INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Domain Adaptive Self-supervised Training of Automatic Speech Recognition

Cong-Thanh Do[1], *Rama Doddipatla*[1], *Mohan Li*[1], *and Thomas Hain*[2]

[1]Cambridge Research Laboratory, Toshiba Europe Ltd., Cambridge, UK
[2]The University of Sheffield, Sheffield, UK

E-mails: {cong-thanh.do, rama.doddipatla, mohan.li}@toshiba.eu, t.hain@sheffield.ac.uk

Abstract

This paper explores domain adaptive self-supervised training of automatic speech recognition (ASR). Unlabeled data from the target domain can either be used in training the self-supervised pre-trained model or in the fine-tuning stage using semi-supervised approaches for the ASR task or both. Here we specifically focus on how semi-supervised approaches can enhance domain adaptation of pre-trained models built using self-supervised learning (SSL). For the purpose of this study, we use variants of English accents as the data from different domains. ASR experiments targeting single domain achieve relative word error rate (WER) reduction in the range 2.7-41.8% based on the extent of domain mismatch, while in the multiple-domain setting we achieve a relative WER reduction of 8% on average using semi-supervised fine-tuning on top of the model pre-trained with target domain using SSL.

Index Terms: Automatic speech recognition, self-supervised learning, domain adaptive training, accented speech.

1. Introduction

Self-supervised learning (SSL) is an unsupervised learning approach that utilizes information extracted from the input data itself as the labels to learn representations useful for downstream tasks [1–3]. In speech representations learning using SSL, a representation model is first trained using unlabeled audio data, also commonly referred to as pre-trained model. The learned representations from the pre-trained models are later used for training a model for the downstream task from scratch [4, 5] or the entire pre-trained model is fine-tuned to downstream task [6–8]. In self-supervised training of automatic speech recognition (ASR) systems, the fine-tuning often requires supervised data, i.e. paired speech and text data.

Unlabeled target domain data which is close to the domain of test data can be collected in several realistic scenarios on the usage of ASR systems. For instance, when users speak to a smart speaker, their voices can be gradually stored as anonymous target domain data. This unlabeled data can subsequently be used to improve the ASR model implemented in the smart speaker. When the ASR system is trained with SSL, the unlabeled target domain data can be efficiently used to improve the training of the ASR system. In [9], the authors explored the use of target domain data in self-supervised pre-training of ASR system and found that this helps improving significantly the ASR performance on test data of domains close to that of the unlabeled data which were added to the pre-training.

The present paper explores the use of unlabeled data for domain adaptation. With the recent advances in SSL, unlabeled target domain data are widely explored for adaptation in the form of retraining the pre-trained model [9,10]. However, semi-supervised approaches can also be applied for adaptation using unlabeled target domain data, which are not widely explored in the context of SSL frameworks. Semi-supervised approaches can be employed to generate pseudo-labels on the target domain data, which can be used in fine-tuning the self-supervised pre-trained model.

In this paper, we primarily focus on improving ASR performance to unseen accents using SSL in combination with pseudo-labeling for fine-tuning with semi-supervised data. To the best of our knowledge, this is the first time semi-supervised fine-tuning is explored in domain adaptive self-supervised training of accented speech recognition. Experiments are conducted using American accents as in-domain data, where models are trained on the Librispeech corpus [11]. Unseen accented data comes from L2-ARCTIC non-native English corpus [12], the British Isles corpus of British English accents [13], and the NPTEL corpus of English spoken with Indian accents [14]. The results show that the proposed approach is effective in improving the ASR performance in both single-domain and multiple-domain scenarios.

The paper is organized as follows. In section 2, we present previous works which are related to the work in our paper. Domain adaptive self-supervised training of ASR is presented in section 3. Experimental settings and results are presented in section 4. Finally, section 5 concludes the paper.

2. Related works

Using unlabeled target domain data, similar or closer to the domain of the test data, in self-supervised pre-training of ASR system was explored in [9, 10] and was found to significantly improve the ASR performance on test data. It was also shown that pre-training on multiple domains increases the robustness to completely unseen domains [9]. In [15], the authors investigated the use of the SSL ASR model to generate pseudo-labels for supervised training of a Transformer-based sequence-to-sequence ASR model, which was shown to surpass the initial SSL ASR model's performance on Librispeech.

The use of pseudo-labels in ASR for semi-supervised training has been widely explored in literature [16–20]. In semi-supervised training, a base or teacher model is used to generate pseudo-labels whereas the student model is tuned on the data that consists of ground truth labels and pseudo-labels. The pseudo-labeling process can be further improved by filtering out pseudo-labels with low-confidence [21] or performing iterative refinements [22–24].

In this paper, we explore domain adaptive self-supervised training of ASR. In contrast to the approaches in [9, 15, 18], we explore the use of target domain data either in pre-training or semi-supervised fine-tuning, or both. In [25], semi-supervised

10.21437/Interspeech.2023-1091

fine-tuning was performed on top of a model trained on supervised data whereas in this paper, we explore the use of unseen target domain data both in pre-training and fine-tuning.

3. Domain adaptive self-supervised training of ASR

In self-supervised training of ASR systems, we assume that large unlabeled training data are available for pre-training, and certain amount of labeled data are available for fine-tuning of the pre-trained model. When unlabeled target domain data are available, we can normally use these data in the pre-training [9]. These data can also be automatically transcribed and combined with the existing labeled data to be used in the semi-supervised fine-tuning of the pre-trained model, trained without or with the target domain data. Fig. 1 shows how the unlabeled target domain data can be used for domain adaptive self-supervised training of an ASR system.

In this paper, Wav2vec 2.0 SSL models [6] are used in the experiments. The feature encoder in Wav2vec 2.0 model consists of several blocks in which a temporal convolution is followed by layer normalization and GELU (Gaussian error linear unit) activation function [26]. The raw waveform is first normalized to zero mean and unit variance before being processed by the feature encoder. The number of time steps which are input to the Transformer context network is determined by the total stride of the encoder [6].

The output of the feature encoder is discretized to a finite set of speech representations via product quantization [27], and is subsequently processed by a context network which has Transformer architecture [28]. Relative positional encoding is implemented via a convolutional layer [29]. The quantization module uses a Gumbel softmax to choose entries from codebooks. During pre-training, speech representations are learned by solving a contrastive task in which a contrastive loss is used. The feature encoder consists of 6 convolutional layers and the context network consists of 12 Transformer layers.

During fine-tuning, the pre-trained model is fine-tuned for ASR task by adding a randomly initialized linear projection on top of the context network into C classes representing the vocabulary of the task. In the present work, we use English characters as output units of the model. There are thus $C = 29$ classes, plus a word boundary token. Connectionist temporal classification loss [30] is minimized to optimized the model parameters during fine-tuning.

During decoding, a world-level Transformer-based language model (LM) is used as an external LM in beam-search decoding [19]. This Transformer-based LM is trained on the manual transcriptions of Librispeech training data. In the present paper, the Wav2vec 2.0 models as well as their pre-training, fine-tuning, and decoding follow the same settings used for the BASE model in [6].

4. Experiments

We explore domain adaptive self-supervised training in the context of accented speech recognition [31, 32]. Accent is considered as one of the most important factors which yield mismatches between training and test in ASR. Addressing accents via model adaptation methods is one of the efficient approaches in accented ASR [33]. In this work, we use data from Librispeech which is of American English accents in the training of the base model, and three other speech corpora of English accents as target domain data. These corpora are L2-ARCTIC corpus of non-native English [12], British Isles corpus of British English accents [13], and NPTEL corpus of English spoken with Indian accents [14]. The speech in Librispeech, L2-ARCTIC, and British Isles are read, while that in the NPTEL corpus is spontaneous lecture speech. The data in each corpus is considered to be from one domain. In this respect, single-domain and multiple-domain data are used in the experiments where data from either one or three corpora are used together with Librispeech data.

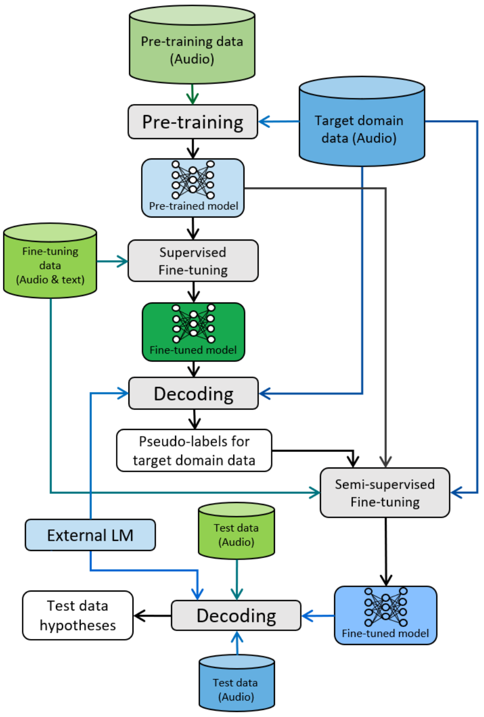

Figure 1: *Domain adaptive self-supervised training of ASR systems. Unlabeled target domain data can be used in pre-training, semi-supervised fine-tuning, or in both stages.*

4.1. Single domain

The ASR models used in this section are named according to the naming convention in Table 1. Librispeech's 960 hours unlabeled training data (the train_960 set) and 100 hours supervised training data (the train-clean-100 set) are used in the pre-training and fine-tuning, respectively, of all the models. Henceforth, the data from Librispeech, L2-ARCTIC, British Isles, and NPTEL are indicated as from domains D0, D1, D2, and D3, respectively. The WERs measured on the target domain data when they are decoded with the D0 (Base) and D0-AUG-Di, $i = 1, .., 3$ models are very similar to those measured on the test data when they are decoded with these models.

The information in Table 1 indicates whether the pre-training and fine-tuning of the models use target domain data in addition to the Librispeech training data. Information about which model is used to decode the target domain data to obtain pseudo-labels is also mentioned. In Table 1, AUG means the target domain data are *augmented* in pre-training, fine-tuning, or in both stages; FT means fine-tuned, and Oracle means the ground truth labels of the target domain data are used during fine-tuning. The results obtained with the D0-FT-Oracle-Di, $i = 1, .., 3$ model are thus used as lower bounds on the WERs.

Table 1: *Names and information on the creation of models used in the experiments with single-domain data Di, $i = 1, .., 3$.*

Model name	Use of domain data in:		Model used to decode domain data
	Pre-train	Fine-tune	
D0 (Base)	No	No	N/A
D0-FT-Di	No	Yes	D0 (Base)
D0-AUG-Di	Yes	No	N/A
D0-FT-AUG-Di	No	Yes	D0-AUG-Di
D0-AUG-Di-FT	Yes	Yes	D0-AUG-Di
D0-FT-Oracle-Di	No	Yes	N/A (Oracle labels)

Table 2: *WERs on Librispeech and L2-ARCTIC (D1) test sets. The average (Avg.) WERs are computed using weights proportional with the numbers of utterances in the test sets.*

Test set Model	Test-clean	Test-other	Test-D1	Avg.
D0 (Base)	2.6	6.6	16.1	8.7
D0-FT-D1	2.8	7.6	11.8	7.6
D0-AUG-D1	2.7	6.9	11.2	7.1
D0-FT-AUG-D1	2.8	7.2	8.5	6.3
D0-AUG-D1-FT	2.8	7.4	**7.9**	**6.2**
D0-FT-Oracle-D1	2.8	7.4	1.6	4.0

4.1.1. L2-ARCTIC corpus (domain D1)

The L2-ARCTIC corpus is a speech corpus of non-native English [12] which contains 26,867 utterances from 24 non-native English speakers with equally distributed number of speakers per accent. The total duration of the corpus is 27.1 hours, with an average of 67.7 minutes of speech per speaker. On average, each utterance is 3.6 seconds in duration. The utterances in L2-ARCTIC are spoken in 6 non-native accents: Arabic, Chinese, Hindi, Korean, Spanish, and Vietnamese.

We separate 3,000 utterances from the whole dataset to create a test set called Test-D1. The remaining 23,867 utterances are used as target domain data which can be incorporated in the pre-training, semi-supervised fine-tuning, or in both stages. The utterances in the test and target domain data sets are not overlapped. Table 2 shows the WERs of the ASR models on Librispeech and L2-ARCTIC test sets. The WERs of the Base (D0) model on `Test-clean` and `Test-other` sets of Librispeech are comparable to state-of-the-art WERs on the same test sets reported in [6].

Compared to the D0 (Base) model, the semi-supervised fine-tuning D0-FT-D1 which uses pseudo-labels generated by the D0 model achieves 12.6% relative reduction of the average WER, and 26.7% relative WER reduction on the Test-D1 set of which target domain data are incorporated in the semi-supervised fine-tuning. Incorporating target domain data only in pre-training (D0-AUG-D1) yields 18.4% relative reduction of the average WER while the semi-supervised fine-tuning D0-AUG-D1-FT which uses the pseudo-labels generated by using the D0-AUG-D1 model reduces further the WER. This model yields 12.5% relative reduction of the average WER and 41.8% relative WER reduction on the Test-D1 set compared to the D0-AUG-D1 model. Using oracle labels of the domain data in fine-tuning (D0-FT-Oracle-D1) yields the lowest average WER.

4.1.2. British Isles corpus (domain D2)

The British Isles speech corpus [13] includes audio of English sentences recorded by volunteers speaking with different accents of the British Isles, namely Ireland, Scotland, Wales, the Midlands, Northern, and Southern of England. The corpus consists of 17,877 utterances spoken by 120 speakers of which 49 are female and 71 are male. We separate 1,800 utterances to create a test set (Test-D2) and use the remaining 16,077 utterances as target domain data for pre-training and fine-tuning. The total duration of the corpus is around 31 hours. The selection of utterances for training and test sets is randomly done, for both L2-ARCTIC and British Isles corpora.

Table 3: *WERs on Librispeech and British Isles (D2) test sets.*

Test set Model	Test-clean	Test-other	Test-D2	Avg.
D0 (Base)	2.6	6.6	14.4	7.1
D0-FT-D2	2.8	7.2	12.3	6.9
D0-AUG-D2	2.6	6.7	11.4	6.4
D0-FT-AUG-D2	2.8	7.2	10.3	6.4
D0-AUG-D2-FT	2.7	7.1	**10.1**	**6.2**
D0-FT-Oracle-D2	2.7	6.8	7.3	5.5

Table 4: *WERs on Librispeech test sets and NPTEL evaluation set (Eval-D3).*

Test set Model	Test-clean	Test-other	Eval-D3	Avg.
D0 (Base)	2.6	6.6	35.7	14.1
D0-FT-D3	2.9	8.0	35.9	14.8
D0-AUG-D3	2.7	7.1	18.3	**9.0**
D0-FT-AUG-D3	2.8	7.9	18.2	9.3
D0-AUG-D3-FT	2.9	7.8	**17.8**	9.3
D0-FT-Oracle-D3	2.8	8.2	9.1	6.7

In Table 3, the effectiveness of domain adaptive self-supervised training is also observed when either the D0 (Base) or the D0-AUG-D2 model is used to generate pseudo-labels. The D0-AUG-D2-FT yields 12.5% relative reduction of the average WER compared to the D0 model. On the Test-D2 set, the semi-supervised fine-tuning model D0-AUG-D2-FT yields 11.4% relative WER reduction compared to the D0-AUG-D2 model, which is 20.8% lower than the WER of the D0 model.

4.1.3. NPTEL corpus (domain D3)

NPTEL (National Programme on Technology Enhanced Learning) Indian English dataset is a collection of online lectures which are freely distributed [14]. This is considered to be the largest online repository of courses in engineering, basic sciences, and selected humanities and social sciences subjects. Total duration of the data from the corpus is around 15,700 hours.

In this paper, 200 hours of NPTEL data which is similar to those used for training ASR systems in the English ASR challenge [34] are used as target domain data. The 5-hour NPTEL evaluation set, which was also used as evaluation set in the English ASR challenge [34], is used as evaluation set (Eval-D3). The speaking styles in both the training and evaluation sets are spontaneous lecture speech.

Experimental results with the NPTEL data are shown in Table 4. It can be seen that including target domain data into the pre-training (D0-AUG-D3) reduces significantly the WERs on the NPEL evaluation set (Eval-D3). The relative WER reduction on the Eval-D3 set is 48.7% compared to the D0 model. The semi-supervised fine-tuning D0-FT-D3 of the D0 model using the pseudo-labels generated by the D0 model does not

Table 5: *WERs of models using single-domain and multiple-domain data only in pre-training.*

Test set Model	Test-clean (D0)	Test-other (D0)	Test-D1	Test-D2	Test-D3	Average
D0	**2.6**	**6.6**	16.1	14.4	35.7	14.6
D0-AUG-D1	2.7	6.9	**11.2**	15.0	34.3	13.4
D0-AUG-D2	2.6	6.7	16.8	**11.4**	36.4	14.5
D0-AUG-D3	2.7	7.1	15.5	15.4	18.3	11.5
D0-AUG-D1,2,3	2.8	7.0	11.8	11.7	**18.2**	**10.0**

Table 6: *WERs of the models obtained by applying semi-supervised fine-tuning on top of the pre-trained models used in Table 5.*

Test set Model	Test-clean (D0)	Test-other (D0)	Test-D1	Test-D2	Test-D3	Average
D0-FT-D1,2,3	2.9	8.4	11.8	13.0	35.9	13.9
D0-AUG-D1-FT	**2.8**	7.4	**7.9**	15.9	35.3	13.0
D0-AUG-D2-FT	2.7	**7.1**	17.3	**10.1**	37.5	14.8
D0-AUG-D3-FT	2.9	7.8	17.4	16.0	17.8	12.1
D0-AUG-D1,2,3-FT	2.9	7.9	8.6	10.6	**17.7**	**9.2**
D0-FT-Oracle-D1,2,3	2.8	7.7	2.9	7.1	9.7	5.9

Table 7: *Data and models used to create the models in Table 6.*

Model name	Data used in:		Model used to decode data
	Pre-train	Fine-tune	
D0-FT-D1,2,3	D0	D0,1,2,3	D0
D0-AUG-D1-FT	D0,1	D0,1	D0-AUG-D1
D0-AUG-D2-FT	D0,2	D0,2	D0-AUG-D2
D0-AUG-D3-FT	D0,3	D0,3	D0-AUG-D3
D0-AUG-D1,2,3-FT	D0,1,2,3	D0,1,2,3	D0-AUG-D1,2,3
D0-FT-Oracle-D1,2,3	D0	D0,1,2,3	Oracle labels

reduce the WER, possibly because the WER of the D0 model on the target domain data is relatively high, around 35%. The semi-supervised fine-tuning D0-AUG-D3-FT of the D0-AUG-D3 model using pseudo-labels generated by the D0-AUG-D3 model helps to reduce the WER on the Eval-D3 set from 18.3% to 17.8%, i.e. 2.7% relative reduction.

4.2. Multiple domains

Using target domain data from one domain in the pre-training and semi-supervised fine-tuning could be effective for recognizing test data from that domain, but could be not optimal for other domains. Therefore, we examine the use of multiple domains data by including L2-ARCTIC, British Isles, and NPTEL data in the pre-training and semi-supervised fine-tuning to check if the domain adaptive self-supervised training using multiple-domain data is still effective. We perform only experiments in which target domain data are used in both pre-training and semi-supervised fine-tuning because this setting yields the best performance, as seen in section 4.1.

Tables 5 and 6 show the WERs, measured on all the test sets, of the models using single-domain and multiple-domain data. All the models in Table 5 are fine-tuned only with 100 hours of labeled data from Librispeech while those in Table 6, except the D0-FT-Oracle-D1,2,3 model, are fine-tuned with semi-supervised data which include 100 hours of labeled data from Librispeech and the corresponding unlabeled target domain data. In Table 5, D0-AUG-D1,2,3 means multiple-domain data are incorporated in the pre-training together with the base domain D0. It can be seen that using multiple-domain data helps to train a better model (D0-AUG-D1,2,3) with an average WER which is 15% relatively lower than that of the best model using single-domain data (D0-AUG-D3).

In Table 6, semi-supervised fine-tunings are applied on top of pre-trained models which incorporate either single-domain or multiple-domain data in pre-training. Table 7 summarizes the data and models used to create the models in Table 6. Applying semi-supervised fine-tuning using multiple-domain data (D0-AUG-D1,2,3-FT) on the pre-trained model which incorporates multiple-domain data in pre-training yields 8% relative reduction of the average WER compared to D0-AUG-D1,2,3 (the best model in Table 5). Incorporating single-domain data in both pre-training and semi-supervised fine-tuning does not necessarily result in a reduction of the average WER, but using multiple-domain data in this setting helps to secure this. Fine-tuning the D0 model using oracle labels of multiple-domain data, D0-FT-Oracle-D1,2,3, yields the lowest average WER, indicating that there is still scope for improvement.

5. Conclusion

The paper presented an approach to perform domain adaptation using unlabeled data by exploiting self-supervised pre-training in combination with semi-supervised fine-tuning. Using data from various accents of English as domain data, experiments were conducted to show how the proposed approach can be used to adapt the model to benefit both single- and multiple-domain scenarios. A systematic evaluation was conducted where target domain data were used only in self-supervised pre-training or only while performing fine-tuning using the semi-supervised approach or a combination of both to leverage unseen target domain data to improve ASR performance. With semi-supervised fine-tuning, one can observe that target single domain scenarios benefit a lot when the extent of domain mismatch is larger, while the multiple-domain scenarios show that the models are robust to variations in accents and can further improve the average performance of the model on multiple accents. These results suggest that semi-supervised fine-tuning can further complement SSL while using unlabeled target domain data to improve ASR performance. While accented speech recognition is studied in this paper, the approach can readily be extended to other domains, such as noise robustness for ASR.

6. References

[1] A. van den Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," in *arXiv preprint arXiv: 1807.03748*, 2018.

[2] A. Mohamed, H.-Y. Lee, L. Borgholt, J. D. Havtorn, J. Edin, C. Igel, K. Kirchhoff, S.-W. Li, K. Livescu, L. Maaløe, T. N. Sainath, and S. Watanabe, "Self-supervised speech representation learning: A review," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1179–1210, 2022.

[3] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao, J. Wu, L. Zhou, S. Ren, Y. Qian, Y. Qian, J. Wu, M. Zeng, X. Yu, and F. Wei, "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.

[4] X. Chang, T. Maekaku, P. Guo, J. Shi, Y.-J. Lu, A. S. Subramanian, T. Wang, S.-W. Yang, Y. Tsao, H.-Y. Lee, and S. Watanabe, "An exploration of self-supervised pretrained representations for end-to-end speech recognition," in *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2021, pp. 228–235.

[5] P. Kumar, V. N. Sukhadia, and S. Umesh, "Investigation of robustness of Hubert features from different layers to domain, accent and language variations," in *Proc. 2022 IEEE ICASSP*, Singapore, May 2022, pp. 6887–6891.

[6] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "Wav2vec 2.0: a framework for self-supervised learning of speech representations," in *Proc. of Advances in Neural Information Processing Systems (NeurIPS 2020)*, December 2020.

[7] 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.

[8] Y. Meng, Y.-H. Chou, A. T. Liu, and H.-Y. Lee, "Don't speak too fast; the impact of data bias on self-supervised speech models," in *Proc. 2022 IEEE ICASSP*, Singapore, May 2022, pp. 3258–3262.

[9] W.-N. Hsu, A. Sriram, A. Baevski, T. Likhomanenko, Q. Xu, V. Pratap, J. Kahn, A. Lee, R. Collobert, G. Synnaeve, and M. Auli, "Robust wav2vec 2.0: analyzing domain shift in self-supervised pre-training," in *Proc. Interspeech 2021*, Brno, Czechia, August-September 2021, pp. 721–725.

[10] A. Misra, D. Hwang, Z. Huo, S. Garg, N. Siddhartha, A. Narayanan, and K. C. Sim, "A comparison of supervised and unsupervised pre-training of end-to-end models," in *Proc. Interspeech 2021*, Brno, Czechia, August-September 2021, pp. 731–735.

[11] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *Proc. 2015 IEEE ICASSP*, Brisbane, Australia, April 2015, pp. 5206–5210.

[12] G. Zhao, S. Sonsaat, A. Silpachai, I. Lucic, E. Chukharev-Hudilainen, J. Levis, and R. Gutierrez-Osuna, "L2-ARCTIC: a non-native English speech corpus," in *Proc. Interspeech 2018*, Hyderabad, India, September 2018, pp. 2783–2787.

[13] I. Demirsahin, O. Kjartansson, A. Gutkin, and C. Rivera, "Open-source multi-speaker corpora of the English accents in the British Isles," in *Proc. of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, Marseille, France, May 2020, pp. 6532–6541.

[14] "NPTEL2020 - Indian English Speech Dataset," https://github.com/AI4Bharat/NPTEL2020-Indian-English-Speech-Dataset, accessed: 2023-02-15.

[15] Q. Xu, A. Baevski, T. Likhomanenko, P. Tomasello, A. Conneau, R. Collobert, G. Synnaeve, and M. Auli, "Self-training and pre-training are complementary for speech recognition," in *Proc. 2021 IEEE ICASSP*, Toronto, Canada, June 2021, pp. 3025–3029.

[16] D.-H. Lee, "Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks," in *Proc. ICML 2013 Workshop: Challenge in Representation Learning (WREPL)*, Atlanta, USA, June 2013.

[17] Y. Chen, W. Wang, and C. Wang, "Semi-supervised ASR by end-to-end self-training," in *Proc. Interspeech 2020*, Shanghai, China, October 2020, pp. 2787–2791.

[18] J. Kahn, A. Lee, and A. Hannun, "Self-training for end-to-end speech recognition," in *Proc. 2020 IEEE ICASSP*, Barcelona, Spain, May 2020, pp. 7084–7088.

[19] G. Synnaeve, Q. Xu, J. Kahn, T. Likhomanenko, E. Grave, V. Pratap, A. Sriram, V. Liptchinsky, and R. Collobert, "End-to-end ASR: from supervised to semi-supervised learning with modern architectures," in *Proc. ICML 2020 Workshop: Self-supervision in Audio and Speech (SAS)*, July 2020.

[20] C.-T. Do, M. Li, and R. Doddipatla, "Multiple-hypothesis RNN-T loss for unsupervised fine-tuning and self-training of neural transducer," in *Proc. Interspeech 2022*, Incheon, Korea, September 2022, pp. 4446–4450.

[21] P. Bell, J. Fainberg, O. Klejch, J. Li, S. Renals, and P. Swietojanski, "Adaptation algorithms for neural network-based speech recognition: an overview," *IEEE Open Journal of Signal Processing*, vol. 2, pp. 33–66, 2021.

[22] Q. Xu, T. Likhomanenko, J. Kahn, A. Hannun, G. Synnaeve, and R. Collobert, "Iterative pseudo-labeling for speech recognition," in *Proc. Interspeech 2020*, Shanghai, China, October 2020, pp. 1006–1010.

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

[24] T. Li, Q. Meng, and Y. Sun, "Improved noisy iterative pseudo-labeling for semi-supervised speech recognition," in *Proc. 2022 IEEE Spoken Language Technology Workshop*, Doha, Qatar, January 2023, pp. 167–173.

[25] L. Lugosch, T. Likhomanenko, G. Synnaeve, and R. Collobert, "Pseudo-labeling for massively multilingual speech recognition," in *Proc. 2022 IEEE ICASSP*, Singapore, May 2022, pp. 7687–7691.

[26] D. Hendrycks and K. Gimpel, "Gaussian error linear units (GELUs)," in *arXiv preprint arXiv: 1606.08415*, 2016.

[27] H. Jegou, M. Douze, and C. Schmid, "Product quantization for neareast neighbor search," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 33, no. 1, pp. 117–128, 2011.

[28] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. of 31st Conference on Neural Information Processing Systems (NIPS 2017)*, Long Beach, CA, USA, December 2017.

[29] A. Mohamed, D. Okhonko, and L. Zettlemoyer, "Transformers with convolutional context for ASR," in *arXiv preprint arXiv: 1904.11660*, 2019.

[30] A. Graves, S. Fernandez, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *Proc. of the 23rd International Conference on Machine Learning*, Pittsburgh, USA, June 2006, pp. 369–376.

[31] A. Hinsvark *et al.*, "Accented speech recognition: a survey," in *arXiv preprint arXiv: 2104.10747*, 2021.

[32] T. Viglino, P. Motlicek, and M. Cernak, "End-to-end accented speech recognition," in *Proc. Interspeech 2019*, Graz, Austria, September 2019, pp. 2140–2144.

[33] M. A. Tugtekin, E. Vincent, and D. Jouvet, "Achieving multi-accent ASR via unsupervised acoustic model adaptation," in *Proc. Interspeech 2020*, Oct. 2020, pp. 1286–1290.

[34] "English ASR Challenge," https://sites.google.com/view/englishasrchallenge/home, accessed: 2023-02-22.