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

The ML-SUPERB 2.0 Challenge: Towards Inclusive ASR Benchmarking for All Language Varieties

William Chen[1], *Chutong Meng*[2], *Jiatong Shi*[1], *Martijn Bartelds*[3], *Shih-Heng Wang*[1], *Hsiu-Hsuan Wang*[4], *Rafael Mosquera*[5,8], *Sara Hincapie*[5,8], *Dan Jurafsky*[3], *Antonis Anastasopoulos*[2,7], *Hung-yi Lee*[4], *Karen Livescu*[6], *Shinji Watanabe*[1]

[1]Carnegie Mellon University, [2]George Mason University, [3]Stanford University, [4]National Taiwan University, [5]ML Commons, [6]TTI-Chicago, [7]Archimedes, Athena RC, [8]Factored AI,

williamchen@cmu.edu

Abstract

Recent improvements in multilingual ASR have not been equally distributed across languages and language varieties. To advance state-of-the-art (SOTA) ASR models, we present the Interspeech 2025 ML-SUPERB 2.0 Challenge. We construct a new test suite that consists of data from 200+ languages, accents, and dialects to evaluate SOTA multilingual speech models. The challenge also introduces an online evaluation server based on DynaBench, allowing for flexibility in model design and architecture for participants. The challenge received 5 submissions from 3 teams, all of which outperformed our baselines. The best-performing submission achieved an absolute improvement in LID accuracy of 23% and a reduction in CER of 18% when compared to the best baseline on a general multilingual test set. On accented and dialectal data, the best submission obtained 30.2% lower CER and 15.7% higher LID accuracy, showing the importance of community challenges in making speech technologies more inclusive.

Index Terms: multilingual, speech recognition

1. Introduction

In the past decade, studies on scaling end-to-end neural networks have led to dramatic improvements in models for Automatic Speech Recognition (ASR) [1, 2]. Importantly, ASR systems are no longer limited to solely the English language: state-of-the-art (SOTA) models achieve strong performance on over 50 languages [3–5].

However, these benefits have not been equally distributed among all languages and language varieties. Many works have shown that SOTA models perform significantly worse on accented speech or dialects that are not considered "standard" [6–8]. This causes many downstream technologies, such as virtual assistants, closed captioning, and translation services to function substantially worse for many subgroups that may already be disadvantaged.

To address this limitation, this paper presents the Interspeech 2025 ML-SUPERB 2.0 Challenge. The goal of the challenge is to encourage the development of systems that can perform well across languages and language varieties. In total, the challenge evaluates ASR models on speech data from *149 languages and 93 language varieties*. Unlike other SUPERB-based benchmarks and challenges [9–15], which focused on smaller probe-based experiments on limited amounts of data, the ML-SUPERB 2.0 Challenge has no restrictions on what data or pre-trained models can be used. Participants can curate training data or leverage the latest advances in speech modeling to train the best possible models. Evaluation of these models is facilitated by our online evaluation server that is hosted by DynaBench [16]. This server performs inference for the participants by having all submissions adhere to an API. As such, the test set remains fully hidden to participants, preventing benchmark overfitting.

The focus on robustness to *both* different languages and language varieties is novel, as research in these areas has often been disjoint. Many works aim to solely increase the amount of languages a system can handle, but ignore variation that occurs within languages [5, 17, 18]. On the other hand, research on improving ASR performance on different accents or dialects often focuses on variation within single languages [6, 19–21]. This benchmark acts as a first step towards better unifying the fields of *multilingual* and *fair* speech processing. We hope such cross-disciplinary interactions can lead to both a more inclusive speech processing community and more inclusive speech processing models. This paper is outlined as follows:

- Section 2 details the collection and cleaning process for the data used in the challenge, while discussing the difficulties in developing a benchmark with such a wide coverage of languages and language varieties.
- Section 3 outlines the challenge rules and design decisions.
- Section 4 presents baseline results with SOTA self-supervised and supervised ASR models, and compares them to the submitted systems.

Our contributions can thus be summarized as:

1. We introduce a new challenge that evaluates multilingual ASR performance across 149 languages and 93 language varieties, representing the broadest coverage of any speech benchmark to date.
2. We compare 5 submitted systems, which all out-performed our baseline systems, showing that community challenges can lead to better-performing systems.
3. Despite these advancements, we find that SOTA ASR systems continue to underperform on accented and dialectal speech.

2. Challenge Data

2.1. General Multilingual Data

The data described in this section is designed to evaluate the general multilingual capabilities of ASR models across 149 languages. We obtain this data by combining previous ML-SUPERB benchmarks [11, 13, 14]. In doing so, we found several issues with the data used in these existing benchmarks or the corpora they were sourced from. Many of these issues originate from cases where a language may use several different orthographies. For example, we removed all Norwegian data (`nor`, `nno`, and `nob`), because (1) some data sources do not clearly indicate which written standard they represent; (2) `nno` and `nob` differ only in writing systems and they do not have a clear relationship with the spoken dialects. Similar issues were

10.21437/Interspeech.2025-1013

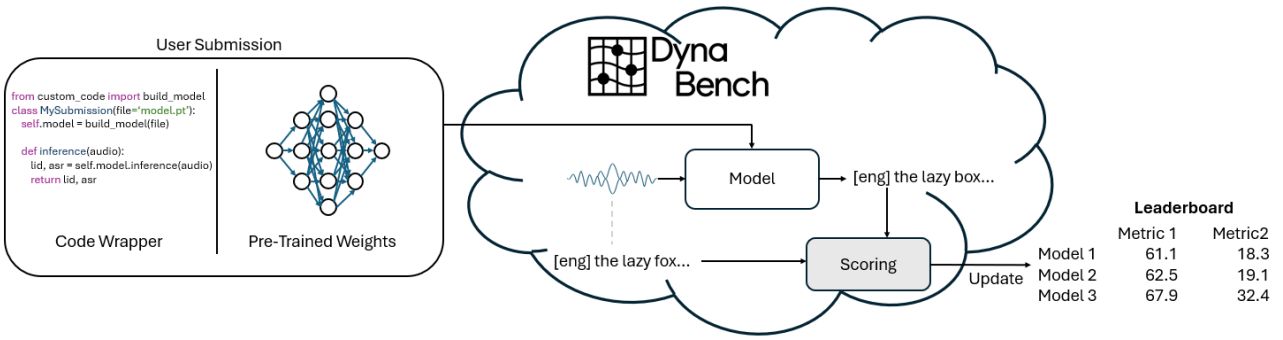

Figure 1: ***Overview of the challenge submission system.*** *Participants upload their model weights and inference code to DynaBench, which runs the evaluation online and returns the evaluation metrics.*

observed in Serbian: FLEURS [22] had Latin transcripts while Common Voice [23] used Cyrillic. Normalizing these factors is particularly important for our challenge, as otherwise our metrics on fairness and robustness will be severely impacted (Section 3.6). Another common problem was the distinction between similar languages. Although Tagalog (`tgl`) and Filipino (`fil`) are technically distinct languages, the source datasets do not explicitly distinguish them. Therefore, we decided to merge them. Also, the choice of ISO code used by different datasets posed a problem, where some datasets use the macro-language designation and others use a dialect-specific code, as seen with Malay (`msa` vs `zlm`) and Odia (`ory` vs `ori`). These issues highlight the difficulties associated with crowd-sourced data collection efforts [22, 23] and show the necessity of linguistic expertise during data development. As multilingual models become increasingly more common in both research and production [3, 5, 17, 18, 24], proper linguistic meta-data is necessary to guarantee that models are configured properly.

After data cleaning, we segment the data into 3 splits: a baseline training set, a development set, and a test set. The baseline training set consists of 1 hour of data per language, while the development and test sets contain 10 minutes of data per language. The baseline training and development sets contain 139 languages, leading to a total of 139 hours of training data and 23 hours of development data. The test set contains an additional 10 languages from the ML-SUPERB Challenge hidden set [13], totaling to 26 hours[1].

2.2. Accented and Dialectal Multilingual Data

To evaluate the performance of ASR models on different accents and dialects, we collect more evaluation data from new data sources, *totaling to 93 accents and dialects*. We create 10-minute development and test subsets for each accent or dialect from each of these corpora in the same style as Section 2. The development set contains data sourced from 9 accented or dialectal speech corpora [25–33]. The hidden test set contains data sourced from the same corpora as the development set along with 4 additional corpora [34–37]. While we list these datasets here for transparency, participants were not made aware of which datasets were used during the challenge. We also did not reveal which accents or dialects were included in the challenge.

[1]Participants were made aware of which exact languages were included in the test set but not the training set.

Table 1: ***Comparison of different ML-SUPERB versions.***

Version	Langs.	Models	Fairness	Dialectal
1.0 Benchmark	143	SSL	✗	✗
1.0 Challenge	154	SSL	✗	✗
2.0 Benchmark	142	ASR+SSL	✓	✗
2.0 Challenge (ours)	149	Any	✓	✓

3. Challenge Task and Rules

3.1. Updates from Previous Challenges and Benchmarks

The goal of this challenge is to encourage the development of ASR systems that are robust to languages, accents, and dialects. Importantly, we avoid constraining participants to certain datasets or modeling approaches. This is distinct from the goals of previous ML-SUPERB benchmarks or challenges [15]. The original benchmark constrained systems to a fixed training set, as it was designed to probe the multilingual capabilities of SSL models and not develop SOTA ASR systems. The ML-SUPERB 1.0 Challenge [13] was proposed to extend the benchmark to new languages, but enforced the same training constraints. Finally, the ML-SUPERB 2.0 benchmark[2] [14] was designed to consider differences in architectures and training strategies when evaluating multilingual SSL models and proposed metrics to evaluate model robustness. However, it was also constrained to a fixed training set and did not consider robustness to accents or dialects. The differences are summarized in Table 1.

3.2. Task

Contrary to previous SUPERB benchmarks [9–15], the ML-SUPERB 2.0 Challenge only features a single evaluation track. Participants are tasked with developing SOTA multilingual ASR systems for 154 languages. During the evaluation phase of the challenge, the participants are given audio files and asked to both transcribe the speech (ASR) and predict the language being spoken (LID prediction). We opted not to include more complex tasks, such as Spoken Language Understanding, as it would lead to many low-resource languages being excluded.

3.3. Submission

We use DynaBench [16] as the challenge submission platform, as it was successfully used in previous competitions that required the test set to remain completely hidden, such as the FLoReS-101 Large-Scale MT challenge at WMT 2021 [38, 39].

[2]We emphasize that [14] is a *benchmark* while this paper implements its findings into a *challenge*.

Listing 1 Pseudo-code for the API used to interface with the DynaBench, which participants customize for their model.

```
def API(waveform):
    """
    Args:
        waveform (np.array): speech waveform
    Returns:
        pred_lid (str): ISO3 code of LID pred
        pred_asr (str): predicted transcript
    """

    # participants must fill in the below
    # lines to connect their model
    pred_lid =
    pred_asr =
    return pred_lid, pred_asr
```

Rather than submitting decoded inference results, participants instead upload their model submission to DynaBench. DynaBench interfaces with the uploaded model via a standardized, pre-defined API (Listing 1) and then performs inference and scoring on the evaluation server. Participants are only able to view the final evaluation scores, and do not have access to the input audio, ground truth text, nor their model outputs. This allows the evaluation to remain fully blind and robust, alleviating issues with community benchmark overfitting. A summary of the submission process is visualized in Figure 1.

3.4. Allowed Models and Methods

Unlike the ML-SUPERB 1.0 challenge that constrained participants to fixed data or training methods [13], participants are allowed to use almost any resource in this benchmark. The only rule is that inference for the final evaluation must be performed on our server without internet connection (See Section 3.3). Furthermore, the submitted model must be able to perform inference within the server's memory limitations (around 46 GB of GPU VRAM). These flexible constraints allow participants to use the latest pre-trained models, such as self-supervised speech encoders [17, 18], supervised ASR models [3, 4, 24, 40], or even Large Language Models (LLMs) [41], while preventing submissions that only use API-based models [42]. We note that participants are allowed to use API-based solutions to aid their own model development, such as for the purpose of distillation or pseudo-labelling.

3.5. Training and Development Data

Similar to Section 3.4, we do not have any explicit constraints on which datasets are allowed: participants are free to use an existing dataset for model development. They are also allowed to collect external data from new resources, as long as they mention how that data was obtained. To ease model development, we provide participants with the baseline training and development set based on the ML-SUPERB 2.0 public set [14], as described in Section 2.1. We also provide participants with a secondary accented and dialectal development set (Section 2.2) to help them assess the robustness of their models to different language varieties. Participants are not required to use any of the provided data.

3.6. Evaluation Data and Metrics

The submissions of participants are evaluated on the accuracy of their speech recognition systems on different languages and language varieties. We separate our metrics into three categories: standard, language robustness, and dialectal robustness. Each of these categories are expanded upon in detail below.

Standard: This category mostly follows the original scoring system of ML-SUPERB's multilingual ASR+LID task with two main metrics: average language identification accuracy (ACC) and ASR character error rate (CER) across languages. The scores for this category are obtained from the inference results on the ML-SUPERB test sets discussed in Section 2.1.

Language Robustness: The language robustness metrics are designed to encourage consistency in performance across languages.[3] We use two metrics to measure this: average ASR CER of a system's worst performing 15 languages and the standard deviation (StD) of a system's ASR CER across all languages. These scores are also obtained from the inference results on the ML-SUPERB test sets discussed in Section 2.1. For both metrics, a lower score indicates a better result.

Dialectal Robustness: The dialectal robustness evaluation metrics encourage systems that are robust to different varieties within languages. The metrics used for robustness mirror those of the standard evaluation category: language identification ACC and ASR CER on the accented and dialectal speech data. Participants are not required to identify the specific dialect spoken in an utterance. Since this evaluation requires accent/-dialect labels, the scores are obtained from inference results on our new accented and dialectal test set discussed in Section 2.2.

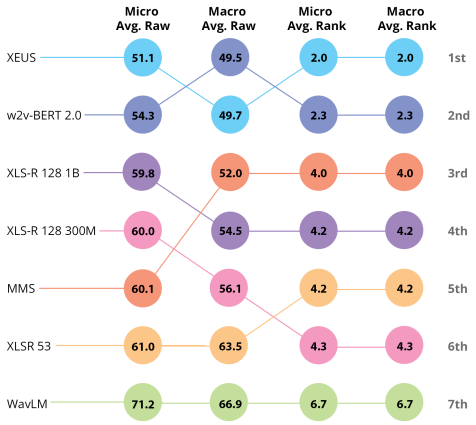

Figure 2: ***Comparison of different aggregation methods used to compute the final rankings across all metrics.*** *Accuracies are inverted by subtracting from 100 when calculating the aggregate raw scores, such that a lower score is better.*

3.7. Ranking

The overall ranking on the challenge is calculated by taking the average rank of each model on each of the 6 metrics. The methodology can be summarized as follows:

1. Calculate rankings for each model on each metric.
2. Calculate the average ranking for each model across metrics.
3. Rank models by average ranking.

We use the average rank of a system rather than the average of the raw metric scores to avoid the ranking being skewed by differences in the metrics' dynamic ranges, which was effective in previous speech challenges [43, 44]. For example, the first two columns in Figure 2 show how rankings can shift significantly between a macro-average across the 3 categories (Standard, Language Robustness, Dialectal Robustness) and micro-

[3]These metrics were first explored in the ML-SUPERB 2.0 benchmark [14], but were not considered for model ranking.

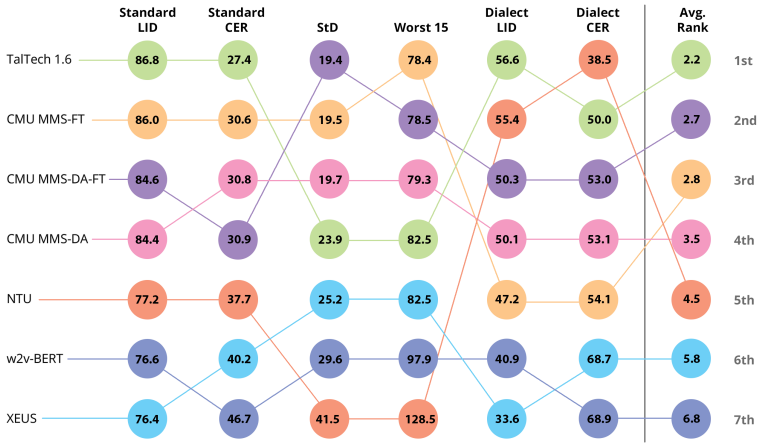

Figure 3: ***Raw scores/rankings of each system submission and the best 2 SSL model baselines on our challenge evaluation metrics.***

Table 2: ***Raw scores of SOTA supervised and SSL models on the ML-SUPERB and Dialectal test sets.***

Model	Standard		Language		Dialectal	
	LID	CER	StD	Worst	LID	CER
WavLM	55.9	72.0	38.3	122.1	27.2	79.0
XLSR 53	53.8	51.9	**19.4**	92.8	18.8	74.5
XLS-R 300M	72.6	53.0	31.4	104.3	30.5	74.5
XLS-R 1B	71.0	53.5	31.4	110.0	32.7	68.0
MMS 1B	74.6	57.1	33.5	118.8	40.4	65.9
w2v-BERT	**76.6**	46.7	29.6	97.9	40.9	68.9
XEUS	76.4	**40.2**	25.2	**82.5**	33.6	68.7
Whisper	42.3	43.2	41.0	122.9	**54.7**	**44.3**
OWSM	64.8	43.0	50.0	122.8	51.0	109.5

average across the 6 individual metrics (Standard CER, Standard LID, Worst 15 CER, StD, Dialectal CER, Dialectal LID) when using raw scores, which is undesirable. On the other hand, there is no change in the overall ranking for the 7 SSL models between a macro-average and micro-average when using per-metric rank (Figure 2, left 2 columns). We therefore use the micro-average across per-metric ranks as the aggregation method for simplicity and consistency, with tie-breakers decided by the micro-average of raw scores.

4. Benchmark and Submission Results

Self-Supervised Models: We obtained baseline results with 7 systems based on SSL speech encoders: WavLM [45], XLSR-53 [46], XLS-R 128 300M [47], XLS-R 128 1B [47], MMS 1B [17], w2v-BERT 2.0 [24], and XEUS [18]. Since all of these models are self-supervised, we develop ASR systems via fine-tuning on the ML-SUPERB 2.0 public set [14]. Each self-supervised speech encoder is frozen during fine-tuning, and a weighted sum of the layer-wise representations is used as input into a 2-layer Transformer encoder [48] that is trained using the CTC loss [49].

Supervised Models: We also evaluate the capabilities of 2 multilingual pre-trained ASR foundation models: OWSM 4B [50] and Whisper Large v3 [3]. These are run in a *zero-shot* manner, as they are designed to be used out-of-the-box.

4.1. Results

Supervised vs Self-Supervised: Table 2 shows the scores of each supervised and SSL model. We find that SSL models generally achieve better scores due to the wide coverage of the fine-tuning dataset; supervised models yield inconsistent performances (high StD) since their original training set does not cover all languages. XEUS is the best overall model, while the English-only WavLM yields the worst results. Table 3 breaks down the two supervised models' scores, separating languages seen in their training set from those that are unseen, showing the substantial impact of unseen languages on CER.

Table 3: ***CERs of supervised models when considering if languages were seen during training.***

Model	Standard		Dialectal	
	Seen	Unseen	Seen	Unseen
Whisper	40.7	76.2	**28.0**	**59.9**
OWSM	**26.0**	**69.3**	102.6	157.7

System Submissions: The 5 systems from the 3 teams are compared to the 2 best SSL baselines in Figure 3. All submitted systems achieved a higher overall ranking than the SSL baselines, showing the progress made during the challenge. No submission obtained the best results across all 6 metrics: each team had a system submission that ranked 1st in at least 1 metric. Comparing the best system per metric to XEUS, the best SSL baseline, we observed an absolute improvement of 12.4, 19.3, 5.8, 4.1, 23.0, and 30.2 on Standard LID, Standard CER, StD, Worst 15 CER, Dialect LID, and Dialect CER, respectively.

5. Conclusion

We propose the ML-SUPERB 2.0 Challenge, a novel speech processing challenge with the goal of inclusive ASR benchmarking for a large number (200+) of languages, accents, and dialects. The challenge introduces a novel multilingual test suite of accented and dialect speech and uses new metrics to test the robustness of ASR systems. The challenge received 5 system submissions from 3 teams, all of which outperformed baselines derived from SOTA SSL models. We hope future work can build on this challenge and its findings to make speech technologies more accessible for all.

Acknowledgments Antonios Anastasopoulos is supported by the National Science Foundation (NSF)under awards IIS-2125466 and IIS-2327143. Parts of this work used the PSC Bridges2 system at and Delta system at NCSA through allocations CIS210014 and IRI120008P from the ACCESS program, supported by NSF grants #2138259,#:2138286, #:2138307, #:2137603, and #:2138296.

6. References

[1] 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.

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

[3] A. Radford *et al.*, "Robust speech recognition via large-scale weak supervision," in *International conference on machine learning*, 2023, pp. 28 492–28 518.

[4] Y. Peng *et al.*, "Owsm v3. 1: Better and faster open whisper-style speech models based on e-branchformer," *arXiv preprint arXiv:2401.16658*, 2024.

[5] Y. Zhang *et al.*, "Google usm: Scaling automatic speech recognition beyond 100 languages," *arXiv preprint arXiv:2303.01037*, 2023.

[6] K. Chang *et al.*, "Self-supervised speech representations still struggle with african american vernacular english," in *INTERSPEECH*, 2024.

[7] S. Feng *et al.*, "Towards inclusive automatic speech recognition," *Computer Speech & Language*, vol. 84, p. 101 567, 2024.

[8] M. Agarwal *et al.*, "Findings of the IWSLT 2023 Evaluation Campaign," in *Proc. IWSLT*, 2023.

[9] H.-S. Tsai *et al.*, "SUPERB-SG: Enhanced Speech processing Universal PERformance Benchmark for Semantic and Generative Capabilities," in *ACL*, 2022.

[10] Y. Tseng *et al.*, "AV-SUPERB: A Multi-Task Evaluation Benchmark for Audio-Visual Representation Models," in *ICASSP*, 2024.

[11] J. Shi *et al.*, "ML-SUPERB: Multilingual Speech Universal PERformance Benchmark," in *Proc. INTERSPEECH*, 2023.

[12] T.-h. Feng *et al.*, "Superb@ slt 2022: Challenge on generalization and efficiency of self-supervised speech representation learning," in *SLT*, 2023, pp. 1096–1103.

[13] J. Shi *et al.*, "Findings of the 2023 ML-SUPERB Challenge: Pre-Training And Evaluation Over More Languages And Beyond," in *ASRU*, 2023.

[14] J. Shi *et al.*, "ML-SUPERB 2.0: Benchmarking Multilingual Speech Models Across Modeling Constraints, Languages, and Datasets," in *Proc. INTERSPEECH*, 2024.

[15] S. Yang *et al.*, "SUPERB: Speech Processing Universal PERformance Benchmark," in *Interspeech*, 2021.

[16] D. Kiela *et al.*, "Dynabench: Rethinking benchmarking in NLP," in *NAACL*, 2021.

[17] V. Pratap *et al.*, "Scaling speech technology to 1,000+ languages," *JMLR*, vol. 25, no. 97, pp. 1–52, 2024.

[18] W. Chen *et al.*, "Towards robust speech representation learning for thousands of languages," *arXiv preprint arXiv:2407.00837*, 2024.

[19] J. Meyer *et al.*, "Artie bias corpus: An open dataset for detecting demographic bias in speech applications," English, in *LREC*, N. Calzolari *et al.*, Eds., 2020, pp. 6462–6468.

[20] S. Weinberger, "Speech Accent Archive," *Retrieved from http://accent.gmu.edu*, 2015.

[21] I.-E. Veliche *et al.*, "Towards measuring fairness in speech recognition: Fair-speech dataset," *arXiv preprint arXiv:2408.12734*, 2024.

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

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

[24] L. Barrault *et al.*, "Seamlessm4t-massively multilingual & multimodal machine translation," *arXiv preprint arXiv:2308.11596*, 2023.

[25] S. Alharbi *et al.*, "Sada: Saudi audio dataset for arabic," in *ICASSP*, 2024, pp. 10 286–10 290.

[26] C. Wang *et al.*, "Voxpopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation," in *ACL*, 2021, pp. 993–1003.

[27] I. Demirsahin *et al.*, "Open-source multi-speaker corpora of the English accents in the British isles," eng, in *LREC*, N. Calzolari *et al.*, Eds., 2020, pp. 6532–6541.

[28] W. Wang, Y. Song, and S. Jha, "Globe: A high-quality english corpus with global accents for zero-shot speaker adaptive text-to-speech," in *Interspeech*, 2024, pp. 1365–1369.

[29] G. Zhao *et al.*, "L2-arctic: A non-native english speech corpus," in *Interspeech*, 2018, pp. 2783–2787.

[30] P. Dogan-Schönberger, J. Mäder, and T. Hofmann, *Swissdial: Parallel multidialectal corpus of spoken swiss german*, 2021.

[31] S. Vakirtzian *et al.*, "Speech recognition for greek dialects: A challenging benchmark," in *Interspeech*, 2024, pp. 3974–3978.

[32] B. M. L. Srivastava *et al.*, "Interspeech 2018 low resource automatic speech recognition challenge for indian languages," in *6th Workshop on Spoken Language Technologies for Under-Resourced Languages (SLTU 2018)*, 2018, pp. 11–14.

[33] A. Guevara-Rukoz *et al.*, "Crowdsourcing Latin American Spanish for low-resource text-to-speech," eng, in *LREC*, N. Calzolari *et al.*, Eds., 2020, pp. 6504–6513.

[34] A. A. B. Abdallah *et al.*, "Leveraging data collection and unsupervised learning for code-switched tunisian arabic automatic speech recognition," in *ICASSP*, 2024, pp. 12 607–12 611.

[35] B. Talafha *et al.*, "Casablanca: Data and models for multidialectal arabic speech recognition," *arXiv preprint arXiv:2410.04527*, 2024.

[36] T. Yu *et al.*, "Automatic speech recognition datasets in Cantonese: A survey and new dataset," in *LREC*, N. Calzolari *et al.*, Eds., 2022, pp. 6487–6494.

[37] M. Hämäläinen *et al.*, "Finnish dialect identification: The effect of audio and text," in *EMNLP*, M.-F. Moens *et al.*, Eds., 2021, pp. 8777–8783.

[38] G. Wenzek *et al.*, "Findings of the WMT 2021 shared task on large-scale multilingual machine translation," in *Proc. WMT*, L. Barrault *et al.*, Eds., 2021, pp. 89–99.

[39] F. Akhbardeh *et al.*, "Findings of the 2021 conference on machine translation (WMT21)," in *Proc. WMT*, L. Barrault *et al.*, Eds.

[40] Y. Peng *et al.*, "OWSM v3.1: Better and Faster Open Whisper-Style Speech Models based on E-Branchformer," in *Proc. INTERSPEECH*, 2024.

[41] H. Touvron *et al.*, "LLaMA: Open and efficient foundation language models," *arxiv:2302.13971*, 2023.

[42] T. Brown *et al.*, "Language models are few-shot learners," in *Proc. NeurIPS*, vol. 33, 2020, pp. 1877–1901.

[43] X. Chang *et al.*, "The interspeech 2024 challenge on speech processing using discrete units," *arXiv preprint arXiv:2406.07725*, 2024.

[44] W. Zhang *et al.*, "NeurIPS 2024 competition proposal: URGENT challenge," in *NeurIPS 2024 Competition Track*, 2024.

[45] 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.

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

[47] A. Babu *et al.*, "Xls-r: Self-supervised cross-lingual speech representation learning at scale," *arXiv preprint arXiv:2111.09296*, 2021.

[48] A. Vaswani *et al.*, "Attention is all you need," in *NeurIPS 2017*, 2017.

[49] A. Graves *et al.*, "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks," in *ICML 2006*, 2006, pp. 369–376.

[50] W. Chen *et al.*, "OWLS: Scaling Laws for Multilingual Speech Recognition and Translation Models," *arXiv preprint arXiv:2502.10373*, 2025.