INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Fast and Efficient Multilingual Self-Supervised Pre-training for Low-Resource Speech Recognition

Zhilong Zhang, Wei Wang, Yanmin Qian[†]

MoE Key Lab of Artificial Intelligence, AI Institute
X-LANCE Lab, Department of Computer Science and Engineering
Shanghai Jiao Tong University, Shanghai, China
{zzdragon,wangwei.sjtu,yanminqian}@sjtu.edu.cn

Abstract

Recent advances in self-supervised learning (SSL) have remarkably improved speech recognition performance for low-resource languages. On the other hand, with data of an increasingly larger scale required for SSL, the pre-training process has become extremely time-consuming. To address this problem, we propose an unsupervised data selection method based on utterance-level language similarity and a curriculum learning strategy to boost the efficiency of multilingual SSL pretraining while maintaining performance. We conduct experiments on five languages in COMMONVOICE dataset. Compared to the baseline with all data for pretraining, we pretrained on only 25% of the data and saved 60% of the training steps with even better performance on the target low-resource language.

Index Terms: low resource speech recognition, data selection, self-supervised pretraining, curriculum learning

1. Introduction

Benefiting from abundant paired data for training, end-to-end (E2E) automatic speech recognition (ASR) models have achieved promising results on rich-resource languages [1–4]. However, significant performance degradation has been observed when E2E models are applied to low-resource languages [5] where curated corpora are hardly available. To address this issue, it has been found effective to adopt the pretraining and fine-tuning paradigm that exploits a significant amount of multi-lingual data and restrained monolingual data from the target language to improve ASR performance.

Multilingual transfer learning and multilingual meta-learning are two approaches that leverage labeled data to pre-train a seed model with multi-lingual data, which is then used for initialization during fine-tuning [6–10]. This strategy narrows the parameter search space, making it easier to converge on data from low-resource target languages [11, 12]. To further improve low-resource automatic speech recognition (ASR), [9] introduced an auxiliary speech-to-text translation task that translates labeled speech from a rich-resource language to text in a low-resource language. Additionally, [10] proposed to optimize the set of parameters in the seed model with meta-learning for fast adaptation to different languages. However, it's important to note that both multilingual transfer learning and meta-learning require paired data throughout the pre-training and fine-tuning paradigm.

While multilingual transfer learning and meta-learning require paired data, self-supervised learning (SSL) has become a popular technique for various tasks because of its ability to extract semantic clues from easily accessible unpaired speech data [13–16]. A pre-trained SSL model can be fine-tuned efficiently with minimal supervised data and training steps to produce an adequate ASR model [17]. For instance, XLSR-53 [18] and XLS [19] were pre-trained on 56k and 500k hours of speech data, respectively, across many languages. Both models achieved impressive results across various languages, demonstrating the capability of SSL models. However, existing SSL works usually train large and general models without considering the similarity between the pre-training data and a specific target language. As a result, it takes significant computational results to train SSL models on enormous amounts of unselected data. Additionally, the mismatch between the pre-training data and the target language for evaluation might lead to performance degradation. Recent works have attempted to address this issue. For example, [20] and [21] intentionally pre-train on data from the same language family as the target language, while [22] and [23] employ a language identification network for language-level data selection. Although the amount of training data is reduced, these methods fail to exploit potentially beneficial data from other languages that are excluded from pre-training.

In this paper, we propose an approach to enhance the data efficiency of SSL pre-training for specific target languages by utilizing data selection based on utterance-level language similarity. Specifically, we rank and select the pre-training data by measuring similarity through the embedding distance to the low-resource target language. Additionally, we introduce an extension to dynamic curriculum learning [24], where the difficulty of training samples is assessed by the weighted sum of the language similarity and running loss. We adopt wav2vec2.0 as the SSL architecture and evaluate the proposed approach through experiments conducted on five languages in the COMMONVOICE dataset.

Our contributions can be summarized as follows: (1) For ASR on a target language, we significantly reduce the data and time required for SSL pre-training. (2) We provide insights into the SSL pre-training data selected for each target language. (3) Compared to the baseline with all data for SSL pre-training, we achieve even better performance with only 25% of data and 60% training time.

2. Methodology

2.1. Utterance-level Language Similarity Evaluation

It has been validated in [22,23] that a LID network trained with a small amount of unpaired speech can effectively distinguish data from different languages. The bottleneck features extracted from the LID network can be used for language similarity evaluation. In this work, we adopt the Time Delay Neural Network (TDNN) [25] as the LID network and evaluate the simi-

[†] corresponding author

10.21437/Interspeech.2023-1630

larity of an utterance to a specific language with the extracted bottleneck features.

Denote the target language as l, the embedding of l as $\mathbf{E_l}$, the set of all utterances in l as L, and the embedding of the utterance u in the dataset as $\mathbf{E_u}$. The similarity between $\mathbf{E_u}$ and $\mathbf{E_l}$ is evaluated as their cosine distance:

$$d(\mathbf{E_u}, \mathbf{E_l}) = \frac{\mathbf{E_u} \cdot \mathbf{E_l}}{\|\mathbf{E_u}\| \cdot \|\mathbf{E_l}\|} \tag{1}$$

where the language embedding $\mathbf{E_l}$ is defined as the averaged embedding of all utterances $u \in L$:

$$\mathbf{E_L} = \sum_{\mathbf{u} \in \mathbf{L}} \mathbf{E_u} \Big/ \sum_{\mathbf{u} \in L} 1 \tag{2}$$

2.2. Data Selection

In prior studies, the main focus has been on language-level pretraining data selection, which involves excluding all data from unselected languages from the pre-training process. However, we argue that this practice may not be the most optimal approach as it could lead to the dismissal of potentially useful data. To test this hypothesis, we extracted embeddings from 400 utterances in each of five languages and conducted T-SNE analysis to visualize the linguistic distance between utterances from different languages. The results of this analysis are presented in Fig. 1. Although embeddings from Italian and French were generally more similar, we found that a certain number of embeddings from Portuguese fell within the range of Italian. Furthermore, a small proportion of embeddings from every language was dispersed throughout the embedding space, indicating that utterance-level selection, irrespective of the actual language, yields more reasonable data than language-level selection in terms of similarity to the target language. We further provide quantitative analysis in experiments regarding the actual language distribution of the selected data that are similar to a specific target language.

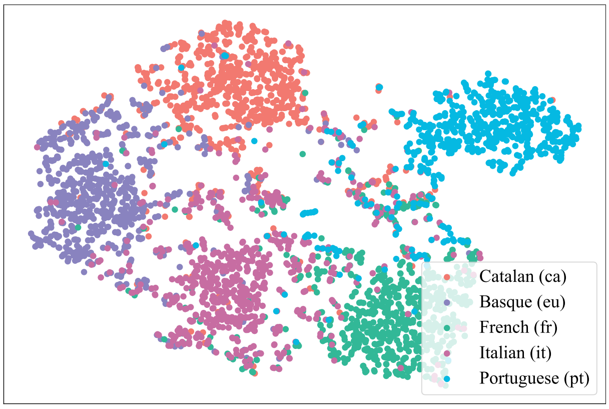


Figure 1: *T-SNE analysis for embeddings of 400 utterances in each of the five languages*

2.3. Extended Dynamic Curriculum Learning (EDCL)

Curriculum Learning (CL) involves defining a criterion for determining the difficulty of training samples. In Dynamic CL (DCL), the difficulty criterion is evaluated by monitoring the running loss dynamically throughout the training process. However, DCL randomly selects a training subset at the beginning of the training, which can result in a subset of unknown difficulty during the critical early stages of training.

To address this issue, we propose an extension to DCL (EDCL), which includes an additional language similarity term in the evaluation of difficulty. This modification helps create a more reasonable training subset at the early stages of training when the running loss is not yet stable. Specifically, we define the difficulty criterion H_u of an utterance u as the weighted sum of the embedding distance between the target language l and the corresponding length-normalized running loss $\mathcal{L}_u$:

$$H_u = \alpha * d(\mathbf{E_u}, \mathbf{E_l}) + \sigma * \mathcal{L}_u \tag{3}$$

During the initial phase of training, we set $\mathcal{L}_u$ to 0 and determine the initial subset for training based on utterance-level language similarity. We periodically re-evaluate the difficulty of the data with H_u and adjust the training set accordingly. This process allows EDCL to gradually include more difficult data as the training progresses. However, we limit the size of the dynamically determined training set to exclude data that has minor similarity to the target language and a large running loss, which can be detrimental to the training process.

3. Experimental Setup

3.1. Data

We conduct experiments by pre-training and fine-tuning on the CommonVoice dataset[1] [26]. CommonVoice is a large multilingual speech corpus, with content taken primarily from Wikipedia articles. Following the setup in [24], we selected five languages for experiment: Catalan (ca), Basque (eu), French (fr), Italian (it), and Portuguese (pt), totaling 1145 hours. The duration of data for each language is summarized in Fig. 2 (a). For data selection, we investigate different percentages of selected data. The selected set of data for the target language l is as D_R^l where $R \in \{12.5\%, 25\%, 50\%, 100\%\}$ is the percentage for the selected data. Note that for a specific target language l, $D_{12.5\%}^l \subset D_{25\%}^l \subset D_{50\%}^l \subset D_{100\%}^l$ always holds.

3.2. LID Network

LID networks typically consist of a TDNN followed by a simple classifier. We utilized a three-layer fully connected classifier with ReLU activation functions. We employed the adam optimizer [27] with a learning rate of 0.0005. To train the language classifiers, we selected only three hours of data from each language as training data and set the number of epochs to 100. Training the language classifiers can be effectively accomplished using only one GPU within two hours.

3.3. Pre-training and Fine-tuning

We utilized FairSeq [28] as our toolkit for both pre-training and fine-tuning experiments. The pre-training stage utilized a wav2vec2.0 base model with 12 layers of conformers, following the hyperparameters of the BASE model [13], with a maximum of 400k updates and a learning rate of 0.0002.

For fine-tuning, we selected the pre-trained model checkpoint with the highest performance for fine-tuning on the target languages. Connectionist temporal classification [29] was used, and we fine-tuned it for 20k updates with a learning rate of 0.00005. To create a low-resource setup, we used only 10 hours of paired data from the target language for fine-tuning.

[1] https://commonvoice.mozilla.org/en/datasets

Table 1: *WER (%) comparison of different selected pre-training data (%).The pre-trained model finetunes over 10 hours of the target language.* ***# Converge Updates*** *stands for the number of update steps till convergence.*

Pre-train data	**# Converge Updates**	**LM**	Catalan (ca)		Basque (eu)		French (fr)		Italian (it)		Portuguese (pt)	
			dev	test	dev	test	dev	test	dev	test	dev	test
100% (1145h)	400k	✗	17.53	17.90	15.46	15.69	30.86	32.96	24.87	26.01	22.62	24.98
		✓	5.94	6.62	9.87	10.46	27.91	29.51	25.09	25.8	9.61	10.45
50% (573h)	400k	✗	17.50	17.84	13.90	14.29	30.69	32.87	24.82	25.96	19.03	20.78
		✓	5.42	6.11	9.77	10.17	26.57	28.07	24.28	24.91	9.46	9.97
25% (286h)	215k	✗	15.20	15.51	13.08	13.83	29.62	31.19	24.73	25.32	17.72	19.87
		✓	5.58	6.11	9.62	10.39	26.99	28.86	24.16	24.83	9.52	10.26
12.5% (143h)	103k	✗	19.25	19.97	17.62	18.10	33.88	34.73	28.25	29.32	24.19	25.92
		✓	8.63	9.97	11.24	11.93	31.12	33.83	27.83	28.51	11.86	12.81
EDCL$_{25\%}$	160k	✗	15.22	15.43	13.75	13.85	29.57	31.05	23.85	25.27	18.09	19.66
		✓	5.35	6.01	9.81	10.19	27.21	28.17	24.63	25.41	9.47	9.95
EDCL$_{25\%\rightarrow50\%}$	200k	✗	14.92	15.08	12.80	12.99	29.07	30.85	23.09	24.47	16.79	18.52
		✓	**5.16**	**5.83**	**9.55**	**9.88**	**25.27**	**26.17**	**22.67**	**23.51**	**9.13**	**9.56**

3.4. Extended Dynamic Curriculum Learning

In our experiment, we explore two different strategies for EDCL, which are denoted as EDCL$_{25\%}$ and EDCL$_{25\%\rightarrow50\%}$. The former strategy involves adjusting the training set during each EDCL evaluation, while maintaining a fixed percentage of data selection at 25%. The latter strategy starts with an initial dataset $D_{25\%}$, and gradually increases the selected data to $D_{50\%}$. The percentage of selected data R in EDCL$_{25\%\rightarrow50\%}$ changes as follows:

$$R = a_0 + \lfloor \frac{t}{\beta} \rfloor * 10\% \; (a_0 = 25\%, \beta = 50000) \qquad (4)$$

where t refers to the number of training updates, a_0 is the initial amount of data, and β is the interval of updates between two EDCL evaluations. For data selection criterion in each evaluation, we set α to 0.5 and σ to 0.9 in Eq. 3.

3.5. Decoding and Langauge Model

For language model rescoring, we use a 4-gram language model built with KenLM [30] and trained on the text part of each language and adopt a beam size of 50 for the Wav2letter++ [31] beam search decoder. Language model weight and word insertion penalty are empirically set to 3.2 and -0.8, respectively.

4. Results and Analysis

4.1. Language Similarity

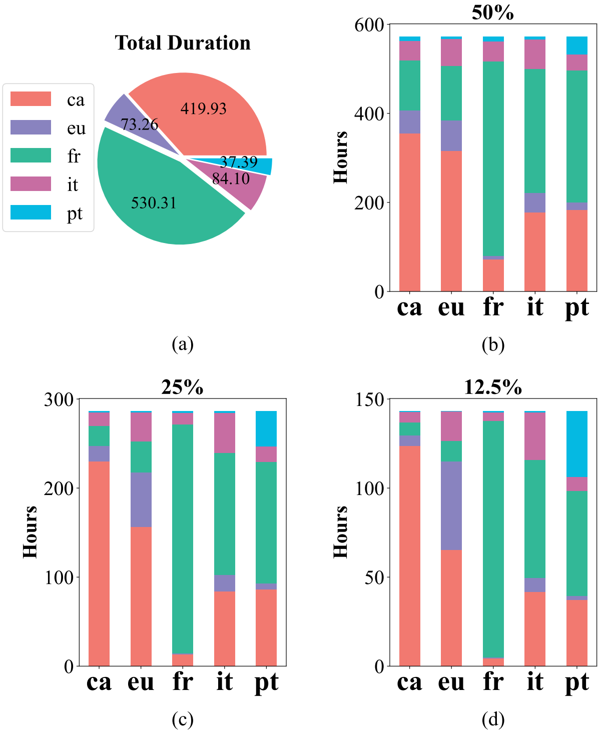
Fig.2 (b)-(d) show the language distribution of selected pre-training data for each target language at different data selection thresholds. While the imbalanced total duration of different languages makes it challenging to analyze language similarity from the histograms, we can still draw some conclusions that are consistent with the T-SNE results in Fig.1. For example, although the total amount of Catalan (ca) data and French (fr) data is similar, the amount of Catalan (ca) data selected for the Basque (eu) language in Fig.2 (d) is significantly larger than the amount of French (fr) data. This suggests that Catalan (ca) is more similar to Basque (eu) than French (fr), which is consistent with the T-SNE embedding distance in Fig.1. Similarly, although the total amount of Italian (it) data and Basque (eu) data is comparative, a significantly larger amount of Italian (it) data is selected for the French (fr) language than the amount of Basque (eu) data, which is consistent with the T-SNE results indicating that Italian (it) is more similar to French (fr) than Basque (eu).

4.2. Data selection

We present results by selecting different amounts of data based on the language similarity criterion in Eq.1, as shown in the first four rows of Table 1. To facilitate comparison, Fig.3 displays the relative WER reduction and convergence steps averaged across results from five languages, at different percentages of data. We observed that reducing the amount of data from 100% to 50% led to reduced WER while the convergence steps were not reduced, which suggests an improvement in data quality but some of the data selected might have impeded convergence. Further reducing the data to 25% resulted in both reduced convergence steps and a further reduction in WER, indicating the effectiveness of the data selection. However, when the data was reduced to 12.5%, we observed significant performance degradation due to the exclusion of a considerable amount of beneficial training data.

Moreover, we found that the quality of Italian data was comparatively poor, with a significant proportion of non-Italian tokens and symbols. Our results also suggest that rescoring the Italian language with a language model did not result in a significant improvement in performance.

Interestingly, we noticed that the degree of improvement achieved by data selection varied across languages and was related to the duration of that language in the original dataset $D_{100\%}$. For instance, reducing the data from $D_{100\%}$ to $D_{25\%}$ for Catalan, which has the most data in $D_{100\%}$, resulted in a relative improvement of 13% / 11% on dev / test sets. However, the improvement increased to 22% / 20% for Portuguese, which has the least data in $D_{100\%}$. We hypothesize that $D_{100\%}$ contains more noisy data for a specific language when it has fewer data in $D_{100\%}$. Therefore, data selection is more critical for such languages as it can exclude more detrimental data. This observation holds for all languages except Italian, which suffers from noisy data in its own dataset.



(a)

(b)

(c)

(d)

Figure 2: *The pie chart shows the total duration and distribution of the pre-training data for five languages. The histograms show the language distribution of selected pre-training data for each target language at different data selection thresholds.*

4.3. Extended Dynamic Curriculum Learning

The last two rows of Table 1 demonstrate the effectiveness of incorporating the EDCL strategy to improve the ASR performance and reduce the number of steps required for convergence. When training with a fixed 25% of the data, using $\text{EDCL}_{25\%}$ resulted in a reduction of 55k convergence steps compared to the third row without a curriculum learning strategy, while maintaining ASR performance. This indicates that the proposed criterion in Equation 3 is a useful estimator of utterance difficulty, which enables the model to converge more efficiently.

In the last row of Table 1, we used $\text{EDCL}_{25\%\rightarrow 50\%}$ to incrementally select the training set from 25% to 50%. This resulted in a further improvement in ASR performance. Compared to the second row, which used 50% of the data without a dedicated curriculum, $\text{EDCL}_{25\%\rightarrow 50\%}$ achieved better convergence steps and WER performance. This finding suggests that the EDCL strategy provides a more effective way to utilize limited training data for ASR pre-training, leading to improved performance with fewer steps.

4.4. Training Efficiency and ASR Performance

In this section, we provide an analysis of the training efficiency and ASR performance of our proposed methods. Figure 3 displays the WER improvement relative to the baseline, which was pre-trained on $D_{100\%}$, and the actual update steps till convergence. Each column group in the figure corresponds to the averaged results in a row of Table 1.

The first three column groups indicate a trade-off between convergence steps and ASR performance. As the amount of data selected reduced from 50% to 12.5%, it took fewer steps for the model to converge, but this came at the cost of degraded ASR performance. Although such a trade-off is likely inevitable, we achieved a better trade-off by incorporating a curriculum learning strategy, as shown in the last two column groups. This strategy effectively reduced the difficulty at the early stages of pre-training, making the model easier to converge in fewer steps. The comparison between the first and last column groups best demonstrates this.

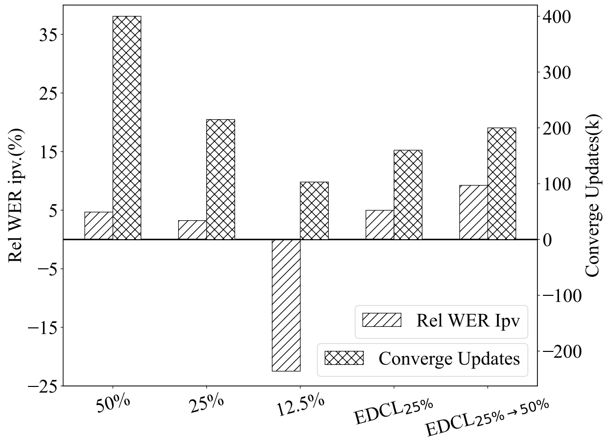


Figure 3: *Relative WER (%) improvements with LM rescoring and Converge Updates (k) averaged over results in Table 1.*

Incorporating the dedicated EDCL strategy led to a reduction in training steps and an improvement in ASR performance. The EDCL strategy thus provides a promising approach for achieving a more efficient and effective pre-training process for ASR of low-resource target languages .

5. Conclusions

In conclusion, this paper proposes an utterance-level language similarity-based data selection method for SSL pre-training, which effectively reduces the amount of required data and training steps while eliminating potentially detrimental data for ASR on low-resource target languages. We provide an insightful analysis of the correlation between our proposed data selection approach and language similarity. Additionally, we introduce a curriculum learning strategy that utilizes a dedicated estimation of language similarity and the running loss as the difficulty criterion for an utterance. We carry out experiments on five languages in the COMMONVOICE dataset with a 10-hour low-resource setup. By incorporating these techniques, we achieve superior performance on the low-resource target language, pre-training on only 25% of the data and saving 60% of the training steps compared to the baseline with all data for pre-training. Our findings highlight the effectiveness and efficiency of our proposed method in SSL pre-training for low-resource ASR, with potential applications in various other SSL tasks.

6. Acknowledgements

This work was supported in part by China NSFC projects under Grants 62122050 and 62071288, and in part by Shanghai Municipal Science and Technology Major Project under Grant 2021SHZDZX0102.

7. References

[1] Albert Zeyer, Kazuki Irie, Ralf Schlüter, and Hermann Ney, "Improved training of end-to-end attention models for speech recognition," in *Proc. Interspeech*, 2018, pp. 7–11.

[2] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever, "Robust speech recognition via large-scale weak supervision," *ArXiv*, vol. abs/2212.04356, 2022.

[3] William Chan, Daniel S. Park, Chris Lee, Yu Zhang, Quoc V. Le, and Mohammad Norouzi, "SpeechStew: Simply mix all available speech recognition data to train one large neural network," in *Workshop on Machine Learning in Speech and Language Processing (MLSLP)*, 2021.

[4] Jinyu Li et al., "Recent advances in end-to-end automatic speech recognition," *APSIPA Transactions on Signal and Information Processing*, vol. 11, no. 1, 2022.

[5] Thomas Reitmaier, Electra Wallington, Dani Kalarikalayil Raju, Ondrej Klejch, Jennifer Pearson, Matt Jones, Peter Bell, and Simon Robinson, "Opportunities and challenges of automatic speech recognition systems for low-resource language speakers," in *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, 2022, pp. 1–17.

[6] Suyoun Kim and Michael L Seltzer, "Towards language-universal end-to-end speech recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 4914–4918.

[7] Shubham Toshniwal, Tara N Sainath, Ron J Weiss, Bo Li, Pedro Moreno, Eugene Weinstein, and Kanishka Rao, "Multilingual speech recognition with a single end-to-end model," in *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2018, pp. 4904–4908.

[8] Jaejin Cho, Murali Karthick Baskar, Ruizhi Li, Matthew Wiesner, Sri Harish Mallidi, Nelson Yalta, Martin Karafiat, Shinji Watanabe, and Takaaki Hori, "Multilingual sequence-to-sequence speech recognition: architecture, transfer learning, and language modeling," in *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2018, pp. 521–527.

[9] Changhan Wang, Juan Pino, and Jiatao Gu, "Improving Cross-Lingual Transfer Learning for End-to-End Speech Recognition with Speech Translation," in *Proc. Interspeech 2020*, 2020, pp. 4731–4735.

[10] Jui-Yang Hsu, Yuan-Jui Chen, and Hung-yi Lee, "Meta learning for end-to-end low-resource speech recognition," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 7844–7848.

[11] Siddharth Dalmia, Ramon Sanabria, Florian Metze, and Alan W Black, "Sequence-based multi-lingual low resource speech recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 4909–4913.

[12] Vikas Joshi, Rui Zhao, Rupesh R Mehta, Kshitiz Kumar, and Jinyu Li, "Transfer learning approaches for streaming end-to-end speech recognition system," *arXiv preprint arXiv:2008.05086*, 2020.

[13] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in neural information processing systems*, vol. 33, pp. 12449–12460, 2020.

[14] Wei-Ning Hsu, Yao-Hung Hubert Tsai, Benjamin Bolte, Ruslan Salakhutdinov, and Abdelrahman Mohamed, "Hubert: How much can a bad teacher benefit asr pre-training?," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6533–6537.

[15] Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, et al., "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, 2022.

[16] Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli, "Data2vec: A general framework for self-supervised learning in speech, vision and language," in *International Conference on Machine Learning*. PMLR, 2022, pp. 1298–1312.

[17] Pablo Peso Parada, Agnieszka Dobrowolska, Karthikeyan Saravanan, and Mete Ozay, "pMCT: Patched Multi-Condition Training for Robust Speech Recognition," in *Proc. Interspeech 2022*, 2022, pp. 3779–3783.

[18] Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli, "Unsupervised cross-lingual representation learning for speech recognition," *arXiv preprint arXiv:2006.13979*, 2020.

[19] Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, et al., "Xls-r: Self-supervised cross-lingual speech representation learning at scale," *arXiv preprint arXiv:2111.09296*, 2021.

[20] Anirudh Gupta, Harveen Singh Chadha, Priyanshi Shah, Neeraj Chhimwal, Ankur Dhuriya, Rishabh Gaur, and Vivek Raghavan, "Clsril-23: cross lingual speech representations for indic languages," *arXiv preprint arXiv:2107.07402*, 2021.

[21] Sandy Ritchie, You-Chi Cheng, Mingqing Chen, Rajiv Mathews, Daan van Esch, Bo Li, and Khe Chai Sim, "Large vocabulary speech recognition for languages of africa: multilingual modeling and self-supervised learning," *arXiv preprint arXiv:2208.03067*, 2022.

[22] Yu Zhang, Ekapol Chuangsuwanich, and James Glass, "Language id-based training of multilingual stacked bottleneck features," in *Proc. Interspeech*. Citeseer, 2014, pp. 1–5.

[23] Samuel Thomas, Kartik Audhkhasi, Jia Cui, Brian Kingsbury, and Bhuvana Ramabhadran, "Multilingual data selection for low resource speech recognition," Tech. Rep., IBM THOMAS J WATSON RESEARCH CENTER YORKTOWN HEIGHTS NY YORKTOWN HEIGHTS ..., 2016.

[24] Yanmin Qian and Zhikai Zhou, "Optimizing data usage for low-resource speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 394–403, 2022.

[25] Vijayaditya Peddinti, Daniel Povey, and Sanjeev Khudanpur, "A time delay neural network architecture for efficient modeling of long temporal contexts," in *Sixteenth annual conference of the international speech communication association*, 2015.

[26] Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M Tyers, and Gregor Weber, "Common voice: A massively-multilingual speech corpus," *arXiv preprint arXiv:1912.06670*, 2019.

[27] Diederik P Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," 2015.

[28] Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli, "fairseq: A fast, extensible toolkit for sequence modeling," *arXiv preprint arXiv:1904.01038*, 2019.

[29] Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *Proceedings of the 23rd international conference on Machine learning*, 2006, pp. 369–376.

[30] Kenneth Heafield, "Kenlm: Faster and smaller language model queries," in *Proceedings of the sixth workshop on statistical machine translation*, 2011, pp. 187–197.

[31] Vineel Pratap, Awni Hannun, Qiantong Xu, Jeff Cai, Jacob Kahn, Gabriel Synnaeve, Vitaliy Liptchinsky, and Ronan Collobert, "Wav2letter++: A fast open-source speech recognition system," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6460–6464.