FUSDOM: COMBINING IN-DOMAIN AND OUT-OF-DOMAIN KNOWLEDGE FOR CONTINUOUS SELF-SUPERVISED LEARNING

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 pre-training (CP) offers multiple advantages, like target domain adaptation and the potential to exploit the continuous stream of unlabeled data available online. However, continued pre-training on out-of-domain distributions often leads to catastrophic forgetting of previously acquired knowledge, leading to sub-optimal ASR performance. This paper presents FusDom, a simple and novel methodology for SSL-based continued pre-training. FusDom learns speech representations that are robust and adaptive yet not forgetful of concepts seen in the past. Instead of solving the SSL pre-text task on the output representations of a single model, FusDom leverages two identical pre-trained SSL models, a teacher and a student, with a modified pre-training head to solve the CP SSL pre-text task. This head employs a cross-attention mechanism between the representations of both models while only the student receives gradient updates and the teacher does not. Finally, the student is fine-tuned for ASR. In practice, FusDom outperforms all our baselines across settings significantly, with WER improvements in the range of 0.2 WER - 7.3 WER in the target domain, while retaining the performance in the earlier domain[1].

Index Terms— speech recognition, self-supervised learning, continued pre-training, continual learning

1. INTRODUCTION

In the recent past, Self-Supervised Learning (SSL) has shown impressive performance on a variety of vision [1], text [2], speech [3, 4, 5], and audio tasks [6]. The primary goal is to learn representations from unlabeled data to learn high-level features that can transfer well across various tasks. In the past couple of years, the Spoken Language Processing (SLP) community has developed several sophisticated algorithms that achieve state-of-the-art (SOTA) performance on popular benchmarks [7]. A primary real-world application of SSL is to overcome the data scarcity problem for under-represented languages [8].

Continued SSL pre-training proves to be an effective solution in many real-world use cases, including domain adaptation to the low-resource target domain [9, 10, 11] and exploiting the continuous stream of unlabeled data online to keep the model's knowledge up-to-date. However, continued SSL pre-training leads to catastrophic forgetting of past knowledge [12] due to data that violates the IID assumption of optimization algorithms [13]. Forgetting past knowledge learned from large-scale pre-training leads to sub-optimal ASR performance in both the current and previous domains.

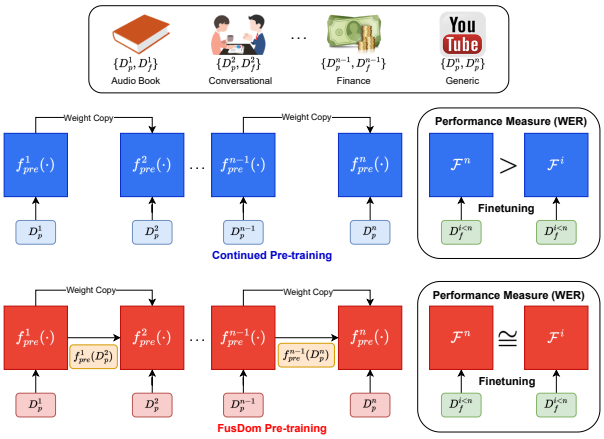

Fig. 1: Illustration of **FusDom**. FusDom facilitates continuous SSL on multiple new distinct domains without forgetting knowledge about past domains. As a result, the resultant model achieves optimal ASR performance in the current and all previous domains.

Main Contributions. In this paper, we propose FusDom, a simple and novel continued pre-training (CP) strategy for pre-training existing SSL models on non-IID data. FusDom brings the best of both worlds by simultaneously adapting a pre-trained model to the target downstream domain and avoiding forgetting past knowledge it has learned with large-scale SSL pre-training. To achieve this, FusDom employs two identical copies of an SSL model and leverages a novel SSL pre-training head to solve the pre-text task for CP. Of these two models, only one receives gradient updates (the student), while the other is always frozen (the teacher). The novel pre-training head employs a transformer layer with cross-attention, where the queries come from the teacher, and the keys and values come from the student. Effectively, the in-domain model representations solve the pre-text in an

* These authors contributed equally to this work.

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

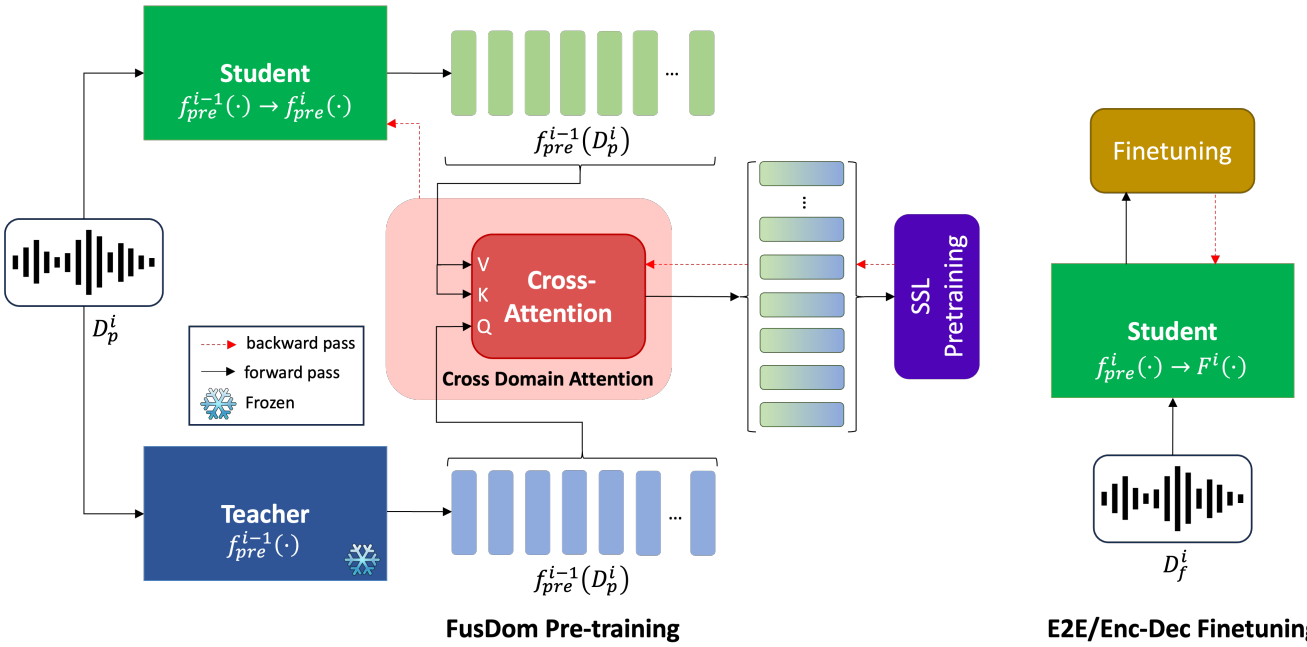

Fig. 2: Illustration of **FusDom**. FusDom employs two identical models, a student and a teacher, for SSL-based CP (instead of 1 in *vanilla* CP). For CP with target dataset $\mathcal{D}_p^i$, we initialize our student and teacher with $f_{pre}^{i-1}(.)$, which comes from the previous stage of pre-training, pre-trained on $\mathcal{D}_p^{i-1}$. The modified pre-training head h_{pre} consists of a standard transformer block with cross-attention between representations of the student and the teachers. Precisely, in h_{pre}, the teacher representations act as the query (Q) and the student representations act as the Key (K) and Value (V). Only the student receives gradient updates for SSL pre-training, while the teacher does not. Finally, the student model is either fine-tuned end-to-end using CTC or used as a feature extractor for fine-tuning an Encoder-Decoder model.

out-of-domain-aware fashion that eventually helps retain past acquired knowledge. To build FusDom, we are inspired by the normal human learning process, where humans leverage past acquired knowledge to infer and learn new and unseen concepts. We build FusDom on the core heuristic that neural networks have enough capacity to store information about every domain it sees in continued SSL [14]. We perform an extensive empirical evaluation to prove the efficacy of FusDom in various settings and show that FusDom achieves relative Word Error Rate (WER) gains in the range of 1.2-7.2 over our baselines on target domain ASR while retaining the performance in the previous domains.

2. RELATED WORK

SSL in Speech. Throughout the past decade, researchers have proposed several SSL algorithms that achieve new SOTA performance on several SLP benchmarks [15, 7]. Some of the most common ones are based on contrastive learning [3], clustering [4], or reconstruction [5]. Despite its success in pushing benchmark performance, SSL models suffer from several fundamental problems which, to the best of our knowledge, lack sufficient research in the speech community. For example, SSL models suffer from catastrophic forgetting when fine-tuning and pre-training domains differ [10]. Continued pre-training on the downstream dataset has shown to be a promising direction to avoid forgetting by adapting to the target domain [16]. However, effectively performing continued pre-training is difficult due to problems like overfitting the training data and catastrophic forgetting [12]. We acknowledge that continued pre-training overall is a relatively understudied problem in literature and specifically in speech representation learning, where the benefits and drawbacks of it are under-explored.

Continual Learning. One popular area of research that handles catastrophic forgetting of neural networks in the transfer learning paradigm is continual learning [17]. Compared to text and vision, continual learning for speech representation learning is a relatively under-explored area, leaving much to explore [18, 19]. Additionally, most prior work focuses on supervised-only settings, ignoring continual learning of models learned with SSL. Different from existing work, FusDom takes a first step toward solving catastrophic forgetting in the continued SSL pre-training paradigm.

3. METHODOLOGY

Problem Formulation. Fig. 2 shows a clear pictorial representation of our proposed approach. . Let's say we have an upstream model $f_{pre}(.)$ and a linear stream of n unlabeled datasets $\mathcal{D}_p \in \{\mathcal{D}_p^i, \cdots, \mathcal{D}_p^n\}$ where each $\mathcal{D}_p^i = (\mathbf{X}_{pre}^i)$

comes from a different domain than the previous. These unlabeled datasets are employed to perform SSL on $f_{pre}(.)$ sequentially, in any order, where at each step, the resultant model can be denoted as $f^i_{pre}(.)$. Additionally, each unlabeled dataset in $\mathcal{D}^i_p$, also has a corresponding downstream ASR dataset $\mathcal{D}^i_f$=($\mathbf{X}^i_{targ}$,$\mathbf{Y}^i_{targ}$). Our primary aim is to obtain a final upstream model $f^n_{pre}(.)$, which, when fine-tuned on any downstream ASR dataset $\mathcal{D}^i_f$, achieves optimal ASR performance, irrespective of the order in which the unlabeled datasets was shown to $f_{pre}(.)$. We denote this fine-tuned model, fine-tuned on $\mathcal{D}^i_f$ as $\mathcal{F}^i$. We either fine-tune the final pre-trained model $f^n_{pre}(.)$ using CTC or use it as a feature extractor to fine-tune a conformer-based Encoder-Decoder model. FusDom proposes an effective methodology for continuous SSL to prevent the model from forgetting previous domains seen in the earlier stage.

3.1. Continued Pre-training with FusDom

During continued pre-training, FusDom tries to avoid forgetting past knowledge by learning target-domain representations that are aware of past knowledge. To achieve this, FusDom employs a standard transformer block with cross-attention between multiple representations as the pre-training head, which we denote as h_{pre}. h_{pre} receives its input from two similar copies of pre-trained SSL model $f^i_{pre}(.)$. We call one of these models a student and the other a teacher. Precisely, the queries for cross-attention in h_{pre} come from the teacher, and the keys and values come from the student. Only the student receives gradient updates during SSL pre-training, while the teacher does not. For simplicity, let's denote the student representations as $\mathbf{S} \in \mathbb{R}^{d \times M}$ and the teacher representation as $\mathbf{T} \in \mathbb{R}^{d \times M}$. Formally, we can denote Cross-Domain Attention or $\mathbf{CDA}$ as:

$$\mathbf{CDA}(\mathbf{S}, \mathbf{T}) = \text{softmax}\left(\frac{[\mathbf{W_{q_i}}\mathbf{T}]^\top [\mathbf{W_{k_i}}\mathbf{S}]}{\sqrt{d/M}}\right)[\mathbf{W_{v_i}}\mathbf{S}]^\top \tag{1}$$

where $\{\mathbf{W_{k_i}}, \mathbf{W_{q_i}}, \mathbf{W_{v_i}}\} \in \mathbb{R}^{d/M \times h}$ denote the query, key, and value weight matrices, respectively, for the i^{th} attention head. Finally, the output of the Cross Domain Attention layer is passed through the standard feed-forward with a skip connection and a non-linear activation. Finally the output representation of the pre-training head h_{pre} is now $\mathbf{F} = (\mathbf{f_0}, \mathbf{f_1}, \cdots, \mathbf{f_{m-1}})$. For simplicity, we make the model solve the same SSL pre-text task that it solved during the earlier pre-training stage for all our experiments.

3.2. Downstream Fine-tuning for ASR

Next, we fine-tune $f^i_{pre}(.)$ on any target dataset, preferably from one of the domains already seen during the various stages of pre-training. For fine-tuning, we either employ End-to-End CTC Fine-tuning where we adjust all weights of $f^i_{pre}(.)$ by introducing a linear CTC head and subsequently optimizing the model with the CTC loss or use $f^i_{pre}(.)$ as a frozen feature extractor for fine-tuning a Conformer-based Encoder-Decoder. For the latter, we jointly optimize CTC and attention-based auto-regressive losses, as [20].

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

4. EXPERIMENTAL SETUP

Datasets. Details on individual datasets used for all our experiments can also be found in Table 1. **(1) MSR.** The MSR speech corpus [21] consists of about ≈150 hours of labeled ASR data for three Indian languages, namely, Gujarati, Tamil, and Telegu. The utterances are sourced from *human conversations*. **(2)** The Gramvani ASR dataset [22] consists of 100 hours of labeled ASR data, with a 100 / 5 / 3 hour train-dev-test split. The utterances are sourced from *telephonic conversations* in Hindi of varying regional dialects. **(3)** The SwitchBoard dataset [23] consists of 330 hours of labeled ASR data sourced from *telephonic conversations* in English. For our experiments, we sample a 30-hour training split. **(4)** The Wall Street Journal dataset [24] consists of 80 hours of labeled ASR data sourced from *read speech* of financial news in the Wall Street Journal.

SSL Pre-trained models. For our experiments, we employ either of these three pre-trained models: **(1) Wav2Vec2-Libri-960:** We use the base variant of Wav2Vec2 [3] pre-trained on 960 hours of LibriSpeech [15]. The model has ≈95M learnable parameters. **(2) XLSR-300:** [25] We use the variant of XLSR with ≈300M learnable parameters. This model is pre-trained on VoxPopuli, MLS, CommonVoice, BABEL, and VoxLingua107 for learning cross-lingual speech representation. **(3) Vakyansh.** [26] The Vakyansh model for pre-trained using the contrastive learning objective, similar to [3], on 4200 hours of Hindi Data from the read-speech domain. The model is built on the base variant of wav2vec-2.0, which has ≈95M learnable parameters.

Baselines. We compare FusDom with: **(1) No Continued Pre-training.** This baseline follows the most common ASR fine-tuning pipeline wherein we use the pre-trained SSL model without any continued pre-training. **(2) Vanilla Continued Pre-training.** This baseline employs an additional step over generic SSL pre-training by performing continued

Table 2: Comparison of FusDom 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		FusDom	
			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	**29.9 / 28.7**	**32.2 / 32.0**
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	**21.2 / 26.6**	**21.4 / 29.4**
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	**26.8 / 26.7**	**29.3 / 29.2**
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	**27.6 / 27.2**	**29.1 / 28.3**
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	**31.6 / 31.4**	**31.0 / 31.1**
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.	**32.4 / N.A.**	**15.2 / 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	**11.1 / 10.2**	**9.6 / 9.2**

Table 3: Comparison of FusDom ASR when CP on diverse target domains (SWBD, WSJ) and finetuned on source domains (Libri 10hr) with our baselines on both Enc-Dec and E2E evaluation settings. All results are in the format of **dev-clean / test-clean**

Pretrained Model	Downstream Dataset	Cont. Pretrained Dataset	No Cont. Pretrain		Vanilla Cont. Pretrain		FusDom	
			Enc-Dec	E2E	Enc-Dec	E2E	Enc-Dec	E2E
XLSR-300	$Libri_{English}$	$SWBD_{English}$	**13.7 / 15.2**	**10.3 / 10.3**	15.6 / 17.2	12.6 / 12.5	13.8 / 15.3	10.4 / 10.4
XLSR-300	$Libri_{English}$	$WSJ_{English}$	**13.7 / 15.2**	**10.3 / 10.3**	14.2 / 16.4	12.1 / 16.6	13.9 / 15.4	10.5 / 10.4
Wav2Vec2-Lb	$Libri_{English}$	$SWBD_{English}$	15.8 / 20.7	12.7 / 17.8	20.8 / 25.3	18.3 / 23.6	**15.2 / 20.4**	**12.3 / 17.4**
Wav2Vec2-Lb	$Libri_{English}$	$WSJ_{English}$	15.8 / 20.7	12.7 / 17.8	18.8 / 22.8	15.2 / 20.1	**15.6 / 20.4**	**12.6 / 17.7**

Table 4: Comparing the downstream performance of FusDom vs. Vanilla CP for sequential CP on diverse domains in E2E evaluation. All results are in format **Vanilla CP / FusDom** test set WER

Pretrained Model	Cont. Pretrain Order	Downstream		
		Libri	SWBD	WSJ
XLSR-300	SWBD → WSJ	15.8 / **11.0**	12.8 / **10.7**	9.3 / **8.1**
XLSR-300	WSJ → SWBD	13.4 / **10.8**	11.7 / **9.8**	9.7 / **9.2**
Wav2Vec2-Lb	SWBD → WSJ	21.0 / **18.3**	16.1 / **13.6**	9.8 / **9.1**
Wav2Vec2-Lb	WSJ → SWBD	24.1 / **18.7**	15.0 / **12.7**	10.1 / **9.7**

pre-training on the target domain.

Hyperparameters. For Continued Pre-training, we train our wav2vec-2.0 base SSL pre-trained model for 100 epochs. For a fair comparison, with our continued pre-training baseline, we also pre-train the FusDom-based model for a total of 100 epochs. We train FusDom with a learning rate of $5e^{-4}$ using Adam optimizer. Our conformer-based encoder-decoder model 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.

5. RESULTS AND ANALYSIS

Table 2 presents a performance comparison of FusDom with our baseline methods on dataset splits mentioned in Table 1. Our experiments mimic real-world scenarios, where the target domain for CP has fewer resources and differs from the source domain, as shown in the Domain Map column of Table 2. On average, FusDom surpasses the baselines by reducing WER by 0.2-7.3 in the Encoder-Decoder (Enc-Dec) setup and 1.1-7.0 in the End-to-End (E2E) setup. Specifically, in the Enc-Dec setup, relative WER improvements of 6.1% and 6.0% are achieved for the dev and test sets when compared to vanilla CP, and improvements of 10.2% and 10.1% when compared to no CP. In the E2E setup, relative WER improvements of 12.2% and 11.9% are observed on the dev and test sets compared to vanilla CP, and improvements of 16.9% and 16.7% when compared to no CP.

Table 3 highlights a performance comparison of FusDom with our baseline methods when CP on diverse domains and finetuned on the source domain. For CP we use *SWBD*, *WSJ*, and for finetuning XLSR-300 and Wav2Vec2-Lb., we use *Libri* 10hr split as our source domain. On average, FusDom gives similar performances as compared to no CP for E2E and Enc-Dec finetuning setups with a slight increase of 0.1-0.6 in absolute WER when compared to vanilla CP with an increase of 0.5-5 in absolute WER. This shows that FusDom is more efficient than vanilla CP in retaining previous domain knowledge. We also show the effect of sequential CP as shown in Fig 1 in Table 4, where FusDom outperforms vanilla CP with a decrease of 0.7-4.8 absolute WER on the E2E evaluation.

6. CONCLUSION AND FUTURE WORK

This paper proposes FuseDom, a novel methodology to continue pre-training an SSL model on a stream of unlabelled non-IID data. FuseDom avoids catastrophic forgetting by learning to solve the pre-text task with representations that are prior knowledge aware. In practice, FuseDom improves downstream ASR performance over all our baselines by a significant margin. As part of future work, we would like to build better learning systems for more effective continued pre-training and perform a layer-wise analysis of the information learned by FusDom to quantify forgetting.

7. REFERENCES

[1] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton, "A simple framework for contrastive learning of visual representations," in *International conference on machine learning*. PMLR, 2020, pp. 1597–1607.

[2] Devlin et al., "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

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

[4] 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.

[5] Liu et al., "Mockingjay: Unsupervised speech representation learning with deep bidirectional transformer encoders," in *IEEE ICASSP 2020*, pp. 6419–6423.

[6] Sreyan Ghosh, Ashish Seth, and S Umesh, "Decorrelating feature spaces for learning general-purpose audio representations," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1402–1414, 2022.

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

[8] Zhang et al., "Google usm: Scaling automatic speech recognition beyond 100 languages," *arXiv preprint arXiv:2303.01037*, 2023.

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

[10] Chen et al., "Recall and learn: Fine-tuning deep pretrained language models with less forgetting," *arXiv preprint arXiv:2004.12651*, 2020.

[11] Aghajanyan et al., "Better fine-tuning by reducing representational collapse," *arXiv preprint arXiv:2008.03156*, 2020.

[12] Purushwalkam et al., "The challenges of continuous self-supervised learning," in *ECCV*. Springer, 2022, pp. 702–721.

[13] Bousquet et al., "Introduction to statistical learning theory," in *Summer school on machine learning*, pp. 169–207. Springer, 2003.

[14] Lai et al., "Parp: Prune, adjust and re-prune for self-supervised speech recognition," *NeurIPS 2021*, pp. 21256–21272.

[15] Panayotov et al., "Librispeech: an asr corpus based on public domain audio books," in *ICASSP*. IEEE, 2015, pp. 5206–5210.

[16] Gururangan et al., "Don't stop pretraining: Adapt language models to domains and tasks," in *ACL 2020*, pp. 8342–8360.

[17] De Lange et al., "A continual learning survey: Defying forgetting in classification tasks," *IEEE transactions on pattern analysis and machine intelligence*, vol. 44, no. 7, pp. 3366–3385, 2021.

[18] Yang et al., "Online continual learning of end-to-end speech recognition models," *arXiv preprint arXiv:2207.05071*, 2022.

[19] Samik Sadhu and Hynek Hermansky, "Continual learning in automatic speech recognition.," in *Interspeech*, 2020, pp. 1246–1250.

[20] Watanabe et al., "Hybrid ctc/attention architecture for end-to-end speech recognition," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 8, pp. 1240–1253, 2017.

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

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

[23] 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.

[24] 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.

[25] Conneau et al., "Unsupervised cross-lingual representation learning for speech recognition," *arXiv preprint arXiv:2006.13979*, 2020.

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