XLS-R: Self-supervised Cross-lingual Speech Representation Learning at Scale

Arun Babu[△], Changhan Wang[△*], Andros Tjandra[△], Kushal Lakhotia[◊†], Qiantong Xu[△], Naman Goyal[△], Kritika Singh[△], Patrick von Platen[♣], Yatharth Saraf[△], Juan Pino[△], Alexei Baevski[△], Alexis Conneau[□‡], Michael Auli[△‡]*

[△] Meta AI [□] Google AI [◊] Outreach [♣] Hugging Face

arbabu@fb.com

Abstract

This paper presents XLS-R, a large-scale model for cross-lingual speech representation learning based on wav2vec 2.0. We train models with up to 2B parameters on nearly half a million hours of publicly available speech audio in 128 languages, an order of magnitude more public data than the largest known prior work. Our evaluation covers a wide range of tasks, domains, data regimes and languages, both high and low-resource. On the CoVoST-2 speech translation benchmark, we improve the previous state of the art by an average of 7.4 BLEU over 21 translation directions into English. For speech recognition, XLS-R improves over the best known prior work on BABEL and CommonVoice. XLS-R also sets a new state of the art on VoxLingua107 language identification. Moreover, we show that with sufficient model size, cross-lingual pretraining can perform as well as English-only pretraining when translating English speech into other languages, a setting which favors monolingual pretraining. We hope XLS-R can help to improve speech processing tasks for many more languages of the world. Models and code are available at `www.github.com/pytorch/fairseq/tree/master/examples/wav2vec/xlsr`.[1]

Index Terms: speech pretraining, speech translation, speech recognition, language identification, crosslingual representations

1. Introduction

Self-supervised learning of generic neural representations has gathered much recent interest with a large body of work in natural language processing (NLP) [1, 2], computer vision [3, 4] as well as speech processing [5, 6, 7]. Self-supervised learning provides general representations that can be used across domains and languages.

Multilingually pretrained NLP models such as mBERT [2], XLM-R [8] or mT5 [9] brought significant improvements in multilingual language understanding [10, 11]. These models offer a promising path towards more ubiquitous NLP technology by improving performance for low-resource languages through leveraging data from high-resource languages. Furthermore, it is only necessary to maintain a single multilingual model instead of a myriad of monolingual models.

For speech processing, self-supervised approaches such as wav2vec 2.0 [7, 12] have also been extended to the multilingual setting [13, 14]. The recent XLSR [14] leverages cross-lingual transfer from high-resource languages to build better representations for languages with little unlabeled data. The largest model, XLSR-53, was trained on about 50K hours of public training data in 53 languages and comprises about 300M parameters [14]. But such models only scratch the surface of self-supervised cross-lingual speech representation learning.

In natural language processing, language models are trained on very large datasets, spanning billions of documents such as CC100 [15] or mC4 [9] to fit models with tens of billions and even trillions of parameters [16] with strong results on established benchmarks. In contrast, scaling efforts in speech have focused either on supervised multilingual models [17] or monolingual self-supervised models, counting a billion or more parameters [18], while cross-lingually pretrained speech models are much smaller in scale.

To this end, we present XLS-R, a large-scale cross-lingually pretrained wav2vec 2.0 model whose name is inspired by XLM-R in NLP. It leverages new publicly available VoxPopuli data, comprising 372K hours of unannotated speech [19], the MLS corpus [20], CommonVoice [21], BABEL [22] and VoxLingua107 [23] to cover 128 different languages from various regions of the world. To our knowledge, this is the largest effort to date, in making speech technology accessible for many more languages using publicly available data.

2. Background

Our work builds on [14] who pretrain wav2vec 2.0 models on data from multiple languages. wav2vec 2.0 contains a convolutional feature encoder $f : \mathcal{X} \mapsto \mathcal{Z}$ to map raw audio $\mathcal{X}$ to latent speech representations $z_1, \ldots, z_T$ which are input to a Transformer $g : \mathcal{Z} \mapsto \mathcal{C}$ to output context representations $c_1, \ldots, c_T$ [24]. Training solves a contrastive task over masked feature encoder outputs. At training time, spans of ten time steps with random starting indices are masked. The objective requires identifying the true quantized latent q_t for a masked time-step within a set of $K = 100$ distractors Q_t sampled from other masked time steps. Specifically, training batches contain samples from multiple languages L [2, 25, 14] by sampling from a distribution $p_l \sim \left(\frac{n_l}{N}\right)^{\alpha}$ where $l = 1, \ldots, L$, while n_l is the amount of unlabeled data for each language, and α is the upsampling factor which controls the trade-off between high- and low-resource languages during pretraining.

3. Training Data

We pretrain our models on a total of 436K hours of publicly available data from the following sources:

VoxPopuli (VP-400K) comprises a total of 372K hours of data in 23 European languages of parliamentary speech from the European parliament [19]. This makes it the largest publicly available speech corpus for semi-supervised learning.

Multilingual Librispeech (MLS) contains data in eight European languages totaling around 50K hours of data [20]. The

[*] Equal contribution.

[†] Work done while at Facebook AI

[‡] Equal advising.

[1] Hugging Face: `https://huggingface.co/models?other=xls_r`

majority of the data is English (44K hours).

CommonVoice (CV) is a corpus of read speech. We use the December 2020 release (v6.1; [21]) which covers 60 languages and over 7K hours of speech audio, ranging from over 1.6K hours for English to less than one hour for languages such as Hindi.

VoxLingua107 (VL) is a dataset of 6.6K hours of data in 107 languages based on YouTube content [23] with an average of 62 hours of data per language.

BABEL (BBL) is a multilingual corpus of conversational telephone speech of about 1K hours of data in 17 African and Asian languages [22].

To the best of our knowledge, this is the largest dataset used for training a publicly available self-supervised speech model to date. There are about 24 high-resource languages with more than 1K hours of data each, almost all of which are European, except for Kinyarwanda which is African. Then there is a small number of 17 mid-resource languages which have more than 100 hours of data (but less than 1K hours) which includes Catalan, Persian, Turkish, Russian, and Basque. Finally, the remaining 88 languages are low-resource and have less than 100 hours of data each.

4. Experiments

4.1. Pretraining

We use the wav2vec 2.0 implementation available in fairseq [26] and evaluate models with between 0.3B parameters to 2B parameters. Models are optimized with Adam [27] and the learning rate is warmed up for the first 32K steps followed by polynomial decay to zero for the remainder of training. Training audio sequences are cropped to a maximum of 320K samples (20 sec) and all models were pretrained for a total of one million updates. XLS-R (0.3B) was trained on 128 GPUs with nearly 2M samples on each GPU, totaling 4.3h of data in a batch. Larger models were trained on 200 GPUs with 800K to 1M samples on each GPU giving an effective batch size of about 2.8-3.6 hours.

To balance data from the different languages and corpora we first upsample the languages within a particular corpus using the strategy outlined in §2 and then balance the different corpora using the same strategy by treating each corpus as a different language. We use $\alpha = 0.5$ in all cases.

4.2. Speech Translation

We conduct experiments on CoVoST-2 [28], a multilingual speech translation benchmark based on CommonVoice [21]. It provides data for translating from English into 15 languages (En $\rightarrow$ X)[2] and from 21 languages into English (X $\rightarrow$ En). [3]

The task entails translating speech audio in one language into another language with text as output. Performance is evaluated in terms of BLEU. Models are simultaneously fine-tuned either on the labeled data of all 21 translation directions with English as target language (X $\rightarrow$ En) or on all the 15 directions where English is the input language, resulting in only two models instead of 36. We stack a decoder network on top of XLS-R which is a Transformer network with 12 layers, embedding size 1024, 16 attention heads and feed forward network dimension 4096. The decoder network is initialized with weights from multilingually fine-tuned mBART [29, 30] and uses the same vocabulary with 250K types. The total size of the decoder network is 459M parameters.

[2]The En $\rightarrow$ X languages are: Arabic (ar), Catalan (ca), Welsh (cy), German (de), Estonian (et), Persian (fa), Indonesian (id), Japanese (ja), Latvian (lv), Mongolian (mn), Slovenian (sl), Swedish (sv), Tamil (ta), Turkish (tr), Chinese (zh) where each direction comprises about 430 hours of training data.

[3]The X $\rightarrow$ En languages include all target languages of En $\rightarrow$ X as well as Spanish (es), French (fr), Italian (it), Dutch (nl), Portuguese (pt).

Table 1: *Speech translation: results for X $\rightarrow$ English directions on CoVoST-2 in terms of average BLEU for 21 directions grouped into high/mid/low-resource labeled data directions.*

	high	mid	low	Avg.
Prior work				
XLSR-53 [14]	30.3	11.1	3.2	10.3
VP-100K [32]	27.7	13.2	4.6	11.1
XMEF-En [30]	32.4	16.8	4.0	12.4
XMEF-X [30]	34.2	20.2	5.9	14.7
This work				
XLS-R (0.3B)	30.6	18.9	5.1	13.2
XLS-R (1B)	34.3	25.5	11.7	19.3
XLS-R (2B)	**36.1**	**27.7**	**15.1**	**22.1**

4.2.1. X $\rightarrow$ English

For X $\rightarrow$ English directions we group languages into high-resource (136-264h of train data; fr, de, es, ca), mid-resource (10-49h of train data; fa, it, ru, pt, zh), and low-resource (2-7h of train data; tr, ar, et , mn, nl, sv, lv, sl, ta, ja, id, cy) for ease of presentation. In order to directly compare to XLSR-53 [14], and VP-100K [19], we fine-tune these publicly available models following the same protocol as XLS-R. We also compare to [30] who either use an English-pretrained wav2vec 2.0 model (XMEF-En) for En $\rightarrow$ X directions or the multilingually pretrained XLSR-53 (XMEF-X) for X $\rightarrow$ En directions.

Table 1 shows a new state of the art with XLS-R (2B), improving over the previous best result [30] by 7.4 BLEU on average over all 21 directions (14.7 BLEU vs. 22.1 BLEU). This is largely due to improvements on mid-resource (+7.5 BLEU) and low-resource (+9.2 BLEU) language directions. **Model capacity has a large impact**: XLS-R (1B) improves over XLS-R (0.3B) by an average of 6.1 BLEU and XLS-R (2B) improves by an average of 2.8 BLEU compared to XLS-R (1B).

There is a trend of **larger capacity in pretrained models enabling few-shot learning for speech translation**, similar to wav2vec 2.0 enabling few-shot speech recognition [7, 31]. For example, on language pairs with only two hours of labeled speech translation data, XLS-R (2B) improves over XLS-R (0.3B) as follows: from 10.3 BLEU to 29.6 BLEU on Swedish-English, from 1.4 BLEU to 16.5 BLEU on Indonesian-English and from 3.0 BLEU to 17.1 BLEU on Arabic-English.

4.2.2. English $\rightarrow$ X

For English $\rightarrow$ X directions we compare to previous cross-lingually pretrained models (XLSR-53, VP-100K) as well as baselines with English-only pretraining: XMEF JT, the best performing setup of [30] for En $\rightarrow$ X directions as well as wav2vec 2.0 pre-trained on 60K hours of English Libri-light data and fine-tuned following the same protocol as XLS-R [33, 7]. The latter has the advantage of being pre-trained on exactly the same language as the input data for all translation directions while cross-lingually pretrained models need to be able to represent many different languages which puts them at a disadvantage.

Table 2: *Speech translation: results for English → X directions on CoVoST-2 in terms of BLEU. We show detailed results for four language pairs: English-German (en-de), English-Catalan (en-ca), English-Arabic (en-ar) and English-Turkish (en-tr) as well as the average performance over all 15 directions. For faster experimental turnaround we do not use self-training and LM-decoding as [34] and we expect these methods to be equally applicable to XLS-R.*

	en-ca	en-ar	en-de	en-tr	Avg. (15 dir)
Prior work					
XLSR-53 [14]	29.0	16.5	23.6	15.3	23.4
VP-100K [32]	26.1	14.5	20.8	13.5	20.9
XMEF JT [30]	30.9	18.0	25.8	17.0	25.1
wav2vec 2.0 [34]	32.4	17.4	23.8	15.4	-
+ self-train + LM [34]	**35.6**	**20.8**	27.2	**18.9**	-
This work - monolingual pretraining					
wav2vec 2.0 (720M)	32.7	19.4	27.0	17.7	26.6
This work - cross-lingual pretraining					
XLS-R (0.3B)	28.7	16.3	23.6	15.0	23.2
XLS-R (1B)	32.1	19.2	26.2	17.1	26.0
XLS-R (2B)	34.2	20.7	**28.3**	18.6	**27.8**

Table 2 shows that XLSR-53 now performs similarly to XLS-R (0.3B) while for X → English XLS-R (0.3B) performed much better (see §4.2.1). This is likely because English data dominates the training corpus of XLSR-53 which is not the case for XLS-R (§3). Both XLS-R (1B) and XLS-R (2B) outperform XMEF JT showing that larger capacity results in better performance.

We also compare to prior work using the English-only pretrained wav2vec 2.0 LV-60K model [34] which additionally uses self-training and a language model for decoding. We do not use these techniques. Their results represent the state of the art on these four directions. [34] achieves an average BLEU of 25.6 on the four directions while as XLS-R (2B) rivals this at an average BLEU of 25.5. We note that self-training and LM decoding methods are equally applicable to our approach.

XLS-R (2B) also performs well compared to English-only pretraining at 27.8 average BLEU compared to 26.6 BLEU for a wav2vec 2.0 model pretrained on 60K hours of Libri-light data and 720M parameters. This confirms that **with sufficient capacity, cross-lingual pretraining can perform as well as strong monolingual models** [14].

4.3. Speech Recognition

4.3.1. BABEL

BABEL is a challenging speech recognition benchmark from IARPA consisting of noisy telephone conversational data.[4] We evaluate on five languages: Assamese (as), Tagalog (tl), Swahili (sw), Lao (lo), and Georgian (ka). Training sets comprise between 30 and 76 hours of annotated data. Following [14], we use 10% of the training set for validation, and report test results on the BABEL dev set. We report word error rate (WER) and use n-gram language models trained on CommonCrawl data.

For all speech recognition experiments, we add a linear layer on top of the pretrained model to predict the output vocabulary and train using Connectionist Temporal Classification (CTC; [35]). We fine-tune using Adam for 20K updates and the learning rate is warmed up for the first 10% of total updates, kept constant for the next 40% and then decayed to zero in the remaining 50% of updates. We also use a language model for decoding.

Table 3: *Speech recognition results on BABEL in terms of word error rate (WER) on Assamese (as), Tagalog (tl), Swahili (sw), Lao (lo) and Georgian (ka).*

	as	tl	sw	lo	ka
Labeled data	55h	76h	30h	59h	46h
Previous work					
Alumae et al. (2016; [37])	-	-	-	-	32.2
Ragni et al. (2018; [38])	-	40.6	35.5	-	-
Inaguma et al. (2019; [39])	49.1	46.3	38.3	45.7	-
XLSR-10 [14]	44.9	37.3	35.5	32.2	-
XLSR-53 [14]	44.1	33.2	26.5	-	31.1
This work					
XLS-R (0.3B)	42.9	33.2	24.3	31.7	28.0
XLS-R (1B)	40.4	30.6	21.2	30.1	25.1
XLS-R (2B)	**39.0**	**29.3**	**21.0**	**29.7**	**24.3**

Table 3 shows that XLS-R (0.3B) outperforms the equally sized XLSR-53, which was the previous state of the art on all languages by an average of 1.4 WER, e.g., on Assamese (as), WER decreases from 44.1 to 42.9, on Swahili (sw) WER decreases from 26.5 to 24.3 and on Georgian (ka) WER drops from 31.1 to 28.0 WER. XLSR-53 and XLS-R were both pretrained on the same BABEL data, and the better performance of XLS-R (0.3B) shows that pretraining on additional out-of-domain datasets such as VoxPopuli does help performance on BABEL. This is similar to findings for monolingual pretraining [36].

Using additional capacity, XLS-R (1B) outperforms XLS-R (0.3B) by 2.5 WER on average. On Georgian (ka), this corresponds to improvements of 6 WER and 7.1 WER compared to [14] and [37], respectively. XLS-R (2B) improves over XLS-R (1B) by 0.8 WER on average showing that additional capacity can further improve performance.

4.3.2. CommonVoice

Following [40], we use ten languages of CommonVoice: Spanish (es), French (fr), Italian (it), Kyrgyz (ky), Dutch (nl), Russian (ru), Swedish (sv), Turkish (tr), Tatar (tt) and Chinese-Hong Kong (zh-HK).[5] CommonVoice contains read speech primarily from Wikipedia sentences. Following prior work [40, 14], we fine-tune models on just one hour of labeled data per language, a few-shot scenario. Results are reported in terms of phoneme error rate (PER) without a language model.

On English speech recognition, pretraining has been shown to be particularly beneficial for low labeled data settings [7]. This is similar to cross-lingual pretraining [14] where pretraining on the large MLS corpus significantly improved performance over pretraining only on CommonVoice data, e.g., on Dutch accuracy improved from 14 PER to 5.8 PER.

Table 4 shows that the additional training data of XLS-R compared to XLSR-53 results in better performance of 1.1 PER on average for XLS-R (0.3B). XLS-R uses the same training data as XLSR-53 plus the very large VP-400K corpus of parliamentary speech as well as the much smaller VoxLingua-107 which consists of YouTube data, both of which are out of domain with

[4] `https://catalog.ldc.upenn.edu/byyear` {LDC2016S06, LDC2016S13, LDC2017S05, LDC2017S08, LDC2016S12}

[5] `https://dl.fbaipublicfiles.com/cpc_audio/common_voices_splits.tar.gz`

Table 4: *Phoneme recognition performance on CommonVoice in terms of phoneme error rate (PER) when using one hour of labeled data to fine-tune each language. We compare to m-CPC [40], [41], XLSR-10 [14] and XLSR-53 [14].*

	es	fr	it	ky	nl	ru	sv	tr	tt	zh	Avg
Labeled data	1h	1h	1h	1h	1h	1h	1h	1h	1h	1h	
Previous work											
m-CPC	38.0	47.1	40.5	41.2	42.5	43.7	47.5	47.3	42.0	55.0	44.5
[41]	36.6	48.3	39.0	38.7	47.9	45.2	52.6	43.4	42.5	54.3	44.9
XLSR-10	7.9	12.6	11.7	7.0	14.0	9.3	20.6	9.7	7.2	22.8	12.3
XLSR-53	2.9	5.0	5.7	6.1	5.8	8.1	12.2	7.1	5.1	18.3	7.6
This work											
XLS-R (0.3B)	3.1	5.4	4.9	5.1	5.8	6.0	7.2	6.0	4.1	17.0	6.5
XLS-R (1B)	**2.0**	**3.9**	**3.5**	4.1	**4.2**	4.1	5.5	4.4	3.4	15.7	5.1
XLS-R (2B)	2.2	4.0	**3.5**	**4.0**	4.7	**3.7**	**5.0**	**4.0**	**2.9**	**14.8**	**4.9**

respect to the read audiobook domain of CommonVoice. This confirms that pretraining on more out of domain data can still improve performance [36].

Furthermore, accuracy improves even on languages for which XLS-R does not add any pretraining data compared to XLSR-53, e.g., Kyrgyz (ky) improves from 6.1 PER to 5.1 PER for XLS-R (0.3B) and 4.1 PER for XLS-R (1B) and both models are pretrained on only about 11 hours of Kyrgyz data - 0.003% of the total pretraining data. This shows that there is cross-lingual transfer that benefits low-resource languages and that additional capacity is important to realize this effect.

Chinese improves the least and gains are particularly large for languages for which the training corpus of XLS-R contains more data due to VoxPopuli, e.g., for Swedish VP-400K adds more than 16K hours of unannotated speech and performance improves from 12.2 PER to 5.5 PER when comparing XLSR-53 to XLS-R (1B). Finally, XLS-R (2B) performs slightly better than XLS-R (1B) on average with some languages improving while as others are performing slightly worse. The modest average improvement is likely because error rates are already low on this benchmark.

4.4. Language Identification

Finally, we evaluate our approach on language identification for which we use our smallest model as the tasks requires less capacity given the lower complexity compared to the structured prediction problems of speech recognition and speech translation. We consider VoxLingua107 [23] which spans 107 languages. It consists of short speech segments automatically extracted from YouTube videos. We train our model on the official train set of 6,628 hours of data, and report results on the development set, comprising 33 languages.

Table 5 shows that our best model outperforms previous work, improving the best known prior work [42] by 1% absolute, a relative error reduction of 15%. For comparison, we also fine-tune the English-only wav2vec 2.0 pretrained on Libri-Light which performs surprisingly well on this multilingual task but XLS-R outperforms it by 1.5% error rate on average.

4.5. Discussion

Cross-lingual training results in a single model for multiple languages compared to a separate model for each language. Training a cross-lingual model requires more effort than a single monolingual model but the resulting model can be used for many different languages. Advances in architectures and training can also be deployed more easily since we only need to retrain a single model rather than many different ones.

Table 5: *Language identification on VoxLingua107. We report the error rate on the development set spanning 33 languages.*

	Error Rate (%)		
	0...5 sec	5...20 sec	Avg
Previous work			
Valk et al. (2020; [23])	12.3	6.1	7.1
Speechbrain (2021; [43])	-	-	6.7
This work			
wav2vec 2.0 LV-60K (300M)	11.5	6.3	7.2
XLS-R (0.3B)	**9.1**	**5.0**	**5.7**

In terms of accuracy, prior work in self-supervised learning for speech established that cross-lingually pretrained models are very competitive to monolingually pretrained models for speech recognition [14]. Our experiments show a similar trend for speech-translation: XLS-R can perform very competitively to English-only pretrained models for English $\rightarrow$ X speech translation where the encoder only needs to encode English speech - a setting which favors monolingually pretrained models.

Overall, XLS-R performs best for low-resource and mid-resource languages. For speech translation, we observe strong improvements for low- and mid-resource X $\rightarrow$ English directions and comparatively smaller gains on high-resource directions. Many low-resource directions which previously had performance in the 1-5 BLEU range improve to over 10-20 BLEU due to the better cross-lingual speech representations. For English $\rightarrow$ X directions, large enough cross-lingual models can even surpass the performance of English-only pretrained models.

Similarly, for speech recognition, we see strong improvements on BABEL and CommonVoice. We find that models trained on more data from more languages can perform as well or better than comparable models of the same size and we we observe this trend across all speech recognition benchmarks. Keeping everything else equal, larger capacity models often further improve performance.

5. Conclusion

XLS-R is a new self-supervised cross-lingual speech representation model which scales the number of languages, the amount of training data as well as model size. The training corpus is an order of magnitude larger than prior work and covers 128 languages in 436K hours of recorded speech audio. The resulting model enables state of the art results for X $\rightarrow$ English speech translation on CoVoST-2, outperforming prior art by a sizeable margin with the largest improvements on mid- and low-resource directions. It also performs competitively to the best English $\rightarrow$ X work, without the use of equally applicable techniques such as self-training and language model decoding.

On speech recognition, XLS-R sets a new state of the art on CommonVoice and several languages of BABEL. Our model also sets a new state of the art on the VoxLingua107 language identification benchmark. The largest XLS-R model comprises 2B parameters which enables it to outperform a strong English-only pretrained model on English $\rightarrow$ X speech translation, a setting which favors monolingually pretrained models. This shows that cross-lingually trained models with sufficient capacity can perform as well as specialized monolingually pretrained models. We hope XLS-R will help catalyze research in speech technology for many more languages of the world. Models and code are publicly available on several platforms.

6. References

[1] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving language understanding by generative pre-training," *arXiv*, 2018.

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

[3] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proc. of ICML*, 2020.

[4] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proc. of CVPR*, 2020.

[5] A. van den Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *Proc. of NIPS*, 2018.

[6] S. Schneider, A. Baevski, R. Collobert, and M. Auli, "wav2vec: Unsupervised pre-training for speech recognition," in *Proc. of Interspeech*, 2019.

[7] 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.

[8] 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," in *Proc. of ACL*, 2020.

[9] L. Xue, N. Constant, A. Roberts, M. Kale, R. Al-Rfou, A. Siddhant, A. Barua, and C. Raffel, "mt5: A massively multilingual pre-trained text-to-text transformer," *arXiv*, 2020.

[10] A. Conneau, G. Lample, R. Rinott, A. Williams, S. R. Bowman, H. Schwenk, and V. Stoyanov, "Xnli: Evaluating cross-lingual sentence representations," *arXiv preprint arXiv:1809.05053*, 2018.

[11] 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," *arXiv preprint arXiv:2104.07412*, 2021.

[12] 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 *Proc. of ICASSP*. IEEE, 2021.

[13] K. Kawakami, L. Wang, C. Dyer, P. Blunsom, and A. van den Oord, "Learning robust and multilingual speech representations," *Proc. of EMNLP*, 2020.

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

[15] G. Wenzek, M.-A. Lachaux, A. Conneau, V. Chaudhary, F. Guzmán, A. Joulin, and É. Grave, "Ccnet: Extracting high quality monolingual datasets from web crawl data," in *Proc. of LREC*, 2020.

[16] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, and et al., "Language models are few-shot learners," in *Proc. of NeurIPS*, 2020.

[17] B. Li, R. Pang, T. N. Sainath, A. Gulati, Y. Zhang, J. Qin, P. Haghani, W. R. Huang, M. Ma, and J. Bai, "Scaling end-to-end models for large-scale multilingual asr," *arXiv*, 2021.

[18] Y. Zhang, J. Qin, D. S. Park, W. Han, C.-C. Chiu, and R. P. et al., "Pushing the limits of semi-supervised learning for automatic speech recognition," in *Proc. of NeurIPS SAS Workshop*, 2020.

[19] 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.

[20] 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.

[21] R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, and et al., "Common voice: A massively-multilingual speech corpus," *Proc. of LREC*, 2020.

[22] 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 *Proc. of SLT*, 2014.

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

[24] A. Baevski, S. Schneider, and M. Auli, "vq-wav2vec: Self-supervised learning of discrete speech representations," in *Proc. of ICLR*, 2020.

[25] G. Lample and A. Conneau, "Cross-lingual language model pre-training," in *Proc. of NeurIPS*, 2019.

[26] M. Ott, S. Edunov, A. Baevski, A. Fan, S. Gross, N. Ng, D. Grangier, and M. Auli, "fairseq: A fast, extensible toolkit for sequence modeling," in *Proc. of NAACL System Demonstrations*, 2019.

[27] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," in *Proc. of ICLR*, 2015.

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

[29] Y. Liu, J. Gu, N. Goyal, X. Li, S. Edunov, M. Ghazvininejad, M. Lewis, and L. Zettlemoyer, "Multilingual denoising pre-training for neural machine translation," *arXiv*, vol. abs/2001.08210, 2020.

[30] X. Li, C. Wang, Y. Tang, C. Tran, Y. Tang, J. Pino, A. Baevski, A. Conneau, and M. Auli, "Multilingual speech translation with efficient finetuning of pretrained models," in *Proc. of ACL*, 2021.

[31] 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 *Proc. of ICASSP*, 2020.

[32] 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.

[33] J. Kahn *et al.*, "Libri-light: A benchmark for asr with limited or no supervision," in *Proc. of ICASSP*, 2020.

[34] C. Wang, A. Wu, J. Pino, A. Baevski, M. Auli, and A. Conneau, "Large-scale self- and semi-supervised learning for speech translation," in *Proc. of Interspeech*, 2021.

[35] A. Graves, S. Fernández, and F. Gomez, "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks," in *Proc. of ICML*, 2006.

[36] W.-N. Hsu, A. Sriram, A. Baevski, T. Likhomanenko, Q. Xu, V. Pratap, J. Kahn, A. Lee, R. Collobert, G. Synnaeve *et al.*, "Robust wav2vec 2.0: Analyzing domain shift in self-supervised pre-training," *arXiv preprint arXiv:2104.01027*, 2021.

[37] 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.

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

[39] 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.

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

[41] R. Fer, P. Matějka, F. Grézl, O. Plchot, K. Veselỳ, and J. H. Černockỳ, "Multilingually trained bottleneck features in spoken language recognition," *Computer Speech & Language*, 2017.

[42] M. Ravanelli, J. Zhong, S. Pascual, P. Swietojanski, J. Monteiro, J. Trmal, and Y. Bengio, "Multi-task self-supervised learning for robust speech recognition," *arXiv*, 2020.

[43] M. Ravanelli, T. Parcollet, P. Plantinga, A. Rouhe, S. Cornell, L. Lugosch, C. Subakan, N. Dawalatabad, and et al., "SpeechBrain: A general-purpose speech toolkit," *arXiv*, 2021.