STABLE DISTILLATION: REGULARIZING CONTINUED PRE-TRAINING FOR LOW-RESOURCE AUTOMATIC SPEECH RECOGNITION

Ashish Seth[1*]   *Sreyan Ghosh*[2*]   *S. Umesh*[1]   *Dinesh Manocha*[2]

[1]IIT Madras, India, [2]University of Maryland, College Park, USA

ABSTRACT

Continued self-supervised (SSL) pre-training for adapting existing SSL models to the target domain has shown to be extremely effective for low-resource Automatic Speech Recognition (ASR). This paper proposes Stable Distillation, a simple and novel approach for SSL-based continued pre-training that boosts ASR performance in the target domain where both labeled and unlabeled data are limited. Stable Distillation employs self-distillation as regularization for continued pre-training, alleviating the over-fitting issue, a common problem continued pre-training faces when the source and target domains differ. Specifically, first, we perform *vanilla* continued pre-training on an initial SSL pre-trained model on the target domain ASR dataset and call it the teacher. Next, we take the same initial pre-trained model as a student to perform continued pre-training while enforcing its hidden representations to be close to that of the teacher (via MSE loss). This student is then used for downstream ASR fine-tuning on the target dataset. In practice, Stable Distillation outperforms all our baselines by 0.8 - 7 WER when evaluated in various experimental settings[1].

Index Terms— speech recognition, self-supervised learning, self-distillation

1. INTRODUCTION

Self-supervised learning (SSL) for learning effective speech representations without expensive labeled data has proven to be extremely successful for a variety of downstream Spoken Language Processing (SLP) tasks [1]. The general formulation of the training pipeline is to first perform large-scale upstream pre-training of a transformer model by solving a pretext task using unlabeled speech data and then fine-tune them on labeled datasets for diverse downstream tasks. The pretext task to be solved can vary, and researchers have continuously found new state-of-the-art methods that push performance on low-resource SLP tasks [2, 3] Despite its huge success in low-resource ASR, prior research has shown that SSL models suffer from various drawbacks: (1) Speech representation models learned with SSL often don't generalize well to
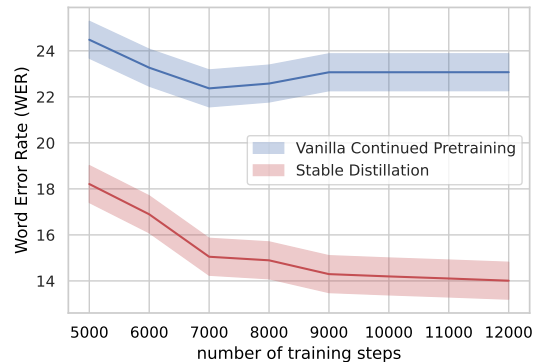
---

[1]https://github.com/cs20s030/stable_distillation

Fig. 1: SwitchBoard test-set WER evaluated at various steps of continued SSL pre-training of a model already pre-trained with LibriSpeech. While vanilla CP starts over-fitting with increasing steps, Stable Distillation avoids it by regularization.

domain shifts between pre-training and fine-tuning data [4], and (2) directly fine-tuning SSL models on downstream data often leads to catastrophic forgetting if there is a significant discrepancy between the pre-training and fine-tuning domains [5, 6]. To overcome this, researchers have proposed *continued SSL pre-training* (CP) on the downstream data to be an effective and complementary solution to bridge upstream pre-training and downstream fine-tuning domains [7], especially when the target domain dataset is low-resource. However, CP of an SSL model often leads to over-fitting on the train set [8]. Fig. 1 shows how, with increasing steps of continued pre-training on a target ASR dataset, the test Word Error Rate (WER) keeps increasing. Over-fitting patterns in the target domain also lead the model to forget useful general-purpose knowledge learned in the past, leading to sub-optimal performance. While it is well known in the deep learning community that training on low-resource data can almost always lead a model to overfit on the train set and poorly generalize, Purushwalkam *et al.* [8] discuss in detail why CP leads to over-fitting and conclude that the key reason is the violation of the IID assumption of optimization algorithms [9].

Main Contributions. In this paper, we present Stable Distillation, a simple and novel continued pre-training strategy to improve ASR performance on the target domain. Our key goal is to learn representations that, while adapting to the tar-
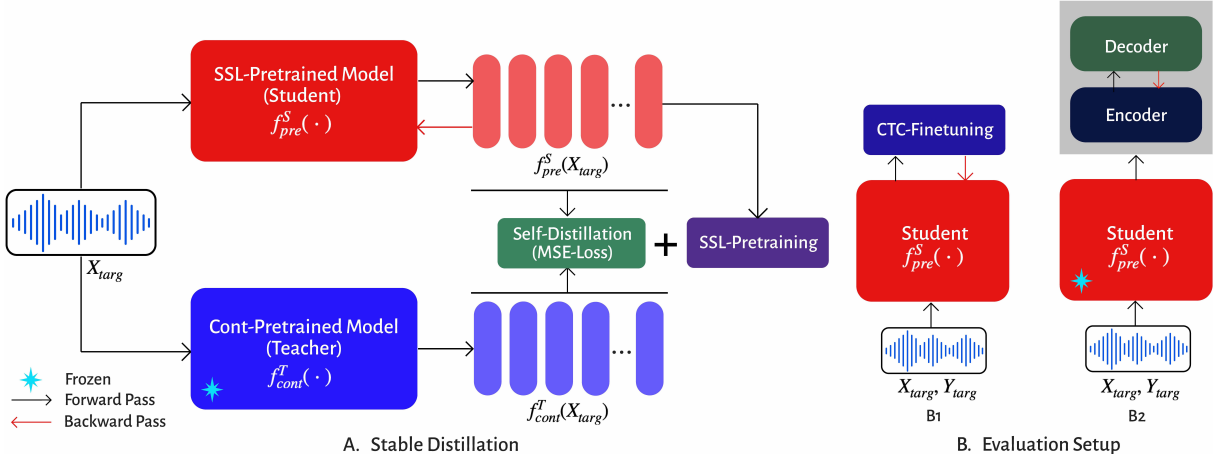
Fig. 2: Illustration of Stable Distillation (SD). SD performs self-distillation with a teacher model while pre-training a student on the target ASR dataset. Both the student and the teacher models start from identical SSL pre-trained models, pre-trained on a large-scale OOD dataset. The teacher is additionally pre-trained on the target dataset before SD. For evaluation, the student model is either fine-tuned end-to-end using CTC (B_1) or used as a feature extractor for fine-tuning an Encoder-Decoder model (B_2).

get domain, have not over-fitted the training dataset and are not forgetful of knowledge acquired during the initial pre-training. Stable Distillation, similar to other model distillation methods, employs two individual models, a teacher $f_{pre}^T(.)$ and a student $f_{pre}^S(.)$, both of which, in our case, are identical and pre-trained with SSL on a large-scale out-of-domain dataset. We first train the teacher with vanilla continued pre-training on the target downstream ASR data and solve the same pre-text task as the original SSL model. Next, we use the teacher to regularize continued pre-training on the student through self-distillation. Specifically, we attempt to match the representations of the student and the teacher by optimizing the MSE loss between both representations. Intuitively, the additional self-distillation regularizes the distance between the initial weight before continued pre-training and the final weight after fine-tuning. Finally, we either fine-tune the student for ASR or use it as a feature extractor. We conduct extensive experiments on several downstream datasets where the domain differs from the actual pre-training data domain and show that Stable Distillation can achieve absolute WER improvements in the range of 0.8 - 7 WER over our baselines that went through no or vanilla continued pre-training. Additionally, we show an implicit benefit of Stable Distillation where our method can be used for better adapting a cross-lingual model for mono-lingual ASR fine-tuning.

2. RELATED WORK

SSL in Speech. The last couple of years have seen a massive rise in SSL research with models constantly outperforming SOTA on benchmark datasets [10, 1]. The most common pre-text tasks include instance discrimination via contrastive learning [2], clustering [3], or reconstruction [11].

However, despite its success, SSL models are not robust to domain shifts between pre-training and fine-tuning datasets [4]. This calls for better and more effective ways to adapt a pre-trained model to a target domain for downstream fine-tuning. Continued pre-training on the downstream dataset has shown to be an effective solution for target domain adaptation. However, it comes at a risk of over-fitting [8]. Overall, through widely employed, continued pre-training of SSL models is a relatively under-explored area of speech representation learning, and we argue that more attention is required by the research community to solve problems related to it.

Domain Adaptation. Supervised domain adaption for ASR is a well-studied problem in literature [12]. On the contrary, unsupervised SSL model domain adaptation to the target task dataset domain is a relatively understudied problem in literature. Fan *et al.* [13] were the first to explore this task and proposed DRAFT, which uses adapters in pre-trained SSL models for effective continued pre-training. Though adapters help effectively mitigate the problem of overfitting and catastrophic forgetting [14], the need and complexity of adding extra parameters to the model impede their widespread usage. Lodagla *et al.* [15] propose a pruning-assisted domain adaptation of SSL representations. However, their framework is complex, requires multiple pre-training and fine-tuning stages, and achieves minimal improvement in performance.

3. METHODOLOGY

Problem Formulation. Fig. 2 illustrates our proposed Stable Distillation. Our problem statement is simple. Given an upstream pre-trained model $f_{pre}(.)$, pre-trained on any unlabeled dataset $\mathcal{D}_p = (\mathbf{X}_{pre})$ using SSL, we would like to employ $f_{pre}(.)$ to learn ASR on a downstream labeled dataset

$\mathcal{D}_f = (\mathbf{X}_{targ}, \mathbf{Y}_{targ})$. For this, we either fine-tune the upstream SSL model using CTC or use it as a feature extractor to fine-tune a conformer-based Encoder-Decoder model to finally obtain an ASR model $\mathcal{F}$. As discussed earlier, though continued pre-training of $\mathcal{D}_f$ on $f_{pre}(.)$ can lead to optimal performance due to the several advantages it offers, like domain adaption to the target domain, but it also suffers from various drawbacks like over-fitting. Thus, we seek an effective solution for CP via our proposed Stable Distillation.

3.1. Stable Distillation

(1) Continue Pre-training the Teacher. The first step towards performing Stable Distillation is to teach the teacher. We initialize our teacher by making a copy of our pre-trained model $f_{pre}(.)$ and denote it as $f_{pre}^T(.)$. $f_{pre}^T(.)$ is then continually pre-trained on $\mathcal{D}_p$, by solving the same pre-text task that $f_{pre}(.)$ was pre-trained on, to finally obtain $f_{cont}^T(.)$

(2) Self-Distillation as Regularization. As a second step, we initialize our student by making a copy of our pre-trained model $f_{pre}(.)$, and denote it as $f_{pre}^S(.)$. Next, parallel to performing continued pre-training on the student, we perform self-distillation between $f_{pre}^S(.)$ and $f_{cont}^T(.)$ by matching their final layer representations of downstream dataset $\mathcal{D}_f$ to finally obtain $f_{dist}^S(.)$. Thus, for stable distillation, we jointly optimize the MSE between student and teacher representations and the student's own pre-text SSL task loss as:

$$\mathcal{L}_{sd} = \|f_{pre}^S(\mathbf{X}_{targ}) - f_{cont}^T(\mathbf{X}_{targ}))\|_2 \\ + \alpha\mathcal{L}_{pretext}(f_{pre}^S(\mathbf{X}_{targ})) \quad (1)$$

where α is a tunable hyper-parameter used to weigh the pre-text loss $\mathcal{L}_{pretext}$. The pre-text task is similar to the original.

(3) Final downstream fine-tuning. Finally, we take the trained student model $f_{dist}^S(.)$ and fine-tune it on the target dataset $\mathcal{D}_f$ using either: (1) End-to-End CTC Fine-tuning: In a more conventional setup, we fine-tune all the weights of $f_{dist}^S(.)$ by adding a linear CTC head and optimizing the CTC loss. (2) We fine-tune a conformer-based encoder-decoder model employing $f_{dist}^S(.)$ as an upstream feature extractor by jointly optimizing CTC and attention-based auto-regressive loss [16]. For the latter, all weights of $f_{dist}^S(.)$ are frozen.

An intuition into why Stable Distillation works. Nagarjan *et al.* [17] show that the model capacity of deep networks is restricted through implicit regularization of the l_2 distance from the initialization. Thus, the distance can affect the generalization bounds of these networks. Mobahi *et al.* [18] perform several experiments to study the regularization effect of self-distillation. Thus, as a derivation of these findings, we use self-distillation to improve generalization via a regularization effect on the l_2 distance. Fig. 3 illustrates how SD minimizes the l_2 distance between fine-tuned and CP models.

Table 1: Detailed Statistics of datasets used in our experiments. **Type** refers to Conversational or Read speech.

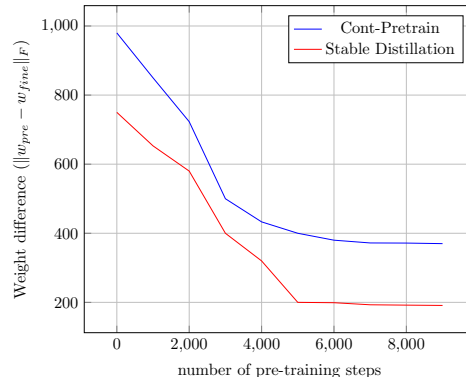| Dataset	Language	Domain	Type	Duration (train, dev, test)
MSR	Gujarati	General	Conv.	40hr, 5hr, 5hr
MSR	Tamil	General	Conv.	40hr, 5hr, 5hr
MSR	Telugu	General	Conv.	40hr, 5hr, 5hr
Gramvani (GV)	Hindi	Call Cent.	Conv.	100hr, 5hr, 3hr
SwitchBoard (SWBD)	English	Call Cent.	Conv.	30hr, 5hr, N.A.
Wall Street Journal (WSJ)	English	Finance	Read	80hr, 1.1hr, 0.4hr

Fig. 3: Frobenius norm of the weight difference between CP (w_{pre}) and fine-tuned models (w_{fine}) for CP of Wav2Vec2-Lb on SWBD.

4. EXPERIMENTAL SETUP

Datasets and SSL Pre-trained models. Details on individual datasets used for all our experiments can also be found in Table 1. We use the splits mentioned in Table 1 for all our experiments in Table 2. Precisely, we evaluate our approach on the MSR low resource Indian language corpus (MSR) [19], Gramvani ASR dataset [20], SwitchBoard dataset [21] (SWBD) and the Wall Street Journal dataset [22]. For SWBD, we take a low-resource split, the rationale behind which is mentioned in Section 5.

For all our experiments in Table 2, we use either of: (1) Wav2Vec2 [2] pre-trained on 960 hours of LibriSpeech [10] (Wav2Vec2-Lb), (2) XLSR-300 model [23] pre-trained on a combination of VoxPopuli, MLS, CommonVoice, BABEL, and VoxLingua107, or (3) Vakyansh model [24] pre-trained on speech from Hindi YouTube videos. All these models are pre-trained on *read* speech from the *general* domain.

Baselines. Due to the lack of prior work in this space, we compare Stable Distillation with other commonly used pipelines that include: (1) No Continued Pre-training: For this baseline, we use the pre-trained SSL model without any continued pre-training on downstream dataset $\mathcal{D}_f$ and (2) Vanilla Continued Pre-training: For this baseline, we perform vanilla continued pre-training of our SSL model on downstream dataset $\mathcal{D}_f$.

Hyper-parameters. For stable distillation, we first perform continued pre-training on our teacher wav2vec-2.0 for a to-

Table 2: Comparison of Stable Distillation ASR results with our baselines on both Enc-Dec and E2E evaluation settings. All results are in the format of **dev / test**. $\mathcal{R}$ and $\mathcal{C}$ indicate Read and Conversational Speech. Domain Map refers to the source pre-training → CP domain.

Pretrained Model	Downstream Dataset	Domain Map (*Source → Target*)	No Cont. Pretrain		Vanilla Cont. Pretrain		Stable Distillation	
			Enc-Dec	E2E	Enc-Dec	E2E	Enc-Dec	E2E
XLSR-300	GV$_{Hindi}$	General$_{\mathcal{R}}$ → Call Cent.$_{\mathcal{C}}$	32.7 / 32.5	37.3 / 37.0	31.6 / 31.4	35.3 / 35.0	**28.9 / 27.7**	**30.3 / 30.1**
XLSR-300	MSR$_{Gujarati}$	General$_{\mathcal{R}}$ → General$_{\mathcal{C}}$	21.7 / 28.5	24.4 / 32.3	21.3 / 27.2	22.1 / 30.3	**20.2 / 26.4**	**20.2 / 28.4**
XLSR-300	MSR$_{Tamil}$	General$_{\mathcal{R}}$ → General$_{\mathcal{C}}$	28.1 / 27.7	33.4 / 32.1	27.8 / 26.9	32.2 / 31.2	**26.7 / 25.7**	**29.1 / 28.4**
XLSR-300	MSR$_{Telugu}$	General$_{\mathcal{R}}$ → General$_{\mathcal{C}}$	28.3 / 28.8	34.1 / 32.8	28.0 / 28.3	32.6 / 32.0	**27.2 / 27.1**	**29.4 / 28.1**
Vakyansh	GV$_{Hindi}$	General$_{\mathcal{R}}$ → Call Cent.$_{\mathcal{C}}$	34.5 / 34.3	33.2 / 34.2	32.7 / 32.5	31.7 / 31.5	**30.6 / 30.4**	**30.0 / 30.1**
Wav2Vec2-Lb	SWBD$_{English}$	General$_{\mathcal{R}}$ → Call Cent.$_{\mathcal{C}}$	39.1 / N.A	22.2 / N.A.	36.2 / N.A.	20.4 / N.A.	**31.4 / N.A.**	**14.3 / N.A.**
Wav2Vec2-Lb	WSJ$_{English}$	General$_{\mathcal{R}}$ → Finance$_{\mathcal{R}}$	12.4 / 11.6	11.4 / 10.9	11.3 / 10.6	10.5 / 10.0	**10.5 / 9.8**	**9.5 / 9.1**

Table 3: Performance comparison in Enc-Dec setting when CP → final fine-tuning language/domain differs. Results show that SD leads to better regularization. For Libri. we report clean | other.
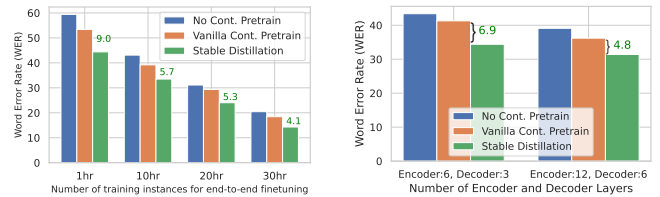
Model	Dataset	Vanilla Cont. Pretrain	Stable Distillation
XLSR-300	GV$_{Hindi}$ → MSR$_{Tamil}$	101.4 / 101.5	**47.8 / 47.8**
XLSR-300	GV$_{Hindi}$ → MSR$_{Telugu}$	99.4 / 99.7	**46.7 / 46.9**
Wav2Vec2-Lb	WSJ$_{English}$ → Libri$_{English}$	18.2 / 28.8 | 23.3 / 29.8	**16.0 / 25.2 | 20.1 / 27.6**

(a) Evaluation-set WER on SWBD on various low-resource settings for E2E setup.

(b) Evaluation-set WER on SWBD for various Enc-Dec layer configs on Wav2Vec2-Lb.

Fig. 4: More results. Stable Distillation outperforms all baselines.

tal of 50 epochs and then perform Stable Distillation on the student for 50 more epochs. For a fair comparison, for our continued pre-training baseline, we pre-train the model for a total of 50 epochs. We train with a learning rate of $5e^{-4}$ using Adam optimizer. Finally, for Stable distillation, we find $\alpha = 0.01$ to give us the best performance searched among $\{0.01, 0.05, 0.1, 0.5, 1\}$. For downstream fine-tuning, we use a conformer-based encoder-decoder model that has 12 encoder layers and 6 decoder layers. We train our models with a learning rate of $1.5e^{-3}$, batch size of 64, and for a total of 100 epochs. We report the Word Error Rate (WER) averaged over three runs with different random seeds.

5. RESULTS AND ANALYSIS

Table 2 compares the performance of Stable Distillation with all our baselines on the splits for all datasets mentioned in Table 1. For all experiments, we follow dataset settings resembling real-world scenarios where the target domain for CP is usually from a different domain and is low-resource compared to the source domain that the model is already pre-trained on (see Domain Map column in Table 2). SD, on average, outperforms our baselines by 0.8-7.7 WER in the Enc-Dec setup and 1-7 WER in the E2E setup. Precisely, for the Enc-Dec setup, compared to vanilla CP, we achieved an average relative WER improvement of 7.2% and 7.1% for dev and test sets, respectively. Compared to no CP, the improvements are 11.3% 11.1%. For the E2E setup compared to vanilla CP, we achieved an average relative WER improvement of 13.0% and 12.9% on dev and test sets, respectively. Compared to no CP, the improvements are 17.3% and 17.2%.
Table 3 confirms our claim that SD improves vanilla CP by adding explicit regularization, thereby preventing overfitting

on the target domain train set. Ideally, if a model has overfitted to the target training dataset after CP, it should perform worse if it is fine-tuned for ASR on the source dataset itself or any other dataset where the domain or language differs. While rows 1 and 2 show examples when the CP and fine-tuning languages differ, row 3 shows that SD leads to better performance than vanilla CP with LibriSpeech ASR when Wav2Vec2-Lb (also pre-trained on LibriSpeech) was further pre-trained on SWBD. Row 6 in Table 2 shows that the same model also performs better than CP with WSJ ASR (due to effectively exploiting past knowledge). Rows 1-6 also show a peripheral benefit of SD, which helps in effective crosslingual to mono-lingual adaptation of SSL models pre-trained on large-scale multi-lingual data [25].

6. CONCLUSION AND FUTURE WORK

In this paper, we propose Stable Distillation, a simple and effective methodology for target domain adaptation of pretrained SSL models. Stable Distillation regularizes continued pre-training of an SSL model on a target domain and prevents overfitting. We empirically show that Stable Distillation proves to be extremely effective in various settings where the initial and target domains differ. As part of future work, we would like to explore better distillation procedures for continued pre-training further to push ASR performance in trivial and simple steps.

7. REFERENCES

[1] Yang et al., "Superb: Speech processing universal performance benchmark," *arXiv preprint arXiv:2105.01051*, 2021.

[2] Baevski et al., "wav2vec 2.0: A framework for self-supervised learning of speech representations," *NeurIPS 2020*, pp. 12449–12460.

[3] Hsu et al., "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.

[4] Hsu et al., "Robust wav2vec 2.0: Analyzing domain shift in self-supervised pre-training," *arXiv preprint arXiv:2104.01027*, 2021.

[5] Sanyuan Chen, Yutai Hou, Yiming Cui, Wanxiang Che, Ting Liu, and Xiangzhan Yu, "Recall and learn: Fine-tuning deep pretrained language models with less forgetting," *arXiv preprint arXiv:2004.12651*, 2020.

[6] Armen Aghajanyan, Akshat Shrivastava, Anchit Gupta, Naman Goyal, Luke Zettlemoyer, and Sonal Gupta, "Better fine-tuning by reducing representational collapse," *arXiv preprint arXiv:2008.03156*, 2020.

[7] Gururangan et al., "Don't stop pretraining: Adapt language models to domains and tasks," *arXiv preprint arXiv:2004.10964*, 2020.

[8] Senthil Purushwalkam, Pedro Morgado, and Abhinav Gupta, "The challenges of continuous self-supervised learning," in *ECCV 2022*, pp. 702–721.

[9] Olivier Bousquet, Stéphane Boucheron, and Gábor Lugosi, "Introduction to statistical learning theory," in *Summer school on machine learning*, pp. 169–207. Springer, 2003.

[10] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *IEEE ICASSP 2015*, pp. 5206–5210.

[11] Andy T Liu, Shu-wen Yang, Po-Han Chi, Po-chun Hsu, and Hung-yi Lee, "Mockingjay: Unsupervised speech representation learning with deep bidirectional transformer encoders," in *IEEE ICASSP 2020*, pp. 6419–6423.

[12] Zhong Meng, Jinyu Li, Yashesh Gaur, and Yifan Gong, "Domain adaptation via teacher-student learning for end-to-end speech recognition," in *IEEE ASRU 2019*, pp. 268–275.

[13] Ruchao Fan and Abeer Alwan, "Draft: A novel framework to reduce domain shifting in self-supervised learning and its application to children's asr," *arXiv preprint arXiv:2206.07931*, 2022.

[14] He at al., "On the effectiveness of adapter-based tuning for pretrained language model adaptation," *arXiv preprint arXiv:2106.03164*, 2021.

[15] Vasista Sai Lodagala, Sreyan Ghosh, and S Umesh, "Pada: Pruning assisted domain adaptation for self-supervised speech representations," in *IEEE SLT 2022*, pp. 136–143.

[16] Shinji Watanabe, Takaaki Hori, Suyoun Kim, John R Hershey, and Tomoki Hayashi, "Hybrid ctc/attention architecture for end-to-end speech recognition," *IEEE Journal of Selected Topics in Signal Processing 2017*, pp. 1240–1253.

[17] Vaishnavh Nagarajan and J Zico Kolter, "Generalization in deep networks: The role of distance from initialization," *arXiv preprint arXiv:1901.01672*, 2019.

[18] Hossein Mobahi, Mehrdad Farajtabar, and Peter Bartlett, "Self-distillation amplifies regularization in hilbert space," *NeurIPS 2020*, pp. 3351–3361.

[19] Srivastava at al., "Interspeech 2018 low resource automatic speech recognition challenge for indian languages," 08 2018, pp. 11–14.

[20] Bhanushali et. al, "Gram vaani asr challenge on spontaneous telephone speech recordings in regional variations of hindi," in *INTERSPEECH 2022*, pp. 3548–3552.

[21] John J Godfrey, Edward C Holliman, and Jane McDaniel, "Switchboard: Telephone speech corpus for research and development," in *IEEE ICASSP 1992*, vol. 1, pp. 517–520.

[22] Douglas B Paul and Janet Baker, "The design for the wall street journal-based csr corpus," in *Speech and Natural Language: Proceedings of a Workshop Held at Harriman, New York, February 23-26, 1992*, 1992.

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

[24] Gupta at al., "Clsril-23: cross lingual speech representations for indic languages," *arXiv preprint arXiv:2107.07402*, 2021.

[25] Sameer Khurana, Antoine Laurent, and James Glass, "Magic dust for cross-lingual adaptation of monolingual wav2vec-2.0," in *IEEE ICASSP 2022*, pp. 6647–6651.