ML-SUPERB 2.0: Benchmarking Multilingual Speech Models Across Modeling Constraints, Languages and Datasets

Jiatong Shi[1], Shih-Heng Wang[2]*, William Chen[1]*, Martijn Bartelds[3]*, Vanya Bannihatti Kumar[1], Jinchuan Tian[1], Xuankai Chang[1], Dan Jurafsky[3], Karen Livescu[3, 4], Hung-yi Lee[2], Shinji Watanabe[1]

[1] Carnegie Mellon University, [2] National Taiwan University,
[3] Stanford University, [4] Toyota Technological Institute at Chicago

* for equal contribution.

Content

- Background and Motivation
- Investigation Details
- Experiments
- Conclusion

Background: Multilingual Speech Processing Benchmark

- Recent multilingual speech processing models
    - Have extended their capbilities to **thousands of languages**

Background: Multilingual Speech Processing Benchmark

- Recent multilingual speech processing models
  - Have reached their power to **thousands of languages**
  - However, they also **raise concerns in evaluation** when the model tested in different experimental setups.

Background: Multilingual Speech Processing Benchmark

- Recent multilingual speech processing models
  - Have reached their power to **thousands of languages**
  - However, they also **raise concerns in evaluation** when the model tested in different experimental setups.

→This results in an increasing need for
multilingual speech processing benchmarks

Background: Multilingual Speech Processing Benchmark

We observe great efforts in the community on spoken multilingual benchmarks:

- XTREME-S (Conneau et al. 2022)
- CL-MASR (Libera et al. 2023)
- IndicSUPERB (Javed et al. 2023)
- ML-SUPERB (Shi et al. 2023)

Background: Multilingual Speech Processing Benchmark

- We observe great efforts in the community on spoken multilingual benchmarks:
  - XTREME-S (Conneau et al. 2022)
  - CL-MASR (Libera et al. 2023)
  - IndicSUPERB (Javed et al. 2023)
  - ML-SUPERB (Shi et al. 2023)

- ML-SUPERB is the most comprehensive benchmark in terms of language coverage, including **143 languages** on
  - Monolingual/multilingual automatic speech recognition (ASR)
  - Language identification (LID)
  - Joint ASR + LID

SUPERB

Motivation in Benchmark Extension

- Strictly constrained benchmark settings with self-supervised learning (SSL) pre-trained models
  - Efficient yet not generalizable enough to various settings (Zaiem et al. 2023; Arora et al. 2024)
- <u>**Flexible constraints**</u> are **needed** to understand and benchmark recent and future modeling in multilingual speech processing.

Introduction of ML-SUPERB 2.0

- A revisit to ML-SUPERB for more various scenarios with a deeper understanding
  - By **relaxing the fixed constraints** in the ML-SUPERB original version
  - By **enriching the evaluation metrics on robustness** across languages and **variations** across datasets.

Introduction of ML-SUPERB 2.0 (Cont’d)

- In this paper, we exemplify ML-SUPERB 2.0 benchmarking by investigating **four new scenarios** that original ML-SUPERB does not consider:
  - Large downstream models
  - SSL model fine-tuning
  - Efficient model adaptation
  - Supervised pre-trained models

Investigation Details

- **Large downstream models**
- SSL model fine-tuning
- Efficient model adaptation
- Supervised pre-trained models

- Two frameworks:
  - CTC-based (CTC)
  - Hybrid CTC/attention-based (ATT-CTC)
- Three model architectures:
  - Transformer (Vaswani et al. 2017)
  - Conformer (Gulati et al. 2020)
  - E-Branchformer (Kim et al. 2023)

Investigation Details

- Large downstream models
- **SSL model fine-tuning**
- Efficient model adaptation
- Supervised pre-trained models

- Two frameworks:
  - CTC-based (CTC)
  - Hybrid CTC/attention-based (ATT-CTC)
- Two fine-tuning schedules:
  - Full fine-tuning
  - Partial fine-tuning

Investigation Details

- Large downstream models
- SSL model fine-tuning
- **Efficient model adaptation**
- Supervised pre-trained models

- Two frameworks:
    - CTC-based (CTC)
    - Hybrid CTC/attention-based (ATT-CTC)
- Two model adaptation strategies:
    - Adapter (Housbly et al. 2019)
    - Low-rank Adaptation (LoRA) (Hu et al. 2021)

Investigation Details

- Large downstream models
- SSL model fine-tuning
- Efficient model adaptation
- **Supervised pre-trained models**

- Two frameworks:
  - CTC-based (CTC)
  - Hybrid CTC/attention-based (ATT-CTC)
- Two supervised models
  - Whisper (Radford et al. 2023)
  - OWSM 3.1 (Peng et al. 2024)

Experimental Design (General Setup)

- Dataset
  - We updated ML-SUPERB dataset by **correcting** some mistakes* in the old versions (8th version in release)

- Some statistics
  - 142 languages across 15 datasets
  - Around 300 hours in total (with 85 hours for validation and test sets)
  - Follow 1-hour configuration in the first version ML-SUPERB

* Please refer to our paper for details updates to the dataset

Experimental Design (General Setup)

- Experimental codebases:
    - ESPnet (Watanabe et al. 2018)
    - S3PRL (Yang et al. 2021)
- Selected pre-trained self-supervised models
    - XLS-R (Babu et al. 2022)
    - MMS (Pratap et al. 2024)
- Additional practice in ML-SUPERB 2.0:
    - the 100 million tunable parameters during ML-SUPERB 2.0 training

Experimental Design (Scenario Setups)

- While aligning with the general setup, for the four testing scenarios:
  - We follow the architecture hyperparameter selections from prior works*
  - We additionally tune learning rate and select the best performing model on validation sets

* Please refer to our paper for complete list of prior works we refer to.

Experimental Design (Evaluation)

- Base metrics:
  - Accuracy for LID
  - Character error rate (CER) for ASR in two subsets (normal and few-shot training set)

- **Enhanced evaluation:**
  - Macro-average over languages/datasets instead of micro-average CER
  - Standard deviation of CER across languages
  - Additional consideration on worst-performing languages
  - Additional evaluation of the CER range across datasets for the same languages

Experimental Results and Discussions

- Effect of introducing four additional scenarios
- Model ranking over different configurations
- Supervised ASR versus SSL pre-trained models
- Variations across languages and datasets

Due to the time limits, we present part of results in the presentation. Please refer to our paper for full details.

Effect of Introducing Four Scenarios

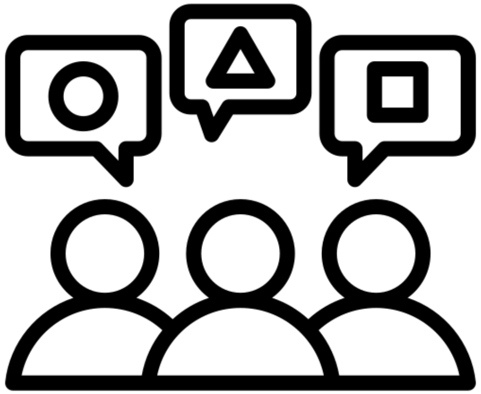
Scenarios	Details	Accuracy	CER (Normal)
Original SUPERB	MMS + Transformer CTC	90.3	24.7 ± 12.3
Large Downstream	MMS + E-Branchformer ATT-CTC	95.2	16.6 ± 11.8
SSL Model Fine-tuning	MMS + 9-14 layers partial fine-tuning CTC	**95.6**	**15.5 ± 10.3**
Efficient Model Adaptation	MMS + LoRA + Transformer ATT-CTC	94.2	18.7 ± 11.5
Supervised Pre-trained Model	Whisper Encoder + Transformer CTC	91.7	21.0 ± 12.5

Compared to the best-performing models in each scenario for multilingual ASR,

in ALL scenarios, we observe **better performance** in LID and ASR (normal).

Model Ranking over Different Configurations

- In original ML-SUPERB,
  - XLS-R reaches better performance in LID
  - MMS achieves better performance in ASR



- However, in **different** training settings, the ranking of upstream models can be **different**

Model Ranking over Different Configurations (Large Downstream Models)

	Transformer	Conformer	E-Branchformer
CTC	XLS-R	MMS	XLS-R
ATT-CTC	MMS	MMS	MMS

XLS-R wins

MMS wins

Compared to original ML-SUPERB,
Different ranks in XLS-R and MMS are observed in the large downstream models

Model Ranking over Different Configurations (Model Fine-tuning)

	Bottom	Middle	Top
CTC	MMS	MMS	MMS
ATT-CTC	MMS	MMS	MMS

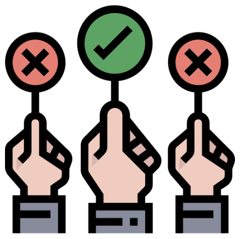
XLS-R wins

MMS wins

Compared to original ML-SUPERB and large downstream models,
Different ranks in XLS-R and MMS are observed in the model fine-tuning

Model Ranking over Different Configurations (Efficient Model Adaptation)

	LoRA	**Adapter**
CTC	XLS-R	XLS-R
ATT-CTC	MMS	MMS



XLS-R wins

MMS wins

Compared to previous experimental settings,
Different ranks in XLS-R and MMS are observed in the efficient adaptation

Supervised ASR vs. SSL Pre-trained Models

Variations across Languages and Datasets

Conclusion of ML-SUPERB 2.0

- Proposing **an updated benchmark** for multilingual speech pre-trained models, built upon and extends ML-SUPERB.
- Investigating **four scenarios** that ML-SUPERB does not consider.
- Introducing **enhanced evaluation metrics** with dataset variation description measures.

Findings of ML-SUPERB 2.0

- All four extended scenarios **show improvements** over the models in the original ML-SUPERB.

- Model fine-tuning achieves **the best performance** on both LID and multilingual ASR tasks.

- We suggest **additional attention** to language/dataset robustness from the experiments.

Acknowledgements

• The work used the Bridges2 system at PSC ad Delta system at NCSA.

Images are generated by DALL-E or directly from Flaticon.com