INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

ML-SUPERB: Multilingual Speech Universal PERformance Benchmark

Jiatong Shi[1], *Dan Berrebbi*[1*], *William Chen*[1*], *Ho-Lam Chung*[2*], *En-Pei Hu*[2*], *Wei Ping Huang*[2*], *Xuankai Chang*[1], *Shang-Wen Li*[3], *Abdelrahman Mohamed*[4], *Hung-yi Lee*[2], *Shinji Watanabe*[1]

[1]Carnegie Mellon University [2]National Taiwan University [3]Meta AI [4]Rembrand

{jiatongs, dberrebbi, wc4, swatanab}@cs.cmu.edu, shangwenl@meta.com, abdo@rembrand.com
hungyilee@ntu.edu.tw

Abstract

Speech processing Universal PERformance Benchmark (SUPERB) is a leaderboard to benchmark the performance of Self-Supervised Learning (SSL) models on various speech processing tasks. However, SUPERB largely considers English speech in its evaluation. This paper presents multilingual SUPERB (ML-SUPERB), covering 143 languages (ranging from high-resource to endangered), and considering both automatic speech recognition and language identification. Following the concept of SUPERB, ML-SUPERB utilizes frozen SSL features and employs a simple framework for multilingual tasks by learning a shallow downstream model. Similar to the SUPERB benchmark, we find speech SSL models can significantly improve performance compared to FBANK features. Furthermore, we find that multilingual models do not always perform better than their monolingual counterparts. We will release ML-SUPERB as a challenge with organized datasets and reproducible training scripts for future multilingual representation research.

Index Terms: speech self-supervised learning, multilingual speech recognition, language identification

1. Introduction

Self-supervised learning (SSL) has been a popular method in the speech community. SSL models have shown promising results by capturing important speech features, such as phonemes and other acoustic units, through training on large amounts of unlabeled speech data [1]. These models have led to significant improvements in downstream tasks, such as speech recognition, speaker identification, and emotion recognition [2]. Over the past few years, researchers have proposed a variety of SSL models with different training objectives, operating under various data conditions, model architectures, and modalities [3, 4].

A major challenge in evaluating SSL models for speech is the difficulty of comparison since most models have been evaluated using different experimental setups. To address this issue, Yang et al. introduced the Speech processing Universal PERformance Benchmark (SUPERB) [2]. Recently, an extension of SUPERB called SUPERB-SG [5] has been introduced. SUPERB provides a comprehensive speech SSL benchmark including tasks such as recognition, detection, semantics, speaker identification, paralinguistics, and generation. With SUPERB, researchers can more easily compare the performance of different SSL models on various speech-related tasks, universally.

While SUPERB covers a wide range of speech tasks, it was designed primarily for English speech. However, there has been growing interest in applying SSL models to multilingual scenarios, such as training multilingual SSL models [6–8] or using SSL models in a cross-lingual manner [9–12]. To support future research in these areas, we propose a new benchmark called multilingual SUPERB (ML-SUPERB).

ML-SUPERB is designed to cover a wide range of languages, including both high-resource languages like English and endangered languages such as Totonac. The benchmark primarily focuses on evaluating SSL models for automatic speech recognition (ASR) and language identification (LID). To accommodate 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 employs frozen SSL models as feature extractors and a lightweight downstream model that can be fine-tuned for different tracks to achieve high training efficiency.

Several existing benchmarks also include multilingual SSL models [13–15]. Lebenchmark primarily evaluates speech tasks in French [13]; IndicSUPERB focuses mostly on Indian languages [14]. XTREME-S focuses on multilingual speech representation benchmarks, including ASR, speech translation, speech classification, and speech retrieval [15]. There are three main differences between XTREME-S and ML-SUPERB. Firstly, ML-SUPERB covers a wider range of languages, with 143 languages compared to XTREME-S's 102. Secondly, ML-SUPERB focuses on ASR and LID, while XTREME-S covers four different tasks. However, ML-SUPERB expands the tasks by evaluating them in four common multilingual research scenarios, while XTREME-S considers multilingual training only. Finally, ML-SUPERB is designed for efficiency, using smaller benchmark datasets and downstream models, and does not include fine-tuning. This lightweight setup allows us to conduct experiments for a dozen of popular speech SSL models, trained with various sizes and pre-training sets, and compare their performances across the proposed tracks. We expect ML-SUPERB would be a valuable complement to existing benchmarks.

2. Benchmark Details

2.1. Data Collection

ML-SUPERB gathers data from a wide range of multilingual speech corpora, including Multilingual Librispeech [16], Commonvoice [17], Voxforge [18], Voxpopuli [19], Googlei18n open-source project [20–22], Nordic Language Technology ASR corpora [23], Fleurs [24], NCHLT Speech [25], Spoken Wikipedia corpus [26], Mexican endangered languages [10, 27, 28], M-AILab multilingual corpora [29], Living Audio dataset [30], ALFFA corpus [31]. All corpora are with either Creative Commons, MIT, GNU, or Free-BSD licenses, which are available for both industrial and academic research, permissively.

For each language-corpus pair denoted as (`lang`, `data`),

*Equal contribution, sorted in alphabetical order.

 10.21437/Interspeech.2023-1316

Table 1: *Statistics of the data used for training, development, and testing in ML-SUPERB. Detailed discussed in Sec. 2.1.*

Dataset	Hours	Normal Langs (123)	Few-shot Langs (20)
10-minute	37.43	~10min × 240 (lang, data)	5 utt. × 20 lang
1-hour	222.46	~1h × 240 (lang, data)	5 utt. × 20 lang
Dev.t	41.82	~10min × 240 (lang, data)	~10min × 31 (lang, data)
Test	44.97	~10min × 240 (lang, data)	~10min × 31 (lang, data)

three 10-minute subsets are randomly extracted for training, development, and testing, along with an additional 1-hour training set that includes the 10-minute training set.[1] The reasons for using a small 10-minute/1-hour training set are: (1) *Challenging design*: using a large training data size could lead to high performance easily and may result in a saturated benchmark in evaluation metrics [3, 4]. Therefore, using a smaller training set size presents a more challenging design for the SSL models, which can help evaluate their robustness and generalization capability. (2) *Reasonable performance*: previous speech SSL works have frequently adopted 10-minute and 1-hour training sizes. Even in such extreme cases, the performances with SSL are generally reasonable [3, 4], indicating that this setting could be a feasible solution to the benchmark as well. (3) *Training efficiency*: with 143 languages coverage, limiting the training size is important to keep the experiments within reasonable computational efforts. Using a smaller training set size can help reduce the computational cost and make the training process more efficient. A full evaluation cycle of ML-SUPERB can take up to 3 days using 4 2080Ti GPUs.

Additionally, the benchmark includes few-shot cases with 20 languages and uses only 5 utterances in training for each language. These reserved few-shot training sets are not used in the monolingual ASR track. A detailed summary of the dataset is shown in Table 1.

2.2. Monolingual Track

The literature suggests that speech SSL models are commonly fine-tuned on monolingual corpora [9–11]. In ML-SUPERB, we introduce a dedicated track for monolingual ASR to facilitate this approach. We select nine languages based on geographical and linguistic considerations to balance language and domain coverage with manageable experimental mass.

In total, we introduce 14 `monolingual_exp`. For a `monolingual_exp` in language `lang` we select one dataset of this language and use it for training the model and for validation[2]. For evaluation of a `monolingual_exp`, we use all the datasets of `lang` to test the trained model on various accent or domain conditions. We select one pair (`lang` , `data`) for training for `lang`$\in$ {`rus`, `swa`, `swe`, `jpn`, `cmn`, `xty`}. For `lang`$\in$ {`eng`, `fra`, `deu`} we select respectively 3, 2 and 2 pairs (`lang`, `data`) in order to evaluate the impact of the training domain on the models' performances. For instance, for `eng` we have 3 `monolingual_exp`, with (`eng` , `MLS`), (`eng` , `NCHLT`) and (`eng` , `VoxPopuli`).

2.3. Multilingual Track

Multilingual ASR task: in the multilingual ASR task, we use the training set where combining text transcriptions from all 143 languages. The multilingual ASR task has two sub-tasks on the 10-minute train set and the 1-hour train set. For both training sets, we reserve 20 languages for few-shot learning scenarios as discussed in Sec. 2.1. In this track, the model is expected to directly predict the correct orthography in the target language.

LID task: LID track focuses on language identification with the same training set of 143 languages in 10 minutes and 1 hour. However, we do not consider evaluation for languages with few-shot settings, given that the identification of those languages is very challenging due to the label biasing.

Joint Multilingual ASR/LID task: A widely used technique in previous literature involves adding the language ID to the start of the speech transcript to facilitate joint training of multilingual ASR and LID models [32–35]. Joint training can improve performance in certain scenarios, and it can also enhance model interpretability by separating language identification errors. Therefore, we have included this task in our multilingual track. The task's design is the same as the multilingual ASR task for ASR and the LID task for language identification.

2.4. Framework and Benchmark Settings

Toolkits: We utilize the S3PRL toolkit [2] for upstream models, which offers a wide range of speech SSL model architectures and APIs that support customized SSL models from Huggingface [36] and user-defined models. For task-specific downstream training, we use ESPnet [37]. We plan to publish ML-SUPERB as an all-in-one recipe in ESPnet's `egs2` recipe collection, encompassing data preprocessing, training, inference, and evaluation[3].

Downstream model and training details: Our downstream model design is based on the SUPERB concept. First, we compute a weighted summation of frozen speech SSL representations using learnable weights. Next, we apply a convolutional downsample layer that reduces the sequence of speech SSL features by half, passing the resulting hidden states to a transformer model consisting of two layers with an attention dimension of 256, a feedforward layer dimension of 1024, and 8 attention heads. A dropout rate of 0.1 is employed, and the model is trained using the connectionist temporal Ccassification loss. We use the Adam optimizer with a learning rate of 0.0001 and 1e-6 weight decay. Specaugment is applied to the representation (i.e., the weighted sum of speech SSL representation) following the SUPERB benchmark. The batch size is set to 8 with the gradient accumulation as 4. The same configuration is used for all tasks in both the monolingual and multilingual tracks.

The number of iterations in training is the only difference across tasks. In the monolingual track, due to the small training size, we set it to 15,000. In the multilingual track, we use 300,000 iterations for the 10-minute train set and 600,000 for the 1-hour train set.

Evaluation metric: In the monolingual track, the phoneme error rate is used for `jpn` and `cmn`, while Character Error Rate (CER) is used for the remaining languages. In the multilingual track, we use CER for ASR evaluation and accuracy rate

[1]We used the original split for source datasets, with the exception of SWC, M-AILABS, LAD, and ALFFA. Therefore, all datasets except these four can be used for SSL pre-training.

[2]Each `monolingual_exp` is made of one experiment with the 10-minute set for training and one with the 1-hour set.

[3]`https://github.com/espnet/espnet/tree/master/egs2/ml_superb/asr1`

Table 2: *Description of the candidate models.*

Model	Params (M)	Pre-Training	
		# Hours	# Langs
wav2vec2-base [3]	95	1k	1
wav2vec2-large [3]	317	60k	1
robust-wav2vec2-large [41]	317	65k	1
wav2vec2-base-23 [19]	95	100k	23
wav2vec2-large-23 [19]	317	100k	23
XLSR-53 [7]	317	56k	53
XLSR-128 [6]	317	400k	128
HuBERT-base [4]	95	1k	1
HuBERT-large [4]	317	60k	1
HuBERT-base-cmn [42]	95	10k	1
HuBERT-large-cmn [42]	317	10k	1
mHuBERT-base [43]	95	14k	3

for LID evaluation, reporting results separately for the normal training set and the few-shot training set.

For overall performance, we use the SUPERB$_s$ metric from the SUPERB benchmark [38]. We denote $s_{t,i}(u)$ as the i^{th} metrics for task t and SSL model u. T is the set of four tasks and I_t is the set of metrics for the task t. SUPERB$_s$ aggregates all task-specific scores $s_t(u)$ with respect to baseline (i.e., FBANK) and state-of-the-art (SOTA) model[4] on the task t. The SUPERB$_s$ is defined as:

$$\text{SUPERB}_s(u) = \frac{1000}{|T|} \sum_t^T \frac{1}{|I_t|} \sum_i^{I_t} \frac{s_{t,i}(u) - s_{t,i}(\text{FBANK})}{s_{t,i}(\text{SOTA}) - s_{t,i}(\text{FBANK})} \quad (1)$$

We expect SUPERB$_s$ can provide a comprehensive view of the model performance on the benchmark and take the difficulty of tasks into consideration.

Analysis support: To facilitate a more comprehensive analysis of the benchmark, we provide various analysis tools. For the multilingual ASR evaluation, we present the character error rate (CER) for each language as well as aggregated scores for different language groups, in addition to the average CER for both normal and few-shot cases. In line with previous studies [39, 40], we also offer visualizations of the learnable layer weights and their learning curve during training.

3. Experiments

3.1. Candidate models

ML-SUPERB welcomes all speech SSL models trained on either monolingual or multilingual data. We believe the analysis of multilingual scenarios for monolingual speech SSLs is also valuable according to previous works [9–11]. In this paper, we show the experimental results of some example model candidates as shown in Table 2.

wav2vec2: wav2vec2 is a popular speech SSL model for speech recognition [3]. Its pre-training uses a contrastive learning approach that prioritizes identifying true quantized latent speech representations over masked time steps from distractors. The wav2vec2 model has also been extended to many other versions for specialized use cases. For example, robust-wav2vec2-large [41] considers the diversity of speech types, such as read speech, conversational speech, and noisy speech, by including additional corpora in the pre-training stage. Wav2vec2-base-23 and wav2vec2-large-23 are pre-trained on Voxpopuli [19], with a focus on European languages. Additionally, XLSR scales up the multilingual training in wav2vec2 by incorporating more languages and data [6, 7].

[4]The SOTA models for each setting are discussed in Sec. 3.2.

HuBERT: HuBERT uses an iterative offline clustering step to generate pseudo labels for each frame. During training, it predicts the pseudo labels of the masked frame, which helps to improve the quality of the learned features. Similar to wav2vec2, HuBERT also has different versions, such as a multilingual HuBERT [43] trained in three European languages (`fra`, `spa`, `eng`) and HuBERT trained on Mandarin [42].

3.2. Experimental Results

The experimental results are shown in Table 3 for 10-minute set and Table 4 for 1-hour set.

Monolingual ASR: In the monolingual ASR task, all speech SSL models outperform the FBANK baseline. XLSR-128 achieves the best performance in the 1-hour set, while HuBERT-large obtains the best performance in the 10-minute set. Several findings are noteworthy: (1) HuBERT-based models outperform wav2vec2-based models when the training data and model size are similar. (2) Large models usually obtain better results than their base versions. (3) While the XLSR series of models deliver impressive performances in the 1-hour set, we have observed their instability in the 10-minute set, particularly on Asian languages such as `cmn`.

Multilingual ASR: In the multilingual ASR task, all models trained using self-supervised learning (SSL) techniques have shown superior performance compared to the baseline model using FBANK features. Among the SSL models, XLSR-128 achieves the best results across all conditions. Our experiments also reveal some interesting findings: (1) Models trained with more languages generally outperform those trained on monolingual datasets, although this may not always be the case. For example, mHuBERT-base performs worse than HuBERT-based models trained on English only. (2) Large models trained on monolingual data do not necessarily have better representations for multilingual scenarios. For instance, HuBERT-large performs worse than HuBERT-base, and wav2vec2-large is less effective than wav2vec2-base. One possible explanation for the lack of performance improvement with larger models is their limited ability to generalize, despite having similar training losses as base models. (3) The robust-wav2vec2-large model achieves decent scores on multilingual ASR, suggesting that our benchmark corpus may need to consider different acoustic environments, as it includes multiple source datasets.

LID: In the LID task, we notice similarities with multilingual ASR, but there are also notable differences. (1) XLSR-128 has been the dominant model for both 10-minute and 1-hour datasets. (2) While most SSL models have improvements over FBANK, some do not, particularly those based on wav2vec2 (e.g., wav2vec2-large-23 for the 10-minute set and wav2vec2-large for the 1-hour set). (3) Larger models with more parameters and pre-trained data do not necessarily lead to better performance compared to base models.

Joint Multilingual ASR + LID: In the joint multilingual ASR+LID task, the results generally align with the other two tasks in the multilingual track. (1) SSL models outperform FBANK on ASR, but some models perform worse on LID. (2) Base models exhibit better generalization ability and often perform better on test sets. (3) There is no single best model that dominates the task, particularly in few-shot cases and LID tasks.

Overall: In terms of overall performance as measured by SUPERB$_s$ in Sec. 2.4, XLSR-128 is the best model for both the 10-minute and 1-hour sets. Major findings include: (1) multilingual training with a broad coverage of languages, as seen in XLSR models that include more than 50 languages, has proven

Table 3: *10-minute set ML-SUPERB benchmark.*

SSL	Monolingual ASR	Multilingual ASR		LID	Multilingual ASR + LID			SUPERB$_s$
		Normal	Few-shot	Normal	Normal		Few-shot	
	CER/PER	CER	CER	ACC	ACC	CER	CER	
FBANK	72.1	62.4	58.3	11.11	35.9	62.0	58.9	0
wav2vec2-base [3]	44.2	43.0	45.7	54.4	66.9	40.6	44.2	712.0
wav2vec2-large [3]	42.0	42.6	45.8	30.9	54.6	45.5	50.3	663.2
robust-wav2vec2-large [41]	44.4	40.1	45.4	50.8	33.1	38.6	44.9	635.9
wav2vec2-base-23 [19]	49.2	37.7	43.4	58.7	45.1	37.2	44.3	660.6
wav2vec2-large-23 [19]	42.0	42.1	44.3	1.1	21.8	43.4	46.1	608.1
XLSR-53 [7]	49.5	33.9	43.6	6.6	45.6	33.4	43.2	692.4
XLSR-128 [6]	39.7	**29.2**	**40.9**	**66.9**	55.6	**28.4**	**42.1**	**886.8**
HuBERT-base [4]	42.8	39.8	44.5	61.2	71.5	39.2	43.8	772.3
HuBERT-large [4]	**38.2**	44.4	48.2	46.5	55.4	45.6	49.3	673.8
HuBERT-base-cmn [42]	43.1	40.8	45.4	49.3	**75.1**	37.7	43.5	771.9
HuBERT-large-cmn [42]	39.4	42.6	45.8	39.5	66.4	41.9	45.2	751.1
mHuBERT-base [43]	41.0	40.5	45.6	52.4	46.6	36.8	44.2	712.0

Table 4: *1-hour set ML-SUPERB benchmark.*

SSL	Monolingual ASR	Multilingual ASR		LID	Multilingual ASR + LID			SUPERB$_s$
		Normal	Few-shot	Normal	Normal		Few-shot	
	CER/PER	CER	CER	ACC	ACC	CER	CER	
FBANK	63.7	59.3	57.4	9.3	43.5	58.6	58.1	0
wav2vec2-base [3]	35.9	35.5	44.3	80.8	83.6	32.1	42.6	743.7
wav2vec2-large [3]	35.4	35.7	43.9	8.0	78.2	34.7	42.2	730.6
robust-wav2vec2-large [41]	35.7	31.1	42.2	72.1	62.9	33.7	46.0	715.7
wav2vec2-base-23 [19]	35.1	32.0	42.2	71.9	66.3	30.9	43.0	744.2
wav2vec2-large-23 [19]	34.2	35.3	42.4	64.2	49.7	35.2	43.1	679.7
XLSR-53 [7]	34.9	26.9	40.6	87.1	76.9	28.6	44.6	813.8
XLSR-128 [6]	**30.6**	**22.0**	**39.3**	**87.9**	85.6	**22.9**	42.4	**937.9**
HuBERT-base [4]	35.3	31.4	42.7	86.1	86.0	30.9	**41.8**	795.2
HuBERT-large [4]	32.2	37.7	43.5	64.1	77.7	35.1	42.2	755.0
HuBERT-base-cmn [42]	35.6	43.2	46.6	85.3	**86.1**	31.8	42.1	700.1
HuBERT-large-cmn [42]	33.7	39.6	45.1	57.3	75.6	37.1	44.4	698.4
mHuBERT-base [43]	33.0	33.4	43.6	72.5	70.9	29.7	43.1	761.6

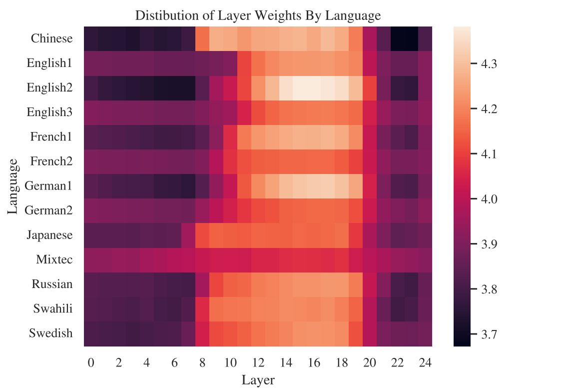

Figure 1: *The layerwise weight analysis of XLSR-128 model in the monolingual track.*

to be useful. However, multilingual training that is limited to a few selective languages may not be as beneficial in larger language groups (e.g., wav2vec2-large-23 and mHUBERT models do not always perform better than their models trained in a single language). (2) The base models tend to generalize better to multilingual cases than their corresponding large versions, such as wav2vec2-base versus wav2vec2-large and HuBERT-base versus HuBERT-large.

3.3. Layerwise analysis

Our benchmark offers tools to guide users in the use of SSL representations according to their needs, including an analysis of the learned weights for layer importance. The results for the XLSR-128 model in monolingual ASR tasks (shown in Fig 1) confirm the conclusions reached by [44] and [45]: the most relevant layers for ASR are not the last few layers. We also observed that English3, French2, and German2 have very similar behavior. These tasks use VoxPopuli data for training, which is the only dataset with lecture speech in our collection. Additionally, Mixtec is the only conversational speech data among our sets, and we can see a distinct behavior in Fig 1. Therefore, the relevance of SSL model layers may be related to the speech domain (in addition to the speech task) rather than the language.

4. Conclusion

This paper introduces ML-SUPERB, a benchmark that extends SUPERB to multilingual tasks. We present the design of the open-source framework and discuss experimental results for some example models. More detailed policies can be found at `https://multilingual.superbbenchmark.org/`. We invite the community to participate in this challenge.

5. References

[1] A. Mohamed *et al.*, “Self-supervised speech representation learning: A review,” *JSTSP*, 2022.

[2] S.-w. Yang *et al.*, “SUPERB: Speech Processing Universal PERformance Benchmark,” in *Proc. Interspeech*, 2021, pp. 1194–1198.

[3] A. Baevski *et al.*, “Wav2vec 2.0: A framework for self-supervised learning of speech representations,” *Proc. NeurIPS*, vol. 33, pp. 12 449–12 460, 2020.

[4] W.-N. Hsu *et al.*, “HuBERT: Self-supervised speech representation learning by masked prediction of hidden units,” *TASLP*, vol. 29, pp. 3451–3460, 2021.

[5] H.-S. Tsai *et al.*, “SUPERB-SG: Enhanced speech processing universal performance benchmark for semantic and generative capabilities,” in *Proc. ACL*, 2022, pp. 8479–8492.

[6] A. Babu *et al.*, “XLS-R: Self-supervised cross-lingual speech representation learning at scale,” *arXiv preprint arXiv:2111.09296*, 2021.

[7] A. Conneau *et al.*, “Unsupervised cross-lingual representation learning for speech recognition,” *arXiv preprint arXiv:2006.13979*, 2020.

[8] P.-A. Duquenne *et al.*, “Speechmatrix: A large-scale mined corpus of multilingual speech-to-speech translations,” *arXiv preprint arXiv:2211.04508*, 2022.

[9] J. Zhao and W.-Q. Zhang, “Improving automatic speech recognition performance for low-resource languages with self-supervised models,” *JSTSP*, vol. 16, no. 6, pp. 1227–1241, 2022.

[10] D. Berrebbi, J. Shi, B. Yan, *et al.*, “Combining Spectral and Self-Supervised Features for Low Resource Speech Recognition and Translation,” in *Proc. Interspeech*, 2022, pp. 3533–3537.

[11] A. Wu *et al.*, “Self-supervised representations improve end-to-end speech translation,” *Proc. Interspeech 2020*, pp. 1491–1495, 2020.

[12] X. Li *et al.*, “ASR2K: Speech Recognition for Around 2000 Languages without Audio,” in *Proc. Interspeech*, 2022, pp. 4885–4889.

[13] S. Evain *et al.*, “ LeBenchmark: A Reproducible Framework for Assessing Self-Supervised Representation Learning from Speech,” in *Proc. Interspeech*, 2021, pp. 1439–1443.

[14] T. Javed *et al.*, “Indicsuperb: A speech processing universal performance benchmark for indian languages,” *arXiv preprint arXiv:2208.11761*, 2022.

[15] A. Conneau *et al.*, “XTREME-S: Evaluating Cross-lingual Speech Representations,” in *Proc. Interspeech*, 2022, pp. 3248–3252.

[16] V. Pratap *et al.*, “MLS: A large-scale multilingual dataset for speech research,” *Proc. Interspeech 2020*, pp. 2757–2761, 2020.

[17] R. Ardila *et al.*, “Common voice: A massively-multilingual speech corpus,” in *Proc. LREC*, 2020, pp. 4218–4222.

[18] K. MacLean, “Voxforge,” *Ken MacLean.[Online]. Available: http://www. voxforge. org/home.[Accessed by 2022]*, 2018.

[19] C. Wang *et al.*, “VoxPopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation,” in *Proc. ACL*, 2021, pp. 993–1003.

[20] K. Sodimana *et al.*, “A step-by-step process for building tts voices using open source data and framework for bangla, javanese, khmer, nepali, sinhala, and sundanese,” in *Proc. SLTU*, 2018, pp. 66–70.

[21] O. Kjartansson *et al.*, “Open-source high quality speech datasets for basque, catalan and galician,” in *Proc. SLTU*, 2020, pp. 21–27.

[22] F. He *et al.*, “Open-source multi-speaker speech corpora for building gujarati, kannada, malayalam, marathi, tamil and telugu speech synthesis systems,” in *Proc. LREC*, 2020, pp. 6494–6503.

[23] G. Rehm and H. Uszkoreit, “Language technology support for norwegian,” in *The Norwegian Language in the Digital Age: Bokmalsversjon*, 2012, pp. 52–70.

[24] A. Conneau *et al.*, “Fleurs: Few-shot learning evaluation of universal representations of speech,” in *Proc. SLT*, 2023, pp. 798–805.

[25] E. Barnard *et al.*, “The nchlt speech corpus of the south african languages,” 2014.

[26] T. Baumann, A. Köhn, and F. Hennig, “The spoken wikipedia corpus collection: Harvesting, alignment and an application to hyperlistening,” *LREC*, vol. 53, pp. 303–329, 2019.

[27] J. Shi *et al.*, “Leveraging end-to-end asr for endangered language documentation: An empirical study on yolóxochitl mixtec,” in *Proc. ACL*, 2021, pp. 1134–1145.

[28] J. Shi *et al.*, “Highland puebla nahuatl speech translation corpus for endangered language documentation,” in *Proc. AmericaNLP*, 2021, pp. 53–63.

[29] I. Solak, “M-ailab speech dataset,” *Imdat Solak.[Online]. Available: https://www.caito.de/2019/01/03/the-m-ailabs-speech-dataset/.[Accessed by 2022]*, 2018.

[30] D. A. Braude *et al.*, “All together now: The living audio dataset.,” in *INTERSPEECH*, 2019, pp. 1521–1525.

[31] N. J. De Vries *et al.*, “A smartphone-based asr data collection tool for under-resourced languages,” *Speech communication*, vol. 56, pp. 119–131, 2014.

[32] S. Watanabe, T. Hori, and J. R. Hershey, “Language independent end-to-end architecture for joint language identification and speech recognition,” in *Proc. ASRU*, 2017, pp. 265–271.

[33] W. Hou *et al.*, “Large-Scale End-to-End Multilingual Speech Recognition and Language Identification with Multi-Task Learning,” in *Proc. Interspeech*, 2020, pp. 1037–1041.

[34] C. Zhang *et al.*, “Streaming End-to-End Multilingual Speech Recognition with Joint Language Identification,” in *Proc. Interspeech*, 2022, pp. 3223–3227.

[35] W. Chen *et al.*, “Improving massively multilingual ASR with auxiliary CTC objectives,” *Proc. ICASSP 2023*, 2023.

[36] T. Wolf *et al.*, “Transformers: State-of-the-art natural language processing,” in *Proc. EMNLP*, 2020, pp. 38–45.

[37] S. Watanabe *et al.*, “ESPnet: End-to-end speech processing toolkit,” in *Proc. Interspeech*, 2018, pp. 2207–2211.

[38] T.-h. Feng *et al.*, “SUPERB@ SLT 2022: Challenge on generalization and efficiency of self-supervised speech representation learning,” in *Proc. SLT*, 2023, pp. 1096–1103.

[39] X. Chang *et al.*, “An exploration of self-supervised pretrained representations for end-to-end speech recognition,” in *Proc. ASRU*, 2021, pp. 228–235.

[40] S. Chen *et al.*, “WavLM: Large-scale self-supervised pre-training for full stack speech processing,” *JSTSP*, vol. 16, no. 6, pp. 1505–1518, 2022.

[41] W.-N. Hsu *et al.*, “Robust wav2vec 2.0: Analyzing Domain Shift in Self-Supervised Pre-Training,” in *Proc. Interspeech*, 2021, pp. 721–725.

[42] S. Liu and P. Guo. “Chinese speech pretraining.” (2023), [Online]. Available: `https://github.com/TencentGameMate/chinese_speech_pretrain` (visited on 06/30/2022).

[43] A. Lee *et al.*, “Textless speech-to-speech translation on real data,” in *Proc. NAACL*, 2022, pp. 860–872.

[44] A. Pasad, B. Shi, and K. Livescu, “Comparative layer-wise analysis of self-supervised speech models,” *Proc. ICASSP 2023*, 2022.

[45] D. Berrebbi, B. Yan, and S. Watanabe, “Avoid overthinking in self-supervised models for speech recognition,” *Proc. ICASSP 2023*, 2022.