducing barriers to self-supervised learning: Hubert pre-training with academic compute," *Proc. Interspeech*, 2023.

[55] Szu-Jui Chen, Jiamin Xie, and John H.L. Hansen, "FeaRLESS: Feature Refinement Loss for Ensembling Self-Supervised Learning Features in Robust End-to-end Speech Recognition," in *Proc. Interspeech*, 2022, pp. 3058–3062.

[56] Tejes Srivastava, Jiatong Shi, William Chen, and Shinji Watanabe, "EFFUSE: Efficient self-supervised feature fusion for E2E ASR in multilingual and low resource scenarios," *arXiv preprint arXiv:2310.03938*, 2023.

[57] Hongfei Xue, Qijie Shao, Kaixun Huang, Peikun Chen, Lei Xie, and Jie Liu, "SSHR: Leveraging self-supervised hierarchical representations for multilingual automatic speech recognition," *arXiv preprint arXiv:2309.16937*, 2023.

[58] Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, Zhaoheng Ni, et al., "Scaling speech technology to 1,000+ languages," *arXiv preprint arXiv:2305.13516*, 2023.