Comparing Self-Supervised Pre-Training and Semi-Supervised Training for Speech Recognition in Languages with Weak Language Models

Comparing Self-Supervised Pre-Training and Semi-Supervised Training for Speech Recognition in Languages with Weak Language Models

Léa-Marie Lam-Yee-Mui[1,2], *Lucas Ondel Yang*[1], *Ondřej Klejch*[3]

[1] University of Paris-Saclay, CNRS, LISN, France,
[2] Vocapia Research, France
[3] The Centre for Speech Technology Research, University of Edinburgh, UK

{lea-marie.lam-yee-mui,lucas.ondel-yang}@lisn.upsaclay.fr, o.klejch@ed.ac.uk

Abstract

This paper investigates the potential of improving a hybrid automatic speech recognition model trained on 10 hours of transcribed data with 200 hours of untranscribed data in low-resource languages. First, we compare baseline methods of cross-lingual transfer with MFCC features and features extracted with the multilingual self-supervised model XLSR-53. Subsequently, we compare two approaches that can leverage the untranscribed data: semi-supervised training with LF-MMI and continued self-supervised pre-training of XLSR-53. Our results on well-resourced English broadcast data derived from MGB show that both methods achieve 18% and 27% relative improvements compared to the baseline, respectively. On the low-resource South African Soap Opera dataset, the relative improvement with semi-supervised training is only 3% due to the inherently weak language model. However, continued pre-training achieves 8.6% relative improvement because it does not rely on any external information.

Index Terms: Low-resource automatic speech recognition, self-supervised training, semi-supervised training

1. Introduction

Automatic speech recognition (ASR) systems have recently demonstrated great accuracy improvements in well-resourced languages [1, 2, 3]. This accuracy has been achieved thanks to the modelling improvements and hundreds of thousands of hours of transcribed speech. However, ASR performance in low-resource languages is still lacking due to limited amounts of transcribed speech for training of acoustic models and limited amounts of text for the training of language models. This lack of training data in low-resource languages is even further exacerbated by code-switching between embedded and matrix languages [4]. In this work, we study how self-supervised training [1, 2, 5] and semi-supervised training [6, 7] can be used to leverage untranscribed audio data to improve the performance of ASR models for underrepresented languages. Consequently, we address the case where only small amounts of manually transcribed speech with an inherently weak language models are available, due to code-switching and the lack of text corpora.

In this paper, we use the South African Soap Opera dataset [8] which contains 14.3 hours of code-switched speech between four Bantu languages (Sesotho, Setswana, isiXhosa, and isiZulu) and English. Building a good language model for this domain is difficult due to code-switching, variation in orthography and lack of text data on the internet. To improve the ASR performance in these languages, previous works explored transfer learning with a multilingual model [9] and semi-supervised training with a weak language model [10, 11]. Following these works, we focus on building a five-lingual model for the South African languages and on taking advantage of 200 hours of untranscribed data for self-supervised and semi-supervised training. We also run contrast experiments on British English broadcast data from the Multi-Genre Broadcast (MGB) Challenge [12], for which we can train a strong language model using the provided historical BBC subtitles. We artificially mimic low-resource settings by sampling 10 hours of transcribed data and another 200 hours as untranscribed data from the MGB dataset. In our experiments on the South African Soap Opera dataset and the English MGB dataset, we show that:

- when training the seed five-lingual South African acoustic model, cross-lingual transfer from a matching domain in a well-resource language (MGB) works better than transfer from a mismatched domain in the same languages (NCHLT).
- cross-lingual transfer is comparable with using multilingual self-supervised features extracted from XLSR-53.
- both semi-supervised training and continued self-supervised pre-training work in well-resourced settings.
- continued self-supervised pre-training works better than semi-supervised training with an inherently weak language model for code-switched speech in South African languages.
- continued self-supervised pre-training and semi-supervised training are complementary even in low-resource languages.

2. Related work

In this section, we review some common methods in speech recognition for low-resource languages: cross-lingual transfer, self-supervised pre-training and semi-supervised training.

2.1. Cross-Lingual transfer

In cross-lingual transfer, we first train an acoustic model for a set of well-resourced languages and then transfer the parameters of the acoustic model to the new low-resource language. Then, we train the final acoustic model by fine-tuning a subset of the parameters on a small amount of available transcribed data [13, 14]. These multilingual models can also be used to extract bottleneck features which are used to train another model with data from a low-resource language[15].

2.2. Self-Supervised Training

In self-supervised learning (SSL), the acoustic model parameters are pre-trained on thousands of hours of untranscribed data. The model learns to recognize latent speech representations from raw signal with a contrastive loss, such as InfoNCE [16]. Architectures based on convolutional layers and Transformers [17] have been proposed and pre-trained with English datasets, such as wav2vec2.0 [1] and HuBERT [18]. For

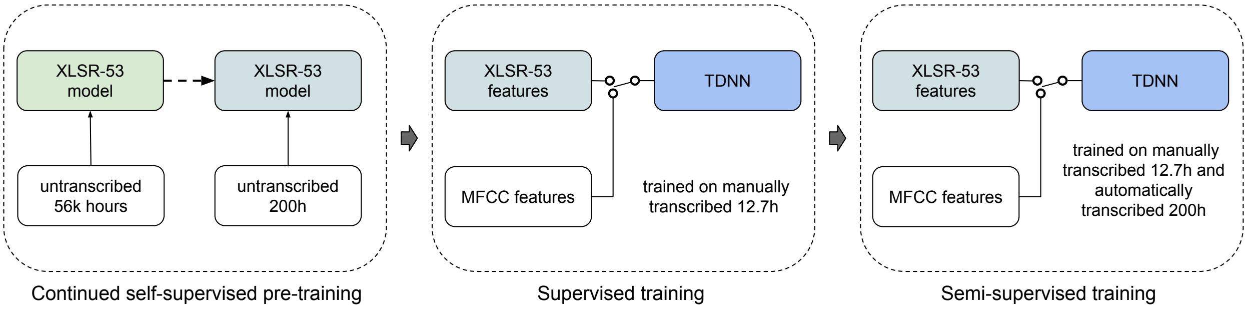

Figure 1: *Diagram of the combination of self-supervised and semi-supervised trainings for* 12.7 *hours of transcribed South African data and* 200 *hours of untranscribed data from other South African soap operas. XLSR*-53 *is further trained with the* 200 *hours of untranscribed data (continued self-supervised pre-training) and used as features extractor to train a TDNN. The TDNN model can also be trained with MFCC features. The obtained TDNN model is the seed model for the subsequent semi-supervised training with both transcribed and untranscribed data. We follow the same procedure when running experiments on MGB.*

speech recognition, these pre-trained self-supervised models are usually fine-tuned on the transcribed training data with a standard supervised loss such as Connectionist Temporal Classification (CTC) loss [19], or lattice-free maximum mutual information (LF-MMI) loss [20, 21]. A multilingual pre-trained model XLSR-53 [2], which is based on the wav2vec2.0 architecture [1], performs well when used to train ASR models for low-resource languages [2, 22, 23, 24]. However, these large pre-trained models can still suffer from mismatch between training and testing conditions. Therefore, continued pre-training with untranscribed data from the target domain can be used to alleviate the domain mismatch [5, 25]. This procedure of continual pre-training [26] alleviates the domain mismatch while requiring orders of magnitude less data by retaining the knowledge from the original dataset. It makes this approach especially useful for low-resource languages.

2.3. Semi-Supervised Training

In semi-supervised training, we start by training a seed model on the transcribed training data. Then, we use this acoustic model with a language model to produce pseudo-labels for the untranscribed data. These pseudo-labels are then used as training targets for fine-tuning the acoustic model [6]. Traditional approaches for semi-supervised training use one-best transcripts as pseudo-labels. However, these transcripts might contain transcription errors negatively affecting the acoustic model. Therefore, it is necessary to perform some form of confidence filtering to discard utterances with noisy transcripts [27]. Another option is to use lattices as pseudo-labels [7, 28]. This approach circumvents the issue of erroneous one-best transcript by representing possible alternative transcriptions and their corresponding uncertainties within the lattice. Popular way of performing lattice-based semi-supervised training is to use lattice-free maximum mutual information criterion [20, 7]. The limiting factor of semi-supervised training is the quality of the language model used to produce the pseudo-labels [29]. A good language model typically needs to be trained on large amounts of text data; in our experience, at least several hundred million words are needed to train a strong general language model. Unfortunately, such large amounts of text are not available online for many low-resource languages [30]. Furthermore, the language modelling in these low-resource languages can further be exacerbated by the presence of code-switching [31]. The performance of semi-supervised training can be improved by starting from a stronger seed acoustic model, trained with cross-lingual transfer [32] or fine-tuned from a self-supervised model [33].

3. Improving the Acoustic Model with Untranscribed Data

The performance of ASR systems in low-resource languages is limited by the small amount of manually transcribed data. In this section, we describe how we collected untranscribed speech and how we used it to improve the performance of ASR systems using self-supervised training and semi-supervised training. Our approach consisted of three steps. First, we used continued self-supervised pre-training of XLSR-53 on our untranscribed speech data. Then, we used the adapted XLSR-53 model to extract features for training of the seed model. Finally, we used this seed model to produce better pseudo-labels for semi-supervised training on the untranscribed speech data. The whole pipeline is illustrated in Figure 1.

3.1. Datasets

We conducted experiments on the South African Soap Operas dataset [8]. This dataset contains 14.3 hours of transcribed code-switched speech with people alternating between four Banto languages and English. We used the official training (12.7 hours) and test splits (1.3 hours). The represented pairs of languages are: English-Sesotho (eng-sot), English-Setswana (eng-tsn), English-isiXhosa (eng-xho) and English-isiZulu (eng-zul). We used the training transcripts (155k tokens) to train a LM. Note that we tried to train a stronger language model using text crawled from internet, but the resulting model was worse than the one trained only on the training transcripts.

We also collected untranscribed speech for these South African languages. To avoid noisy collected data for South African languages, we identified South African soap operas on Wikipedia and we downloaded trailers for these soap operas. This method ensured that the crawled data was in-domain and contained relevant languages together with other South African languages used in the soap operas. In total we were able to collect 200 hours of raw recordings which were segmented with WebRTC VAD[1] for further processing.

As a contrast, we also experimented with the BBC broadcasts from the MGB dataset [12]. We selected 10 hours from

[1] https://github.com/wiseman/py-webrtcvad

the MGB dataset for the training dataset and used another 200 hours from the MGB dataset as untranscribed data. We trained a language model on the provided BBC subtitles (650M words). We evaluated the performance of the models trained with the official MGB development set using the manual segmentation.

3.2. ASR model training

We trained a five-lingual (four Bantu languages + English) South African acoustic models which used either 40-dimensional MFCC features or 1024-dimensional XLSR-53 features as inputs. Both types of models were trained using Kaldi toolkit [34] and used the same alignments obtained with a standard GMM model. We used a CNN-TDNN architecture with 16.7M parameters for the acoustic model using MFCC features and a TDNN-F architecture with 23.7M parameters for the acoustic model using XLSR-53 features. We did not use convolutional layers with the XLSR-53 features because the XLSR-53 model acts as a convolutional front-end. Both types of models were trained with the LF-MMI criterion [20] for six epochs on speed-perturbed training data. The pronunciation dictionaries for the four South African languages and English were built using the NCHLT dictionaries and corresponding grapheme-to-phoneme rules [35]. The final merged dictionary had 88 phones, including 12 phones shared by all five languages. The vocabulary size was 18.8k tokens. Furthermore, we trained a 3-gram language model on the available training transcripts.

To improve the performance of the model using MFCC features, we transferred parameters from a model trained either on the NCHLT dataset [35], which contains read speech from 11 official South African languages, or on the 200 hours of English broadcasts from the MGB dataset [12]. To deal with the mismatch in acoustic units, we replaced the final layer of the pre-trained model and retrained the whole acoustic model. Based on our experience with fine-tuning of hybrid models, we used a 10-times smaller learning rate than during training from scratch, for all layers except for the newly initialized final layer.

The acoustic model for MGB has the same CNN-TDNN or TDNN-F architecture depending on the input features. The language model for MGB was trained on all available MGB subtitles making it a very strong language model.

3.3. Continued self-supervised pre-training

We used the collected 200 hours of South African speech untranscribed data or the 200 hours of MGB for self-supervised training. However, since training the self-supervised models from scratch is computationally expensive and requires a lot of data, which might not be available for low-resource languages, we only performed continued pre-training of a self-supervised model using these 200 hours with the contrastive loss of wav2vec2.0 [1]. As pre-trained model, we chose to use the multilingual model XLSR-53, which is based on wav2vec2.0 LARGE architecture and is trained on 53 languages using a total of 56k hours of training data. Instead of using the pre-trained model directly as an acoustic model, we used it as a multilingual bottleneck feature extractor. We experimented with various pre-trained self-supervised models and extracted the features from different layers and we found that the last layer of XLSR-53 worked best in our scenario.[2] We extracted these representations from the last layer with the S3PRL toolkit [36] and used them as inputs for a standard hybrid TDNN-F model [37]. We performed the continued self-supervised pre-training of the XLSR-53 model with the fairseq toolkit [38] using the 200 hours of untranscribed data. We kept the hyperparameters identical to the ones used to train wav2vec2.0 LARGE [1] on Librivox. We continued pre-training for 8k iterations equivalent to 67 epochs, with a batch size of at most 1.4M tokens and we used gradient accumulation to simulate training on a bigger batch size using only two Tesla V100 GPUs.

3.4. Semi-supervised training

In semi-supervised training, we used the seed acoustic model trained on the manually transcribed data together with a language model trained on available text data to produce pseudo-labels for the untranscribed speech. The semi-supervised training was done using the lattice-free maximum mutual information (LF-MMI) training criterion [20] and followed the semi-supervised training approach proposed in [7]. We performed semi-supervised training on a combination of the manually transcribed training data and the untranscribed 200 hours. We decoded the untranscribed data with the seed model trained on the transcribed data and we used the decoded lattices as pseudo-labels for semi-supervised training. Note that, we filtered the untranscribed data with the minimum mean recording confidence threshold of 0.8 and the minimum speaking rate threshold 1.25 words per second prior to the semi-supervised training as suggested in [39]. We trained the semi-supervised models for six epochs.

4. Results

4.1. Baselines acoustic models

In the first set of experiments, we assessed how well cross-lingual transfer works for South African languages. We compared training the acoustic model from scratch, denoted as (1) in Table 1, using cross-lingual transfer from a model trained on the NCHLT dataset (2) and using cross-lingual transfer from a model trained on the MGB challenge dataset (3). Our results demonstrated that a domain-match and data diversity in the MGB dataset (3) is more important than training on additional data for the target languages (2) with overall word error rates (WER) of 50.8% and 52.0% respectively. This is because the NCHLT data contains clean read speech, which is acoustically very different from the speech in the Soap Operas dataset. In addition to the noticeable background noise in the Soap Operas dataset, the speech characteristics are also very different, for example the average speaking rate in the NCHLT dataset is three times slower than in the Soap Operas dataset. Subsequently, we compared the model trained with cross-lingual transfer from MGB (3) with a model using the multilingual self-supervised representations obtained with XLSR-53 as input features (5). We found that these two approaches achieved similar WER of 50.8% and 50.9%. The benefit of (3) is that it is trained on very well matched data and (5) benefits from being pre-trained on large amounts of multilingual data. We hypothesize that (3) would improve even further if cross-lingual transfer was done from a model trained on a diverse multilingual dataset, not only on British English dataset.

4.2. Semi-supervised vs self-supervised training

In the next set of experiments, we investigated how 200 hours of untranscribed data can help improve the performance of the

[2] Note that XLS-R [22] achieved better results than XLSR-53 in our preliminary experiments. However due to its size we were not able to continue pre-training it and therefore we left it for future work.

Table 1: *Word Error Rate (WER) on the test set of the South African Soap Operas dataset and the development set of MGB. For the South African Soap Operas, the results are split by language pairs: English-Sesotho (eng-sot), English-Setswana (eng-tsn), English-isiXhosa (eng-xho) and English-isiZulu (eng-zul).*

		South African Soap Operas					**MGB**
		eng-sot	eng-tsn	eng-xho	eng-zul	all	dev
(1)	CNN-TDNN baseline	55.9	46.6	63.9	56.6	54.7	29.4
(2)	CNN-TDNN with cross-lingual transfer from NCHLT	52.8	44.5	60.4	54.1	52.0	-
(3)	CNN-TDNN with cross-lingual transfer from MGB	50.1	43.8	60.8	52.8	50.8	-
(4)	+ semi-supervised training	**48.3**	**42.6**	**59.2**	**51.1**	**49.3**	**24.0**
(5)	TDNN-F using XLSR-53 bottleneck features	49.8	45.1	61.7	51.6	50.9	25.4
(6)	+ continued self-supervised pre-training	45.5	40.9	**56.2**	47.6	46.5	21.2
(7)	+ semi-supervised-training	**44.3**	**38.5**	56.5	**46.4**	**45.2**	**19.6**

initial acoustic model. We observed that semi-supervised training with the seed models using MFCC features (3) in Table 1 performed worse than continued self-supervised pre-training of the multilingual model XLSR-53 (6). It is worth noting that it was crucial to combine the transcribed 10 hours of data with the 200 hours of untranscribed data to make semi-supervised training work on the South African dataset. Without the transcribed data the semi-supervised model was worse than the seed model. When we combined the continued self-supervised pre-training with semi-supervised training we achieved further gains (7). The WER of the Soap Operas dataset is reduced by 11% relative compared to the baseline model trained with cross-lingual transfer from MGB (3 → 7) while the WER of the MGB dataset has a higher relative 33% gain compared to the baseline trained with MFCC from a flat-start initialization (1 → 7), thanks to the stronger language model. Also note that the comparison between semi-supervised training and continued self-supervised pre-training are not completely fair, because XLSR-53 is a much bigger model than our CNN-TDNN acoustic model. We plan to conduct a fair comparison in future.

To overcome the code-switched aspect of the Soap Operas dataset, we tried using a language model with South African monolingual texts crawled from the web and MGB [12] transcriptions but the resulting WER were worse than using the in-domain language model trained only on the Soap Operas transcriptions. This shows that the language model domain match is very important, especially since this in-domain language follows clear transcription conventions consistent with the test transcription. This explains why the in-domain language model is stronger because not all South African languages have standardised orthography and the crawled online texts might follow different spelling rules, which negatively affects the WER. Code-switching is also a mostly unwritten phenomenon, which makes language modelling even more difficult. Furthermore, the performance of the model could be improved by fine-tuning the model for each language pair individually, but we chose not to do it since we were interested in building a unified five-lingual model. Previous works on this South African data indeed demonstrated that the semi-supervised training batch by batch yields improvement over training in a single pass, as does training bilingual acoustic models instead of a five-lingual model [10]. Improvement also comes from adding generated texts [40] and automatic transcriptions to build a strong LM. However, this type of bilingual training with a strong LM is less practical because it requires to obtain untranscribed data with a specific type of code-switching to train the model.

5. Conclusions

In this paper, we explored using untranscribed speech data to improve the accuracy of the speech recognition, using small amounts of manually transcribed speech data. We added a constraint of using a weak language model coming from the difficulty of finding accurate code-switched training texts to build a stronger language model for South African languages. We also evaluated our approach on a simulated low-resource setting on English, with a strong language model. We found that the best approach to improve the initial acoustic models is using features from the multilingual XLSR-53 model with continued self-supervised pre-training with the unstranscribed data, which does not require any language model. This approach is particularly useful in cases like South African code-switching, where we can only train a weak language model due to the lack of sufficient amount of in-domain text. When we subsequently used this model as a seed model for semi-supervised training, we obtained a relative improvement of 11% relative compared to the baseline model trained with cross-lingual transfer from MGB for the Soap Operas dataset while for English, we obtained a relative gain of 33% compared to the baseline trained with MFCC. The improvements on the Soap Operas dataset were much smaller than we would expect with a language model with external language resources. For this reason, in the future, we would like to explore ways of training better language models in low-resource and especially code-switched settings to improve the performance during semi-supervised training and decoding. Finally, we would also like to follow [5] and use better regularization methods during continued self-supervised pre-training. Both these methods should allow for a more efficient continued pre-training.We believe that our findings will be applicable to other low-resource languages with limited amounts of text corpora available.

6. Acknowledgements

This work was supported by JSALT 2022 at JHU, with gift-funds from Amazon, Microsoft and Google. We would like to thank Ahmed Ali, Hung-yi Lee and Matthew Wiesner for their insightful comments and support during the JSALT 2022 workshop. Ondřej Klejch was also supported by EPSRC Project EP/T024976/1 (Unmute). For the purpose of open access, the author has applied a Creative Commons Attribution (CC BY) licence to any Author Accepted Manuscript version arising from this submission.

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 *NeurIPS*, 2020.

[2] A. Conneau, A. Baevski, R. Collobert, A. Mohamed, and M. Auli, "Unsupervised Cross-Lingual Representation Learning for Speech Recognition," in *Interspeech*, 2021.

[3] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," *arXiv preprint arXiv:2212.04356*, 2022.

[4] S. Sitaram, K. R. Chandu, S. K. Rallabandi, and A. W. Black, "A survey of code-switched speech and language processing," *arXiv preprint arXiv:1904.00784*, 2019.

[5] J.-H. Lee, C.-W. Lee, J.-S. Choi, J.-H. Chang, W. K. Seong, and J. Lee, "CTRL: Continual Representation Learning to Transfer Information of Pre-trained for WAV2VEC 2.0," *Interspeech*, 2022.

[6] L. Lamel, J.-L. Gauvain, and G. Adda, "Lightly supervised and unsupervised acoustic model training," *Computer Speech and Language*, vol. 16, no. 1, pp. 115–229, 2002.

[7] V. Manohar, H. Hadian, D. Povey, and S. Khudanpur, "Semi-supervised training of acoustic models using lattice-free MMI," in *ICASSP*, 2018.

[8] E. van der Westhuizen and T. Niesler, "A first South African corpus of multilingual code-switched soap opera speech." in *LREC*, 2018.

[9] E. Yılmaz, A. Biswas, E. van der Westhuizen, F. de Wet, and T. Niesler, "Building a Unified Code-Switching ASR System for South African Languages," in *Interspeech*, 2018.

[10] A. Biswas, E. Yilmaz, F. De Wet, E. Van der westhuizen, and T. Niesler, "Semi-supervised Development of ASR Systems for Multilingual Code-switched Speech in Under-resourced Languages," in *LREC*, 2020.

[11] N. Wilkinson, A. Biswas, E. Yilmaz, F. De Wet, E. Van der westhuizen, and T. Niesler, "Semi-supervised Acoustic Modelling for Five-lingual Code-switched ASR using Automatically-segmented Soap Opera Speech," in *SLTU & CCURL*, 2020.

[12] P. Bell, M. J. Gales, T. Hain, J. Kilgour, P. Lanchantin, X. Liu, A. McParland, S. Renals, O. Saz, M. Wester *et al.*, "The MGB challenge: Evaluating multi-genre broadcast media recognition," in *ASRU*, 2015.

[13] J.-T. Huang, J. Li, D. Yu, L. Deng, and Y. Gong, "Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers," in *ICASSP*, 2013.

[14] A. Ghoshal, P. Swietojanski, and S. Renals, "Multilingual training of deep neural networks," in *ICASSP*, 2013.

[15] K. Veselý, M. Karafiát, F. Grézl, M. Janda, and E. Egorova, "The language-independent bottleneck features," in *SLT*, 2012.

[16] A. van den Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *CoRR*, vol. abs/1807.03748, 2018.

[17] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is All you Need," in *NeurIPS*, 2017.

[18] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.

[19] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *ICML*, 2006.

[20] D. Povey, V. Peddinti, D. Galvez, P. Ghahremani, V. Manohar, X. Na, Y. Wang, and S. Khudanpur, "Purely sequence-trained neural networks for ASR based on lattice-free MMI," in *Interspeech*, 2016.

[21] A. Vyas, S. Madikeri, and H. Bourlard, "Comparing CTC and LF-MMI for out-of-domain adaptation of wav2vec 2.0 acoustic model," in *Interspeech*, 2021.

[22] A. Babu, C. Wang, A. Tjandra, K. Lakhotia, Q. Xu, N. Goyal, K. Singh, P. von Platen, Y. Saraf, J. Pino, A. Baevski, A. Conneau, and M. Auli, "XLS-R: Self-supervised Cross-lingual Speech Representation Learning at Scale," in *Interspeech*, 2022.

[23] K. D. N, P. Wang, and B. Bozza, "Using Large Self-Supervised Models for Low-Resource Speech Recognition," in *Interspeech*, 2021.

[24] M. Wiesner, D. Raj, and S. Khudanpur, "Injecting Text and Cross-Lingual Supervision in Few-Shot Learning from Self-Supervised Models," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2022, pp. 8597–8601, iSSN: 2379-190X.

[25] S. Kessler, B. Thomas, and S. Karout, "An Adapter Based Pre-Training for Efficient and Scalable Self-Supervised Speech Representation Learning," in *ICASSP*, 2022.

[26] G. I. Parisi, R. Kemker, J. L. Part, C. Kanan, and S. Wermter, "Continual lifelong learning with neural networks: A review," *Neural Networks*, vol. 113, no. C, May 2019.

[27] H. Y. Chan and P. Woodland, "Improving broadcast news transcription by lightly supervised discriminative training," in *ICASSP*, 2004.

[28] T. Fraga-Silva, J.-L. Gauvain, and L. Lamel, "Lattice-based unsupervised acoustic model training," in *ICASSP*, 2011.

[29] E. Wallington, B. Kershenbaum, P. Bell, and O. Klejch, "On the learning dynamics of semi-supervised training for ASR," in *Interspeech*, 2021.

[30] I. Caswell, T. Breiner, D. van Esch, and A. Bapna, "Language id in the wild: Unexpected challenges on the path to a thousand-language web text corpus," in *ACL*, 2020.

[31] T. Reitmaier, E. Wallington, D. Kalarikalayil Raju, O. Klejch, J. Pearson, M. Jones, P. Bell, and S. Robinson, "Opportunities and challenges of automatic speech recognition systems for low-resource language speakers," in *CHI*, 2022.

[32] A. Carmantini, P. Bell, and S. Renals, "Untranscribed web audio for low resource speech recognition." in *Interspeech*, 2019.

[33] 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.

[34] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The Kaldi Speech Recognition Toolkit," in *ASRU*, 2011.

[35] E. Barnard, M. H. Davel, C. van Heerden, F. De Wet, and J. Badenhorst, "The nchlt speech corpus of the south african languages," in *SLTU*, 2014.

[36] S.-W. Yang, P.-H. Chi, Y.-S. Chuang, C.-I. J. Lai, K. Lakhotia, Y. Y. Lin, A. T. Liu, J. Shi, X. Chang, G.-T. Lin, T.-H. Huang, W.-C. Tseng, K. tik Lee, D.-R. Liu, Z. Huang, S. Dong, S.-W. Li, S. Watanabe, A. Mohamed, and H. yi Lee, "SUPERB: Speech Processing Universal PERformance Benchmark," in *Interspeech*, 2021.

[37] D. Povey, G. Cheng, Y. Wang, K. Li, H. Xu, M. Yarmohammadi, and S. Khudanpur, "Semi-Orthogonal Low-Rank Matrix Factorization for Deep Neural Networks," in *Interspeech*, 2018.

[38] 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 *NAACL-HLT: Demonstrations*, 2019.

[39] O. Klejch, E. Wallington, and P. Bell, "The CSTR System for Multilingual and Code-Switching ASR Challenges for Low Resource Indian Languages," in *Interspeech*, 2021.

[40] J. J. van Vüren and T. Niesler, "Code-Switched Language Modelling Using a Code Predictive LSTM in Under-Resourced South African Languages," in *SLT*, 2022.