A Novel Self-training Approach for Low-resource Speech Recognition

*Satwinder Singh, Feng Hou, Ruili Wang**

School of Mathematical and Computational Sciences, Massey University, Auckland, New Zealand

{s.singh4,f.hou,ruili.wang}@massey.ac.nz

Abstract

In this paper, we propose a self-training approach for automatic speech recognition (ASR) for low-resource settings. While self-training approaches have been extensively developed and evaluated for high-resource languages such as English, their applications to low-resource languages like Punjabi have been limited, despite the language being spoken by millions globally. The scarcity of annotated data has hindered the development of accurate ASR systems, especially for low-resource languages (e.g., Punjabi and Māori languages). To address this issue, we propose an effective self-training approach that generates highly accurate pseudo-labels for unlabeled low-resource speech. Our experimental analysis demonstrates that our approach significantly improves word error rate, achieving a relative improvement of 14.94% compared to a baseline model across four real speech datasets. Further, our proposed approach reports the best results on the Common Voice Punjabi dataset.

Index Terms: self-training, low-resource, Punjabi ASR

1. Introduction

The development of end-to-end (E2E) automatic speech recognition (ASR) systems for high-resource languages has made significant strides in recent years, resulting in an excellent performance [1]. The key factors contributing to this high performance include the presence of large annotated datasets, advancements in deep learning algorithms, and easy access to high-performance computational resources. However, the same cannot be said for ASR systems designed for low-resource languages. These systems still fall short in terms of performance when compared to their high-resource counterparts. The primary reason for this discrepancy is the lack of annotated speech datasets and their availability in the public domain. Without these resources, it is difficult for ASR systems to be trained effectively and perform at the same level as those designed for high-resource languages.

The research community has been actively exploring various ways to address the scarcity of annotated datasets for speech recognition systems [2–6]. The process of annotating highly accurate speech datasets is a labor-intensive task that requires a significant amount of time and resources. To overcome these challenges, many researchers have proposed self-training or pseudo-labeling (PL) based methods [7]. These methods leverage the vast amount of available unlabeled data in conjunction with a small labeled dataset to improve the performance of ASR systems. Self-training has gained increasing interest in a wide range of research areas, including computer vision [8, 9], natural language processing [10, 11], and more recently, in the field of speech recognition [2, 12–15]. These methods have shown promise in addressing the shortage of annotated data and improving the performance of ASR systems.

The fundamental concept behind self-training is to first train an initial seed model, either from scratch or by fine-tuning an existing pre-trained model, using a limited amount of labeled data. This seed model is then used to generate pseudo-labels for unlabeled data. The process can be repeated over multiple iterations in order to improve the quality of the generated pseudo-labels. To ensure that only highly accurate pseudo-labels are used, various filtering methods are applied to filter out any incorrect labels [12]. Additionally, a robust language model can be utilized to achieve accurate decoding of ASR transcriptions, further improving the quality of the generated pseudo-labels.

Most of the previous and recent work in self-training/pseudo-labeling is based on and demonstrated on high-resource languages like English [2, 12–14, 16]. However, most of the languages of the world are low-resource and the ASR system designed for them performs poorly when compared to those with high-resource languages. With that in mind, in this work, we focus to design a self-training approach for one of the most widely spoken languages in the world, i.e., the Punjabi language. Although the Punjabi language is spoken natively across India and Pakistan region, however, Punjabi speakers are spread across the globe, especially in countries such as Canada, the United States, the United Kingdom, Australia, and New Zealand. Despite having more than 100 million speakers, the Punjabi language still lacks mature ASR systems due to a lack of annotated datasets.

In the paper, we propose a self-training approach to leverage the available large unlabeled audio data. We adopt a pre-trained self-supervised crosslingual wav2vec 2.0 model (XLSR-53) [17] as our seed model to generate pseudo-labels across multiple iterations. The proposed model first fine-tunes the XLSR-53 seed model on limited Punjabi datasets in a supervised fashion with language model decoding. Afterwards, the fine-tuned model is used to produce pseudo-labels for unlabeled audio data. As we are fine-tuning on limited data, it is common to get a mixed bag of pseudo-labels. To sieve out the erroneous pseudo-labels, we propose a length-normalized confidence score. The pseudo-labels with high confidence scores above the certain cut-off threshold are selected to be included into the dataset. Finally, the model is fine-tuned again from scratch using labeled data and highly confident pseudo-labels. This process is repeated over multiple iterations with gradual confidence score filtration to refine the pseudo-labels. Our empirical results demonstrate that our proposed approach is able to achieve significant WER reduction compared with the baseline and also reports the best results on the Common Voice Punjabi dataset. To the best of our knowledge, this work is the first to explore the self-training approach for the Punjabi language on the mix of datasets (public, non-public, and synthesized datasets).

2. Related Work

Semi-supervised approaches such as self-supervision and self-training have been active areas of research for the past few years. The goal of these approaches is to leverage the large unlabeled speech data to improve the overall accuracy of the ASR systems. These approaches showed significant results for low-resource languages. Self-training is a fairly simple yet effective semi-supervised approach for utilizing unlabeled audio data. Self-training has been applied in computer vision and natural language processing tasks. However, it is recently adopted for E2E speech recognition.

Kahn et al. [7] proposed a self-training approach for E2E ASR model. The proposed system incorporate a pseudo-label filtering method and language model decoding to improve the word error rates. Additionally, an ensemble of multiple models showed further performance improvement. Further, inspired by noisy student training (NST) from the image processing domain, Park et al. [12] adopted this method for speech recognition. The proposed approach exploited the adaptive SpecAugment [18] augmentation method. The NST approach incorporated filtering, balancing, and mixing techniques to carefully select highly accurate pseudo-labels. Furthermore, rather than selecting pseudo-labels over just one iteration, iterative pseudo-labeling (IPL) [13] refines the acoustic model over multiple iterations. IPL effectively generates the pseudo-labels for a subset of unlabeled data rather than labeling the entire unlabeled data. Also, IPL fine-tuned an existing model instead of training an entirely new model.

Recently, Higuchi et al. [14] presented momentum pseudo-labeling (MPL), which incorporated online and offline models. At different time steps, weights of multiple models are averaged using the momentum-based moving average to refine pseudo-labels. The online model (also called the teacher model), hypothesizes the pseudo-labels produced by the offline model (also referred to as the student model) on the go. The MPL approach outperformed standard pseudo-labels and IPL approaches. The extension to this work is presented in [19], where MPL is combined with IPL. The resulting approach improved the seed model by exploiting language model knowledge using the IPL strategy. Additionally, the Conformer-based ASR architecture further enhanced the overall accuracy of the system.

While most of the self-training approaches utilized language models for decoding, the SlimIPL [16] method performed pseudo-labeling without any language model. The proposed model is built upon the IPL approach, which produced pseudo-labels over multiple iterations using a single model. SlimIPL also incorporated a dynamic cache for better and more stable model training. Furthermore, Lugosch et al. [20] presented a self-training approach to multilingually generate pseudo-labels for 60 languages. The proposed approach first pre-trained a single multilingual model followed by fine-tuning it on the target language using semi-supervision. The fine-tuned model is then used to produce pseudo-labels for that particular language, which in turn is utilized to train the final model. Overall, the multilingual pseudo-labeling demonstrated good performance and better generalization to the data from different domains.

Xu et al. [21] presented a study, which showed that self-training and pre-training can effectively be combined and are complementary to each other. Pre-training based self-supervised approaches like wav2vec [1] and its variants demonstrated excellent performance for low-resource languages. They leveraged unlabeled audio data and learn representations using contrastive learning. When combined with self-training, self-supervision showed a better generation of pseudo-labels and achieved significant results. Similarly, Jin et al. [22] exploited a self-supervised pre-trained wav2vec 2.0 model for pseudo-labeling. The proposed method filtered out the low-probability pseudo-labels and only selected the pseudo-labels with the highest probability scores. Further, self-training has also been used to improve representation learning in crosslingual settings [23]. Furthermore, Khurana et al. [24] proposed the DUST (Dropout-based Uncertainty-driven Self-Training) approach to alleviate the issues caused by domain mismatch. The proposed approach efficiently rejected pseudo-labels that exhibit high levels of uncertainty.

Until recently, limited work has been carried out in the Punjabi language. Previous efforts relied heavily on statistical and hybrid models, using mostly non-public datasets [25–28]. However, in our previous work [6], we demonstrated that E2E models trained with synthetic speech and real speech datasets could enhance the effectiveness of E2E ASR models for Punjabi.

3. Proposed Approach

We propose a very simple, yet effective self-training method for the Punjabi language. However, the proposed method could work for any language having a small amount of annotated data and large unlabeled audio data.

3.1. Pre-trained seed model

For the strong baseline seed model, we adopt a self-supervised pre-trained wav2vec 2.0 based crosslingual XLSR-53 model [17]. The XLSR-53 model is pre-trained on 53 languages using a contrastive learning approach to learn audio representations in an unsupervised manner. The crosslingual wav2vec model maps raw input speech to latent speech representations, which are then transformed into a set of discrete representations using a quantization module. These discrete representations are shared across multiple languages (in this case 53 languages excluding the Punjabi language) and used as the target for the self-supervised objective. Conneau et al. [17] showed that XLSR-53 could greatly improve the character error rate (CER) and word error rate (WER) in the case of low-resource languages.

Although pre-trained XLSR-53 does not include Punjabi in the pre-training setup, the model is still able to learn powerful shared feature representation across many languages. The crosslingual nature of the model made it easy to fine-tune the model on any new language with limited labeled data. Following that, rather than pre-training the model using the Punjabi language, we leverage the effective feature representations of the pre-trained model and only fine-tune the model on available Punjabi labeled data. This way, we keep our self-training approach simple and fast.

3.2. Self-training approach

Our proposed self-training approach is fairly simple as shown in Figure 1. Initially, we fine-tune our pre-trained seed acoustic model on available limited labeled Punjabi datasets. Once the model is fine-tuned on the Punjabi language then we use this model to generate the pseudo-labels for unlabeled Punjabi data. The details of labeled and unlabeled datasets are outlined in Section 4.1. At this stage, we also employ the 5-gram KenLM language model [29] for decoding the seed model's outputs. Since the seed model is only trained on limited Punjabi
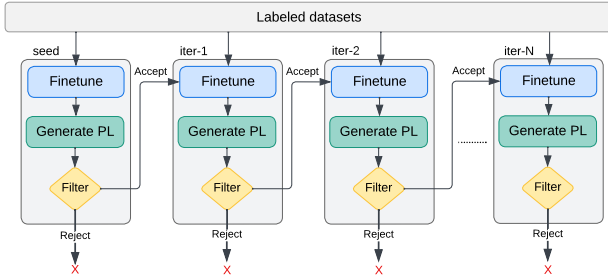
Figure 1: *Overview of proposed self-training approach.*

Table 1: *List of datasets used for our experimentations.*

Datasets	#Hours	#Utterances
Real speech		
Shrutilipi	94	50K
Common Voice	1	1210
Punjabi Speech	4	2431
50Languages	3	3955
Synthesized speech		
Google-synth	38	50K
CMU-synth	170	80K
Unlabeled speech		
Audiobooks	450	322K

data, it is fairly common to produce inaccurate pseudo-labels. To tackle the erroneous pseudo-labels, we incorporated confidence based scoring strategy [7]. Our language model produces sentence-level shallow fusion scores while decoding. We leverage the shallow fusion score as a confidence score (CS) and normalize it to the length of the sentence. We filter pseudo-labels with different filtration thresholds of confidence scores in each iteration of self-training, i.e., [0.5, 0, -0.5, -1, -1.5, -2, -2.5]. We start with a very strict (0.5) threshold and then with each iteration of self-training, we gradually relax the filtration threshold. Afterwards, we fine-tune an entirely new model using a combination of labeled and pseudo-labeled datasets. We repeat this whole process over multiple iterations until the ASR model is not improving on the labeled test datasets.

Our approach is similar to the IPL approach [13], but with few distinctions. Firstly, we compute pseudo-labels for the entire unlabeled dataset instead of a subset of it. Although it is time-consuming to pseudo-label the entire unlabeled dataset, it would give us more high-quality filtered pseudo-labels for fine-tuning. Secondly, since the XLSR-53 model does not show any improvement over multiple continued fine-tuning phases, we find that fine-tuning a new model from pre-trained XLSR-53 model produces the best results. Therefore, rather than continuing fine-tuning like in IPL, in every iteration, we fine-tune an entirely new model.

4. Experimental Setup

4.1. Datasets

To train our seed model we use a range of labeled Punjabi datasets. We utilize four low-resource Punjabi real speech datasets and two synthesized datasets. For real speech datasets, we use the Shrutilipi dataset [30], Common Voice (CV) [31], Punjabi Speech [6], and 50Languages[1]. Previously, we have demonstrated the effectiveness of synthetic datasets in improving the performance of the Punjabi ASR system on real speech datasets [6]. Thus, following that, we use two synthesized datasets i.e. Google-synth and CMU-synth [6].

- **Shrutilipi**: The Shrutilipi corpus is a collection of audio and text pairs sourced from public resources such as All India Radio news bulletins, and compiled through data mining techniques. The corpus includes labeled speech data for a total of 12 Indian languages, including Punjabi.
- **Common Voice**: The Common Voice dataset is a crowd-sourced dataset available to the public, featuring speech data in numerous languages. For our purposes, we utilize the Punjabi data made available in the April 2022 release of Common Voice. The dataset comprises of a range of speech samples from a total of 39 male and 12 female speakers, providing a

diverse array of recordings.

- **Punjabi Speech**: The Punjabi Speech dataset is a read speech dataset like Common Voice. The dataset consists of speech data from 2 male speakers.
- **50Languages**: The dataset is extracted from the 50Languages learning platform[1]. The dataset comprises speech utterances from a single male speaker.
- **Google-synth**: The Google-synth dataset is synthesized labeled speech data produced using Google's Cloud text-to-speech (TTS) API[2]. The dataset contains speech samples from 2 male and 2 female speakers.
- **CMU-synth**: The CMU-synth dataset is synthesized using CMU's Clustergen TTS model. The dataset contains speech samples from a single female speaker.
- **Audiobooks**: The Audiobooks dataset consists of speech recordings gathered from open Punjabi audiobooks featured on various YouTube channels. It is an unlabeled dataset used primarily for self-training purposes. The dataset contains audio recordings from multiple speakers and several audiobooks. To facilitate self-training, we preprocess the audio recordings, chunking them down into smaller audio segments no longer than 15 seconds.

The synthesized datasets (Google-synth and CMU-synth) are produced using Punjabi text available in the Old Newspaper corpus[3]. We ensure that there is no overlap between data used to train the ASR model, self-training, language model, and synthesized datasets.

4.2. Methodology

For our self-training approach, we adopt a pre-trained wav2vec 2.0 based crosslingual XLSR-53 model. The XLSR-53 model is pre-trained on 53 languages (50K hours) with 300 million parameters. To save time, we only fine-tune the pre-trained model on our labeled datasets instead of pre-training it from scratch. We adopt the pre-trained model by adding a fully connected layer on top of the Transformer. The fully connected layer outputs the characters using Connectionist Temporal Classification (CTC) [32]. During the fine-tuning, we freeze the weights of the feature encoder. We fine-tune a single XLSR-53 model on combined data from 6 of the labeled datasets as mentioned in Table 1. For all the datasets except Common Voice, we use 8:1:1 splits for train, validation, and test sets. For the Common Voice dataset, we use pre-defined data splits. We fine-tune our model for 10 epochs with a batch size of 32. Further, we

---

[1]https://www.50languages.com/phrasebook/en/pa

[2]https://cloud.google.com/text-to-speech
[3]https://www.kaggle.com/datasets/alvations/old-newspapers

Table 2: *Experimental results in terms of WER (%). The best pseudo-label (PL) results are obtained by choosing the best-performing model on labeled datasets over multiple iterations of PL with different confidence thresholds.*

Model	Datasets					
	Shrutilipi	Common Voice	Punjabi Speech	50Languages	Google-synth	CMU-synth
Seed without LM	17.23	29.88	18.88	28.08	13.39	6.82
Seed with LM (baseline)	16.87	14.04	9.84	24.16	2.93	4.96
Raw PL (no filter)	15.23	12.30	9.10	24.00	2.75	5.12
Best PL	**14.67**	**11.42**	**8.57**	**20.55**	**2.69**	**4.92**

optimize the training process using the Adam optimizer [33] with an initial learning rate of 3e-4, which is warmed up for the first 10% of the updates and afterward, linearly decays during the rest of the updates. All the experiments are conducted on NVIDIA A100 GPUs and results are reported in terms of word error rate (WER).

Additionally, we decode the ASR model's output with a 5-gram KenLM language model [29]. We train the Punjabi language model using the IndiCorp dataset [34]. The dataset contains 29.2 million Punjabi sentences (773 million tokens). We fine-tune α and β parameters and empirically set $\alpha = 0.7$ and $\beta = 4.0$. Here α represents the shallow fusion weight for the language model and β refers to the weight to adjust the score as per the length of the decoded sequence [35]. Further, for selecting the highly confident pseudo-labels, we adopt a normalized confidence score filtering.

5. Experimental Results and Discussion

5.1. Comparative analysis

We report the results in Table 2. To show the effectiveness of our 5-gram language model (LM), we demonstrate the results on the seed model with and without the language model. We find that our LM decoding further improves the seed model, consequently, we choose seed with LM as our baseline model. All of our experimental results (except seed without LM) are carried out with LM decoding. Raw PL uses all available pseudo-labels, whereas Best PL selects the best-performing pseudo-labels over multiple iterations of pseudo-labeling with different confidence thresholds. The results show that using pseudo-labels with a filter (Best PL) outperforms Raw PL on all datasets, indicating that selecting the best pseudo-label is important to improve the performance of the ASR system. We find that iteration 5 with confidence scores (CS$\geq$-1.5) produces the best results as shown in Table 2 (Best PL) and Figure 2. As compared to the baseline, we find our approach achieves relative improvements of 13.04%, 18.66%, 12.91%, and 14.94% on real speech Shrutilipi, Common Voice, Punjabi Speech, and 50Languages datasets, respectively. Our PL approach demonstrates the best results on the Common Voice Punjabi dataset with a relative improvement of 50.5% over previously reported results (23.07 % vs 11.42% WER) [6].

Further, on synthetic datasets (i.e., Google-synth and CMU-synth), our PL approach also achieves better results; nevertheless, the improvements are marginal. We attribute these minimal improvements to the quality of synthetic data. Following our previous study [6], we only incorporated these datasets to improve the quality of the overall ASR model.

5.2. Effectiveness of gradual filtration

Figure 2 illustrates the performance of the model with different CS thresholds and the amount of PL data selected with filtration over multiple PL iterations. Initially, we start with aggressive
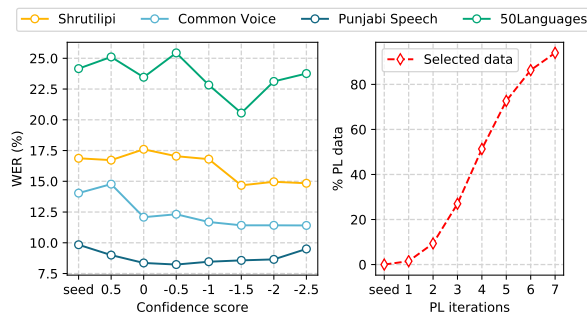

Figure 2: *Performance on real speech datasets against various confidence score thresholds (left) and % of selected data over the PL iterations (right).*

filtering with (CS$\geq$0.5), which selects only 1.5% of the total unlabeled Audiobooks dataset. With this threshold, although we select high-quality PL, however, it results in marginal improvements overall (2.8% of relative improvement). As we gradually relax the confidence score threshold, our model indeed continues to produce better results over each iteration of PL. With (CS$\geq$0), selects 9.37% of PL, which results in 9.5% of overall relative improvement compared to baseline. Overall, as Figure 2 shows, (CS$\geq$-1.5), which selects 72.66% of PL, produces the lowest WERs across most datasets. Most consistent improvements are seen on the Common Voice and Punjabi Speech datasets. Further, our experimental analysis reveals that PL generated with (CS$\geq$-2 or greater) selects more than 80% of the total PL data. However, the performance of the model gradually starts to diminish as compared to the Best PL.

6. Conclusions

The paper presents a self-training (pseudo-labeling) approach for automatic speech recognition for low-resource settings, specifically focusing on the Punjabi language. The proposed self-training approach generates highly accurate pseudo-labels for unlabeled Punjabi speech, resulting in a significant improvement in word error rate (WER) compared to a strong baseline. The experimental analysis demonstrates the effectiveness of our approach and highlights the potential to improve the performance of ASR systems for low-resource languages such as Punjabi.

7. Acknowledgement

8. References

[1] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in Neural Information Processing Systems*, vol. 33, pp. 12 449–12 460, 2020.

[2] A. Ragni, K. M. Knill, S. P. Rath, and M. J. Gales, "Data augmentation for low resource languages," in *Proc. Interspeech*, 2014, pp. 810–814.

[3] S. Zhou, S. Xu, and B. Xu, "Multilingual end-to-end speech recognition with a single transformer on low-resource languages," *arXiv preprint arXiv:1806.05059*, 2018.

[4] C. Yi, J. Wang, N. Cheng, S. Zhou, and B. Xu, "Applying wav2vec2. 0 to speech recognition in various low-resource languages," *arXiv preprint arXiv:2012.12121*, 2020.

[5] S. Singh, R. Wang, and F. Hou, "Improved meta learning for low resource speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 4798–4802.

[6] S. Singh, R. Wang, F. Hou, and Z. Ma, "Enhancing end-to-end automatic speech recognition for low-resource Punjabi language using synthesized datasets," *Available at SSRN 4181844*, 2022.

[7] J. Kahn, A. Lee, and A. Hannun, "Self-training for end-to-end speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7084–7088.

[8] D.-H. Lee *et al.*, "Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks," in *Workshop on challenges in representation learning, ICML*, vol. 3, no. 2, 2013, p. 896.

[9] D. Berthelot, N. Carlini, I. Goodfellow, N. Papernot, A. Oliver, and C. A. Raffel, "MixMatch: A holistic approach to semi-supervised learning," *Advances in neural information processing systems*, vol. 32, 2019.

[10] J. He, J. Gu, J. Shen, and M. Ranzato, "Revisiting self-training for neural sequence generation," in *International Conference on Learning Representations*, 2019.

[11] X. Li and B. Yang, "A pseudo label based dataless naive bayes algorithm for text classification with seed words," in *Proceedings of the 27th International Conference on Computational Linguistics*, 2018, pp. 1908–1917.

[12] D. S. Park, Y. Zhang, Y. Jia, W. Han, C.-C. Chiu, B. Li, Y. Wu, and Q. V. Le, "Improved Noisy Student Training for Automatic Speech Recognition," in *Proc. Interspeech*, 2020, pp. 2817–2821.

[13] Q. Xu, T. Likhomanenko, J. Kahn, A. Hannun, G. Synnaeve, and R. Collobert, "Iterative Pseudo-Labeling for Speech Recognition," in *Proc. Interspeech*, 2020, pp. 1006–1010.

[14] Y. Higuchi, N. Moritz, J. L. Roux, and T. Hori, "Momentum pseudo-labeling for semi-supervised speech recognition," in *Proc. Interspeech*, 2021, pp. 726–730.

[15] N. Moritz, T. Hori, and J. Le Roux, "Semi-supervised speech recognition via graph-based temporal classification," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6548–6552.

[16] T. Likhomanenko, Q. Xu, J. Kahn, G. Synnaeve, and R. Collobert, "slimIPL: Language-model-free iterative pseudo-labeling," in *Proc. Interspeech*, 2020, pp. 741–745.

[17] A. Conneau, A. Baevski, R. Collobert, A. Mohamed, and M. Auli, "Unsupervised cross-lingual representation learning for speech recognition," in *Proc. Interspeech*, 2021, pp. 2426–2430.

[18] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition," in *Proc. Interspeech*, 2019, pp. 2613–2617.

[19] Y. Higuchi, N. Moritz, J. Le Roux, and T. Hori, "Advancing momentum pseudo-labeling with conformer and initialization strategy," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 7672–7676.

[20] L. Lugosch, T. Likhomanenko, G. Synnaeve, and R. Collobert, "Pseudo-labeling for massively multilingual speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 7687–7691.

[21] 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 *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 3030–3034.

[22] Z. Jin, D. Zhong, X. Song, Z. Liu, N. Ye, and Q. Zeng, "Filter and evolve: progressive pseudo label refining for semi-supervised automatic speech recognition," *arXiv preprint arXiv:2210.16318*, 2022.

[23] Z.-Q. Zhang, Y. Song, M.-H. Wu, X. Fang, and L.-R. Dai, "XLST: Cross-lingual self-training to learn multilingual representation for low resource speech recognition," *arXiv preprint arXiv:2103.08207*, 2021.

[24] S. Khurana, N. Moritz, T. Hori, and J. Le Roux, "Unsupervised domain adaptation for speech recognition via uncertainty driven self-training," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6553–6557.

[25] M. Dua, R. Aggarwal, V. Kadyan, and S. Dua, "Punjabi automatic speech recognition using HTK," *International Journal of Computer Science Issues (IJCSI)*, vol. 9, no. 4, p. 359, 2012.

[26] J. Guglani and A. N. Mishra, "Continuous Punjabi speech recognition model based on Kaldi ASR toolkit," *International Journal of Speech Technology*, vol. 21, no. 2, pp. 211–216, 2018.

[27] V. Kadyan, A. Mantri, R. Aggarwal, and A. Singh, "A comparative study of deep neural network based Punjabi-ASR system," *International Journal of Speech Technology*, vol. 22, no. 1, pp. 111–119, 2019.

[28] Y. Kumar, N. Singh, M. Kumar, and A. Singh, "Autossr: an efficient approach for automatic spontaneous speech recognition model for the punjabi language," *Soft Computing*, vol. 25, pp. 1617–1630, 2021.

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

[30] K. S. Bhogale, A. Raman, T. Javed, S. Doddapaneni, A. Kunchukuttan, P. Kumar, and M. M. Khapra, "Effectiveness of mining audio and text pairs from public data for improving asr systems for low-resource languages," *arXiv preprint arXiv:2208.12666*, 2022.

[31] R. Ardila, M. Branson, K. Davis, M. Kohler, J. Meyer, M. Henretty, R. Morais, L. Saunders, F. Tyers, and G. Weber, "Common Voice: A massively-multilingual speech corpus," in *Proceedings of the 12th Language Resources and Evaluation Conference*, 2020, pp. 4218–4222.

[32] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks," in *International conference on Machine learning (ICML)*, 2006, pp. 369–376.

[33] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *International Conference on Learning Representations (ICLR)*, 2015.

[34] D. Kakwani, A. Kunchukuttan, S. Golla, N. Gokul, A. Bhattacharyya, M. M. Khapra, and P. Kumar, "IndicNLPSuite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for Indian languages," in *Findings of the Association for Computational Linguistics: EMNLP 2020*, 2020, pp. 4948–4961.

[35] A. Kannan, Y. Wu, P. Nguyen, T. N. Sainath, Z. Chen, and R. Prabhavalkar, "An analysis of incorporating an external language model into a sequence-to-sequence model," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 1–5828.