Interspeech 2022
18-22 September 2022, Incheon, Korea

XTREME-S: Evaluating Cross-lingual Speech Representations

Alexis Conneau[△], *Ankur Bapna*[△], *Yu Zhang*[△], *Min Ma*[△], *Patrick von Platen*[♣], *Anton Lozhkov*[♣], *Colin Cherry*[△], *Ye Jia*[△], *Clara Rivera*[△], *Mihir Kale*[△], *Daan Van Esch*[△], *Vera Axelrod*[△], *Simran Khanuja*[△], *Jonathan H. Clark*[△], *Orhan Firat*[△], *Michael Auli*[□], *Sebastian Ruder*[△], *Jason Riesa*[△], *Melvin Johnson*[△]

[△] Google Research [♣] Hugging Face [□] Meta AI

{aconneau,ankurbpn,ngyuzh,ruder,riesa,melvinp}@google.com; patrick@huggingface.co

Abstract

We introduce XTREME-S, a new benchmark to evaluate universal cross-lingual speech representations in many languages. XTREME-S covers four task families: speech recognition, classification, speech-to-text translation and retrieval. Covering 102 languages from 10+ language families, 3 different domains and 4 task families, XTREME-S aims to simplify multilingual speech representation evaluation, as well as catalyze research in "universal" speech representation learning. This paper describes the new benchmark and establishes the first speech-only and speech-text baselines using XLS-R and mSLAM on all downstream tasks. We motivate the design choices and detail how to use the benchmark. Datasets and fine-tuning scripts are made easily accessible through the HuggingFace platform.[1]

1. Introduction

In the past two decades, the exploding amount of content on the Internet has led to a pressing urgency to build systems that can understand text, speech, and videos in all of the world's approximately 6,900 languages. Making speech technology available in all languages is especially important to give speakers of under-represented languages an equal voice on the Internet, and the possibility to make their content and culture known outside of their language cluster. Building speech systems for such a large number of languages is especially challenging but recent advances in self-supervised learning (SSL) present great opportunities to achieve this goal.

Speech pre-training techniques like wav2vec 2.0 [1] have emerged as the predominant approach for automatic speech recognition (ASR) and direct speech-to-text translation (ST), and have made speech models much more data efficient: ASR models can be learnt with as little as a few hours of labeled data [2, 3]. Multilingual pre-training helps build better representations for languages that lack unannotated data, and thus enables the same data-efficient strategies for low-resource languages. Approaches like XLS-R [4, 5], for example, have shown particularly strong results on several tasks, including ASR on BABEL and multilingual LibriSpeech, and AST on CoVoST-2. Following a recent trend in natural language processing, the speech community has made these multilingual pre-trained models publicly available to accelerate research in multilingual speech understanding.

To support this rapid development and to make better speech technology available in all languages of the world, the community requires high-quality datasets and a unified evaluation benchmark that is shared across researchers and practitioners. There has been significant progress in the past few years towards building publicly available multilingual evaluation datasets for speech understanding [6, 7, 8]. Many research studies have, however, designed models on different tasks, and evaluated on a small and often disparate set of languages. This makes comparisons across methods difficult, slows down the development of multilingual representations, and hinders the evaluation of the generalization capabilities of such pre-trained models. The goal of this paper is to structure the evaluation of multilingual speech representation learning.

To address these issues and incentivize the rapidly-evolving research on general-purpose multilingual speech representation learning, we introduce XTREME-S, the Cross-lingual Transfer Evaluation of Multilingual Encoders for Speech benchmark. XTREME-S builds on top of the XTREME series of evaluation benchmarks for text understanding, with XTREME [9] and XTREME-R [10], which specialize in the evaluation of multilingual text representations and have helped the community improve multilingual language understanding, with impressive performance improvements on a variety of tasks.

XTREME-S is meant to be a more exhaustive, thorough and complete evaluation of learned speech representations. It covers 102 diverse languages spanning more than 10 language families and includes four different task families: recognition, translation, classification and retrieval. The seven downstream tasks of XTREME-S also cover various domains, from read-speech to parliamentary speech. It also includes a new general-purpose massively multilingual evaluation dataset dubbed Fleurs in all of the 102 languages.

2. Related work

Multilingual representations Self-supervised learning methods like BERT [11], wav2vec 2.0 [1] or w2v-BERT [12] have been extended to the cross-lingual setting through mBERT [11], XLM-R [13] or XLS-R [4, 5]. These methods demonstrate the effectiveness of multilingual understanding in improving low-resource language representation through unsupervised cross-lingual transfer from higher-resource languages. Combined with the few-shot learning capability of wav2vec 2.0 [2], strong self-supervised speech representations can be built in low-resource languages, enabling training speech recognition systems with just a few hours of labeled data. XLS-R models demonstrate data-efficient capabilities in both speech recognition and speech translation for low-resource languages. Recently, mSLAM [14] built a pre-trained multilingual model for both speech and text, leading to strong improvements on speech translation and even better data efficiency in low-resource languages. mSLAM is evaluated on text downstream tasks from XTREME [9] and tasks from our new XTREME-S benchmark.

Multilingual speech evaluation There has been a sig-

[1] https://hf.co/datasets/google/xtreme_s

10.21437/Interspeech.2022-10007

nificant body of work on building trusted multilingual evaluation datasets for speech. IARPA introduced BABEL [15] for evaluating speech models in low-resource languages. This dataset has been widely used in the speech community and covers real-world conversational telephone speech in 17 African and Asian low-resource languages. Recent work revived this dataset with different preprocessing [16, 17, 18, 4]. The CommonVoice effort [19] offers a wide coverage of speech recognition data in more than 70 languages, with read speech of Wikipedia and other sentences. CommonVoice has been used namely for phoneme recognition [20]. The Multilingual LibriSpeech [6] dataset extends the classical LibriSpeech task [21] to seven other European languages. VoxPopuli builds semi-supervised learning data from European Parliament session [7] in 23 languages, and includes speech transcriptions and translations for 16 languages, as well as speech-to-speech translations. With more than 400k hours of unlabeled speech, VoxPopuli is also used as a public pre-training corpus [5, 14]. In speech-to-text translation, CoVoST-2 [8] has become one of the go-to datasets for multilingual evaluation, covering 21 language directions into English and English into 15 languages. Europarl-ST [22], Must-C [23] and mTEDX [24] also provide common evaluation of speech translation. LangID can be evaluated using VoxLingua107 [25] on YouTube data in 107 languages, and CMU Wilderness [26] on New Testament data in 700+ languages. Fleurs is a new multilingual speech understanding evaluation dataset in 102 languages.

Multilingual benchmarks For text understanding, GLUE [27] and SuperGLUE [28] provide common benchmarks for representation learning [11, 29, 30]. Methods like BERT, or T5 leverage GLUE to show the generalization ability of self-supervised learning on a variety of tasks. In the multilingual setting, new evaluation datasets like XNLI [31], MLQA [32] or TyDi QA [33] are grouped in the XTREME benchmarks [9, 10], on which methods like mBERT, XLM-R or mT5 show their generalization capabilities across languages. SUPERB [34] attempts to transpose GLUE to the speech setting, by grouping several common speech tasks to evaluate English speech models while LeBenchmark [35] is designed for the evaluation of French self-supervised speech models. Our new XTREME-S benchmark groups several multilingual speech datasets and is the speech version of XTREME. The choice of tasks in XTREME-S is motivated by several factors explained in this work. Most tasks have been already used in previous work as evaluation for multilingual speech SSL.

3. XTREME-S

In this section, we describe the design decisions we made that led to the choice of tasks, domains and languages for our benchmark. Then we describe task families and their corresponding datasets.

3.1. Design principles

Given XTREME's goal of providing an accessible benchmark for the evaluation of cross-lingual transfer learning on a diverse and representative set of tasks and languages, we select the tasks and languages that make up the benchmark based on the following principles:

Task difficulty Tasks should be sufficiently challenging that they are not saturated by the strongest existing baselines. The data should also be representative of the challenges faced by practitioners, under the constraint that the data should be publicly accessible.

Diversity We aim for task, domain and language diversity. Tasks should be diverse and cover several domains to provide a reliable evaluation of model generalization and robustness to noisy naturally-occurring speech in different environments. Languages should be diverse to ensure that models can adapt to a wide range of linguistic and phonological phenomena. Language coverage should not be unnecessarily large so as to avoid cumbersome evaluations. We note that the tasks are focused particularly on linguistic aspects of speech, while nonlinguistic/paralinguistic aspects of speech relevant to e.g. speech synthesis or voice conversion are not evaluated.

Data efficiency The training sets of XTREME-S range from a few hours to a few hundred hours of labeled data per language. This is a few-shot setting suited for low-resource understanding. XTREME-S strongly encourages data-efficient self-supervised representation learning.

Training efficiency Tasks should be trainable with a reasonable amount of time (few days) and compute (few GPUs). We enforce that constraint by having datasets focused on few-shot learning (e.g. Fleurs or MLS). This is to make the benchmark accessible, in particular to practitioners working under resource constraints. We also minimize the number of required fine-tuning runs where we can, for instance by encouraging multilingual fine-tuning over monolingual fine-tuning.

Monolingual data Unlabeled speech is available publicly through corpora already used in past work (e.g. MLS, VoxPopuli, CommonVoice). Unlabeled text data is available in all languages, for instance, through Common Crawl data as in the mC4 dataset . Speech data is however not abundant for all languages, so multilinguality is important to build strong representations for those languages.

Accessibility Each task should be available under a permissive license that allows the use and redistribution of the data for research purposes. When needed, we provide scripts to download and easily reproduce the preprocessing steps. Tasks have also been selected based on their usage by pre-existing multilingual pre-trained models, for simplicity.

Reproducibility We encourage submissions that leverage publicly available speech and text datasets. Users should detail which data they use. In general, we encourage settings that can be reproduced by the community, but also encourage the exploration of new frontiers for speech representation learning.

3.2. Tasks

We present in this section the four task families of XTREME-S and their corresponding datasets.

3.2.1. Speech Recognition (ASR)

For speech recognition, we use three datasets: Fleurs, MLS and VoxPopuli, which cover more than 100 languages.

Fleurs-ASR Fleurs is the speech version of the FLoRes machine translation benchmark [36]. We use 2009 n-way parallel sentences from the FLoRes dev and devtest publicly available sets, in 102 languages. We collect between one and

Task	Corpus	Train	Dev	Test	Lang.	Fine-tune	Eval	Task	Metric	Domain
Speech recognition	FLEURS	999h	122h	293h	102	Multi	1	ASR	CER	Read-speech
	MLS	80h	10h	10h	8	Multi	1	ASR	WER	Read-speech
	VoxPopuli	1300h	240h	240h	14	Multi	1	ASR	WER	Euro Parl
Speech translation	CoVoST-2	566h	144h	153h	21	Multi	1	AST	BLEU	Read-speech
Speech classification	FLEURS	999h	122h	293h	102	Multi	1	LangID	Acc.	Read-speech
	Minds-14	2h	1h	1h	14	Multi	1	Intent Cl.	Acc.	E-banking
Speech retrieval	FLEURS	49h	6h	14h	5	Either	1/5	Mining	P@K	Read-speech

Table 1: *Characteristics of the datasets in XTREME-S. We report the number of hours for each train, dev and test set, and the number of languages. We specify the type of fine-tuning (monolingual or multilingual), which coincides with the number of fine-tuning runs. We also include the task, the metric and the speech domain.*

three recordings for each sentence (2.3 on average), and build new train-dev-test splits with 1509, 150 and 350 sentences for train, dev and test respectively. Training sets have around 10 hours of supervision. Speakers of the train sets are different than speakers from the dev/test sets. Multilingual fine-tuning is used and "unit error rate" (characters, signs) of all languages is averaged. Languages and results are also grouped into seven geographical areas: Western Europe (WE), Eastern Europe (EE), Central-Asian/Middle-East/North-Africa (CMN), Sub-Saharan Africa (SSA), South Asia (SA), South-Eastern Asia (SEA) and CJK languages (CJK) [37].

MLS The Multilingual LibriSpeech (MLS) dataset is a large corpus derived from read audiobooks of Librivox and consists of 8 languages: *Dutch (nl), English (en), French (fr), German (de), Italian (it), Polish (pl), Portuguese (pt), Spanish (es)*. The latest version of this corpus contains around 50k hours including 44k hours in English. The task consists of the official 10-hour splits provided by [6] to evaluate few-shot learning capabilities. We use multilingual fine-tuning on all languages at once.

VoxPopuli VoxPopuli is a multilingual speech dataset for semi-supervised learning [7] . It contains 400k hours of unannotated speech as well as speech transcriptions and translations. We use the 14 languages with more than 10 hours of data from the ASR task. Models are fine-tuned on all 14 languages at once, ranging from 543 hours of supervision for English to 10 hours for Slovenian. Word Error Rate (WER) is reported. The language modeling data is provided by VoxPopuli.

3.2.2. Speech Translation (ST)

For speech translation, we use all the 21 language pairs into English from the CoVoST-2 dataset.

CoVoST-2 CoVoST-2 is a large-scale multilingual speech translation corpus covering translations from 21 languages into English. This represents the largest open dataset available to date from total volume and language coverage perspective. We consider all languages to English, grouped into high/mid/low labeled data directions. The task has been widely used in recent speech representation learning [5, 14] and has been recently expanded to cover speech-to-speech translation [38].

3.2.3. Speech classification

For speech classification, we include LangID and intent classification. After hyperparameter tuning, we encourage reporting the average result over 5 random seeds.

Fleurs-LangID We use Fleurs as a LangID dataset by using the same train, dev and test splits as used for ASR. We report over classification accuracy over the 102 languages.

Minds-14 MINDS-14 [39] is an intent classification task from spoken data. It covers 14 intents extracted from the e-banking domain, with spoken examples in 14 language varieties. We merge monolingual datasets into a single multilingual dataset, with a 30-20-50% train-dev-test split.

3.3. Languages

Our 102 languages cover various language families and geographical locations, from Western Europe/Americas, Eastern Europe, Central-Asia, Middle-East, North-Africa, Sub-Saharan Africa, South Asia, South-East Asia to CJK languages. We have 36 languages covered by at least two evaluation datasets. The language coverage provides a good estimate of the generalization ability of multilingual models.

4. Results

In this section, we describe our baselines and the corresponding results. We also comment on the specificities of each downstream task and offer remarks on how results can be improved.

4.1. Baselines

We present two baselines. The first is a 600M parameter speech-only pre-trained wav2vec-BERT model trained on 429k unlabeled data in 51 languages from VoxPopuli, MLS, CommonVoice and BABEL, similar to XLS-R. The second is the 600m parameter mSLAM speech-text pre-trained model that leverages the same speech data, as well more than 10TiB of unlabeled text data from mC4 and some ASR supervision. More details on these baselines, including fine-tuning details can be found in [14]. For some tasks, we also report results of the XLS-R models from [5]. If capacity constraints become an issue, we encourage practitioners to use same-capacity apples-to-apples comparisons with the smaller XLS-R (0.3B) and w2v-bert-51 (0.6B) models.

4.2. Speech recognition

In Table 2, we report average unit and word error rates on Fleurs, MLS and VoxPopuli. Pre-trained models obtain strong performance across domains and on both high-data regimes datasets

Model	Speech recognition Fleurs	 MLS	 VoxPopuli	Speech translation CoVoST-2	Speech classification Fleurs-LID	 Minds-14	Speech retrieval Fleurs-R5	Avg
Metrics	UER	WER	WER	BLEU	Acc.	F1	P@1	-
XLS-R (0.3B)	-	12.8	12.8	13.2	-	-	-	
w2v-bert-51 (0.6B)	17.0	9.9	9.3	20.4	71.4	82.7	-	58.7
mSLAM (0.6B)	17.0	10.1	9.2	20.6	73.3	86.9	-	59.4

Table 2: *Table of results for XTREME-S.*

like VoxPopuli as well as low-data regimes tasks like Fleurs and MLS. Overall, mSLAM obtains 17.0 UER on Fleurs, and 10.1 WER on MLS, as well as 9.2 WER on VoxPopuli which cover different domains.

4.3. Speech translation

For most low-resource languages, only a couple of hours are available as supervision. Specifically, w2v-bert-51 (0.6B) obtains 13.4 and mSLAM obtains 15.6 average BLEU on low-resource languages, 35.6 and 36.3 on high-res languages. Overall, those models obtain 20.4 and 20.6 average BLEU respectively on all languages (see Table 2). On this dataset, only one multilingual fine-tuning run is done to simplify the evaluation.

4.4. Speech classification

We report our baselines on the two speech classification datasets in Table 2. We see that the mSLAM model obtains the best performance overall. Each of these datasets only require a single fine-tuning run; we build a multilingual training set from Minds-14 to reduce its inherent variance.

On Minds-14, mSLAM obtains around 86.9% accuracy, and 73.3% accuracy on Fleurs LangID, while w2v-bert-51 (0.6B) obtains 82.7 and 71.4 respectively. We note that on Fleurs-LangID, speakers are different between train sets and dev/test sets. Avoiding overfitting on speaker ID for the LangID task is essential for obtaining good performance.

5. Discussion

In this section, we discuss several components of the XTREME-S benchmark.

On test sets: Test sets are available in open-source and are not hidden to the public. We trust practitioners to perform all hyperparameter search and checkpoint selection on the dev set, and eventually report performance on the test set. Results are however double-checked through the submission of the predictions of the model for each task.

On speech data: We encourage the community to use similar unlabeled speech datasets across submissions when possible to encourage apple-to-apple comparisons across models. We do encourage submissions that also use different unlabeled speech, although preferably only in the case where there is a substantial difference (e.g. much smaller or much larger, or from more diverse sources, or using TTS-augmented data etc). Additional unlabeled speech data can be used for pre-training but also for self-training and other methods.

On text data: The mC4 and Wikipedia datasets should cover all the languages of the XTREME-S benchmark, including low-resource ones. We encourage the use of these datasets for learning language models, for training text-augmented speech models, or using TTS augmentation for example. We hope the community can also develop smarter ways to adapt these very large unlabeled text datasets to each particular task and domain through filtering methods.

On language modeling: The use of language model decoding is allowed. When using LMs, results should also be reported without LM fusion for comparison. The dataset and the type of LM used should be explicitly detailed in submissions and papers for reproducibility. When doing smart filtering of unlabeled text data, the technique should be explained clearly and the data released in open-source when possible.

On the average score: We weight differently each task of the XTREME-S benchmark. Speech recognition and translation each have a weight of 40%, and speech classification has a weight of 20%. The average score is computed in the following way:

$$0.4 * \left(100 - \frac{\text{Fleurs} + \text{MLS} + \text{VP}}{3}\right)_{\text{(WER)}} +$$

$$0.4 * \text{CoVoST-2}_{\text{(BLEU)}} + 0.2 * \left(\frac{\text{F-LID} + \text{M-14}}{2}\right)_{\text{(Acc)}}$$

This is to give more importance to the core recognition and translation tasks.

6. Conclusion

We presented XTREME-S, an evaluation benchmark meant to evaluate the generalization ability of multilingual speech pre-trained models. The benchmark consists of four key task types: recognition, translation, classification and retrieval. In total, XTREME-S covers 102 languages with various language families, from high-resource to low-resource, and different scripts. Tasks cover several domains and data regimes, from a few hours of supervision to more than a thousand hours, and are all directly open-sourced and made easily accessible. We presented two baselines: one speech-only pre-trained model and one speech-text pre-trained model that obtain strong results on each task. We also built a new dataset named Fleurs, in 102 languages, covering many low-resource languages. We hope XTREME-S will enable the community to build better speech representations in many languages, and enable rapid access to data-efficient speech technology for all the world's languages.

7. References

[1] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Proc. of NeurIPS*, 2020.

[2] Q. Xu, A. Baevski, T. Likhomanenko, P. Tomasello, A. Conneau, R. Collobert, G. Synnaeve, and M. Auli, "Self-training and pre-training are complementary for speech recognition," in *ICASSP 2021*. IEEE, 2021, pp. 3030–3034.

[3] A. Baevski, W.-N. Hsu, A. Conneau, and M. Auli, "Unsupervised speech recognition," *NeurIPS 2021*, 2021.

[4] A. Conneau, A. Baevski, R. Collobert, A. Mohamed, and M. Auli, "Unsupervised cross-lingual representation learning for speech recognition," in *Proc. of Interspeech*, 2021.

[5] A. Babu, C. Wang, A. Tjandra, K. Lakhotia, Q. Xu, N. Goyal, K. Singh, P. von Platen, Y. Saraf, J. Pino *et al.*, "Xls-r: Self-supervised cross-lingual speech representation learning at scale," *arXiv preprint arXiv:2111.09296*, 2021.

[6] V. Pratap, Q. Xu, A. Sriram, G. Synnaeve, and R. Collobert, "Mls: A large-scale multilingual dataset for speech research," in *Proc. of Interspeech*, 2020.

[7] C. Wang, M. Riviere, A. Lee, A. Wu, C. Talnikar, D. Haziza, M. Williamson, J. Pino, and E. Dupoux, "VoxPopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation," in *Proc. of ACL*, 2021.

[8] C. Wang, A. Wu, and J. Pino, "Covost 2 and massively multilingual speech-to-text translation," *arXiv*, 2020.

[9] J. Hu, S. Ruder, A. Siddhant, G. Neubig, O. Firat, and M. Johnson, "Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation," in *ICML*, 2020.

[10] S. Ruder, N. Constant, J. Botha, A. Siddhant, O. Firat, J. Fu, P. Liu, J. Hu, G. Neubig, and M. Johnson, "Xtreme-r: Towards more challenging and nuanced multilingual evaluation," *EMNLP*, 2021.

[11] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *Proc. of NAACL*, 2019.

[12] Y.-A. Chung, Y. Zhang, W. Han, C.-C. Chiu, J. Qin, R. Pang, and Y. Wu, "W2v-bert: Combining contrastive learning and masked language modeling for self-supervised speech pre-training," *arXiv preprint arXiv:2108.06209*, 2021.

[13] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov, "Unsupervised cross-lingual representation learning at scale," *ACL 2020*, 2019.

[14] A. Bapna, C. Cherry, Y. Zhang, Y. Jia, M. Johnson, Y. Cheng, S. Khanuja, J. Riesa, and A. Conneau, "mslam: Massively multilingual joint pre-training for speech and text," in *arXiv*, 2022.

[15] M. J. F. Gales, K. M. Knill, A. Ragni, and S. P. Rath, "Speech recognition and keyword spotting for low-resource languages: Babel project research at cued," in *n Spoken Language Technologies for Under-Resourced Languages*, 2014.

[16] T. Alumäe, D. Karakos, W. Hartmann, R. Hsiao, L. Zhang, L. Nguyen, S. Tsakalidis, and R. Schwartz, "The 2016 bbn georgian telephone speech keyword spotting system," in *ICASSP*, 2017.

[17] A. Ragni, Q. Li, M. J. F. Gales, and Y. Wang, "Confidence estimation and deletion prediction using bidirectional recurrent neural networks," in *SLT*, Athens, 2018.

[18] H. Inaguma, J. Cho, M. K. Baskar, T. Kawahara, and S. Watanabe, "Transfer learning of language-independent end-to-end asr with language model fusion," in *ICASSP*, 2019.

[19] R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, and G. Weber, "Common voice: A massively-multilingual speech corpus," *LREC*, 2020.

[20] M. Rivière, A. Joulin, P.-E. Mazaré, and E. Dupoux, "Unsupervised pretraining transfers well across languages," in *ICASSP*, 2020.

[21] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *ICASSP*, 2015.

[22] J. Iranzo-Sánchez, J. A. Silvestre-Cerdà, J. Jorge, N. Roselló, A. Giménez, A. Sanchis, J. Civera, and A. Juan, "Europarl-st: A multilingual corpus for speech translation of parliamentary debates," in *ICASSP 2020*, 2020.

[23] M. A. Di Gangi, R. Cattoni, L. Bentivogli, M. Negri, and M. Turchi, "Must-c: a multilingual speech translation corpus," in *NAACL*, 2019.

[24] E. Salesky, M. Wiesner, J. Bremerman, R. Cattoni, M. Negri, M. Turchi, D. W. Oard, and M. Post, "The multilingual tedx corpus for speech recognition and translation," *arXiv preprint arXiv:2102.01757*, 2021.

[25] J. Valk and T. Alumäe, "Voxlingua107: a dataset for spoken language recognition," in *Proc. of SLT*, 2020.

[26] A. W. Black, "Cmu wilderness multilingual speech dataset," in *ICASSP 2019*, 2019.

[27] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. R. Bowman, "GLUE: A multi-task benchmark and analysis platform for natural language understanding," in *ICLR*, 2019.

[28] A. Wang, Y. Pruksachatkun, N. Nangia, A. Singh, J. Michael, F. Hill, O. Levy, and S. R. Bowman, "Superglue: A stickier benchmark for general-purpose language understanding systems," *arXiv preprint arXiv:1905.00537*, 2019.

[29] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. Salakhutdinov, and Q. V. Le, "Xlnet: Generalized autoregressive pretraining for language understanding," *arXiv*, vol. abs/1906.08237, 2019.

[30] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *JMLR*, 2019.

[31] A. Conneau, R. Rinott, G. Lample, A. Williams, S. R. Bowman, H. Schwenk, and V. Stoyanov, "Xnli: Evaluating cross-lingual sentence representations," in *EMNLP 2018*, 2018.

[32] P. Lewis, B. Oğuz, R. Rinott, S. Riedel, and H. Schwenk, "Mlqa: Evaluating cross-lingual extractive question answering," *arXiv preprint arXiv:1910.07475*, 2019.

[33] J. H. Clark, E. Choi, M. Collins, D. Garrette, T. Kwiatkowski, V. Nikolaev, and J. Palomaki, "Tydi qa: A benchmark for information-seeking question answering in typologically diverse languages," *TACL*, 2020.

[34] S. Yang, P. Chi, Y. Chuang, C. J. Lai, K. Lakhotia, Y. Y. Lin, A. T. Liu, J. Shi, X. Chang, G. Lin, T. Huang, W. Tseng, K. Lee, D. Liu, Z. Huang, S. Dong, S. Li, S. Watanabe, A. Mohamed, and H. Lee, "SUPERB: speech processing universal performance benchmark," *Interspeech*, 2021.

[35] S. Evain, M. H. Nguyen, H. Le, M. Z. Boito, S. Mdhaffar, S. Alisamir, Z. Tong, N. Tomashenko, M. Dinarelli, T. Parcollet *et al.*, "Task agnostic and task specific self-supervised learning from speech with lebenchmark," *arXiv*, 2021.

[36] N. Goyal, C. Gao, V. Chaudhary, P.-J. Chen, G. Wenzek, D. Ju, S. Krishnan, M. Ranzato, F. Guzmán, and A. Fan, "The flores-101 evaluation benchmark for low-resource and multilingual machine translation," in *TACL*, 2021.

[37] A. Conneau, M. Ma, S. Khanuja, Y. Zhang, V. Axelrod, S. Dalmia, J. Riesa, C. Rivera, and A. Bapna, "Fleurs: Few-shot learning evaluation of universal representations of speech," *arXiv preprint arXiv:2205.12446*, 2022.

[38] Y. Jia, M. T. Ramanovich, Q. Wang, and H. Zen, "CVSS corpus and massively multilingual speech-to-speech translation," *arXiv preprint arXiv:2201.03713*, 2022.

[39] D. Gerz, P.-H. Su, R. Kusztos, A. Mondal, M. Lis, E. Singhal, N. Mrkšić, T.-H. Wen, and I. Vulić, "Multilingual and cross-lingual intent detection from spoken data," *arXiv preprint arXiv:2104.08524*, 2021.