Sample-Efficient Unsupervised Domain Adaptation of Speech Recognition Systems: A case study for Modern Greek

Georgios Paraskevopoulos [1], Theodoros Kouzelis [1], Georgios Rouvalis [1], Athanasios Katsamanis [1], Vassilis Katsouros [1], and Alexandros Potamianos [1]

[1]Affiliation not available

October 30, 2023

Abstract

Modern speech recognition systems exhibits rapid performance degradation under domain shift. This issue is especially prevalent in data-scarce settings, such as low-resource languages, where diversity of training data is limited.

In this work we propose M2DS2, a simple and sample-efficient finetuning strategy for large pretrained speech models, based on mixed source and target domain self-supervision. We find that including source domain self-supervision stabilizes training and avoids mode collapse of the latent representations. For evaluation, we collect HParl, a 120 hour speech corpus for Greek, consisting of plenary sessions in the Greek Parliament. We merge HParl with two popular Greek corpora to create GREC-MD, a test-bed for multi-domain evaluation of Greek ASR systems. In our experiments we find that, while other Unsupervised Domain Adaptation baselines fail in this resource-constrained environment, M2DS2 yields significant improvements for cross-domain adaptation, even when a only a few hours of in-domain audio are available. When we relax the problem in a weakly supervised setting, we find that independent adaptation for audio using M2DS2 and language using simple LM augmentation techniques is particularly effective, yielding word error rates comparable to the fully supervised baselines.

Sample-Efficient Unsupervised Domain Adaptation of Speech Recognition Systems: A case study for Modern Greek

Georgios Paraskevopoulos *Student Member, IEEE*, Theodoros Kouzelis, Georgios Rouvalis, Athanasios Katsamanis *Member, IEEE*, Vassilis Katsouros *Member, IEEE*, Alexandros Potamianos *Fellow, IEEE*

***Abstract*—Modern speech recognition systems exhibits rapid performance degradation under domain shift. This issue is especially prevalent in data-scarce settings, such as low-resource languages, where diversity of training data is limited. In this work we propose M2DS2, a simple and sample-efficient finetuning strategy for large pretrained speech models, based on mixed source and target domain self-supervision. We find that including source domain self-supervision stabilizes training and avoids mode collapse of the latent representations. For evaluation, we collect HParl, a 120 hour speech corpus for Greek, consisting of plenary sessions in the Greek Parliament. We merge HParl with two popular Greek corpora to create GREC-MD, a test-bed for multi-domain evaluation of Greek ASR systems. In our experiments we find that, while other Unsupervised Domain Adaptation baselines fail in this resource-constrained environment, M2DS2 yields significant improvements for cross-domain adaptation, even when a only a few hours of in-domain audio are available. When we relax the problem in a weakly supervised setting, we find that independent adaptation for audio using M2DS2 and language using simple LM augmentation techniques is particularly effective, yielding word error rates comparable to the fully supervised baselines.**

***Index Terms*—Unsupervised Domain Adaptation, Automatic Speech Recognition, Multi-Domain Evaluation, Greek Speech**

I. Introduction

Automatic Speech recognition (ASR) models have matured to the point where they can enable commercial, real-world applications, e.g., voice assistants, dictation systems, etc., thus being one of machine learning's success stories. However, the performance of ASR systems rapidly deteriorates when the test data domain differs significantly from the training data. Domain mismatches can be caused by differences in the recording conditions, such as environmental noise, room reverberation, speaker and accent variability, or shifts in the target vocabulary. These issues are extenuated in the case of low-resource languages, where diversity in the training data is limited due to poor availability of high-quality transcribed audio. Therefore, specialized domain adaptation approaches need to be employed when operating under domain-shift.

Unsupervised Domain Adaptation (UDA) methods are of special interest, as they do not rely on expensive annotation of domain-specific data for supervised in-domain training. In contrast to supervised approaches, where the existence of labeled data would allow to train domain-specific models, UDA methods aim to leverage data in the absense of labels to improve system performance in the domain of interest [1], [2]. In the context of speech recognition the importance of UDA is extenuated, as the transcription and alignment process is especially expensive and time-consuming. Adaptation methods have been explored since the early days of ASR, at different levels of the system and different deployment settings [3]. UDA has been used to improve the robustness of ASR on a variety of recording conditions including far-field speech, environmental noise and reverberation [4], [5], [6]. Furthermore, UDA has been used for speaker adaptation, and to improve performance under speaker, gender and accent variability [7], [8]. UDA has also been employed for multilingual and cross-lingual ASR, in order to improve ASR models for low-resource languages [9], adapt to different dialects [10], and even train speech recognition systems for endangered languages [11].

Classical speech adaptation techniques involve feature-based techniques, e.g., speaker normalization [12], feature-based approaches [13]–[15], or multi-condition training [16]. Generally, traditional approaches require some knowledge about the target domain, and the domain mismatch, e.g., regarding the noise and reverberation variability [17], and require specific engineering for each adaptation scenario.

Modern ASR pipelines, increasingly rely on end-to-end neural networks, e.g., [18], [19], or large pretrained models with self-supervised objectives [20], [21]. The key approaches employed for UDA of end-to-end ASR models can be grouped in three categories, namely, teacher-student learning [10], domain adversarial training [22], and target domain self-supervision [23]. The benefit of these techniques is that they do not require any special knowledge about the source or the target domain. This makes end-to-end UDA approaches versatile and able to be utilized in a larger array of adaptation scenarios. In particular, adaptation through self-supervision has been shown to be a robust, simple and efficient technique for adaptation of state-of-the-art speech models [24].

Here, we leverage in-domain self-supervision to propose the Mixed Multi-Domain Self-Supervision (M2DS2) finetuning strategy, enabling sample-efficient domain adaptation of wav2vec2 [20] based speech recognition models, even when available in-domain data are scarce. Our key contributions are

G. Paraskevopoulos is with the Graduate School of ECE, National Technical University of Athens, Athens, Greece

G. Paraskevopoulos, T. Kouzelis, G. Rouvalis, A. Katsamanis, V. Katsouros are with the Institute for Speech and Language Processing, Athena Research Center, Athens, Greece

A. Potamianos is with the Faculty of ECE, National Technical University of Athens, Athens, Greece

TABLE I
SUMMARY OF RELATED WORKS ON UNSUPERVISED DOMAIN ADAPTATION FOR ASR.

Work	Method	Model	Adaptation Setting	Language
[23], [25], [26]	Teacher-Student Hard and soft labels	Conformer RNN-T [27] Transformer CTC RNN-T [19]	News speech, Voice search, Far-field, Telephony, YouTube	English
[4], [5]	Teacher-Student Soft labels	TDNN-LSTM [28]	Noise, Far-field	English
[29]	Teacher-Student Hard and soft labels	NiN-CNN [30]	Dialects Children speech	Japanese
[31]	Teacher-Student Soft labels	Streaming RNN-T [32]	Multilingual	English, Brazilian Portuguese, Russian, , Turkish, Nordic/Germanic
[6], [33], [34]	Domain Adversarial Training	TDNN Kaldi [35], [36] DNN-HMM DNN-HMM	Noise, Channel	English
[37]	Domain Adversarial Training	RNN-CTC [38]	Far-field	English
[8], [39]	Domain Adversarial Training	TDNN Kaldi RNN-T	Accent	Mandarin
[7], [40]	Domain Adversarial Training	DNN-HMM CNN-DNN	Speaker, Gender, Accent	English
[9]	Domain Adversarial Training	DSN [41]	Multilingual	Hindi, Sanskri
[24], [42]	Continual Pre-Training	wav2vec2 [20]	Audiobooks, Accents, Ted Talks, Telephony, Crowd-sourced, Parlamentary speech	English
[43]	Continual Pre-Training	wav2vec2	Cross-lingual	Korean
[11], [44]	Continual Pre-Training	XLSR-53 [21] wav2vec2	Low resource languages	Ainu Georgian, Somali, Tagalog, Farsi

organized as follows:

1) Inspired by recent advances on UDA for Natural Language Processing systems [45], we propose a finetuning strategy for speech models, where the self-supervised objective is based on a contrastive loss in Section III. Contrary to prior works, who leverage only in-domain self-supervision, we find that in this contrastive setting this leads to mode-collapse of the latent representations, and mixed source and target domain self-supervision is essential. We demonstrate this empirically in Section VII-B.
2) We collect and curate HParl, the largest publicly available[1] speech corpus for Greek, collected from plenary sessions in the Greek Parliament between 2018 and 2022. We establish a data collection, pre-processing and alignment pipeline that can be used for continuous data integration, as the parliamentary proceedings get regularly uploaded. We provide a detailed description of our data collection process and the dataset statistics in Section IV-A. HParl is merged in Section IV with two popular Greek corpora (Logotypografia and Common-Voice) to create GREC-MD, a testbed for multi-domain evaluation of ASR systems in Greek.
3) We demonstrate that, while other baselines fail at UDA in our resource-constrained setting, M2DS2 can improve model performance in the target domain in multiple adaptation scenarios in Section VII. Specifical emphasis is given in the sample efficiency of our approach in Section VII-A, where we demonstrate successful adaptation even when we reduce the available in-domain data.
4) When we relax the problem to a weakly supervised adaptation setting, where some in-domain text is available but the pairing between audio and text is unknown, we find that M2DS2 can be effectively combined with simple N-gram adaptation techniques to get comparable performance with the fully supervised baseline in Section VIII. Furthermore we find that a simple text augmentation approach, based on perplexity filtering of a large corpus can produce strong adaptation results, even for small amounts of in-domain text.

Additionally, we provide a formulation of the UDA problem for ASR in Section II-A and link prior works to this formulation in Sections II-B, II-C and II-D. We provide detailed experimental settings for reproducibility in Section V, and an upper-bound estimation for UDA performance with fully supervised finetuning in Section VI.

[1]We plan to release this version of HParl under the CC BY-NC 4.0 license upon publication. The other corpora used in this work are available through their respective distributors.

II. BACKGROUND

We start by formally defining the Unsupervised Domain Adaptation (UDA) problem. Initially, we formulate the problem in a classification setting and then we extend it for speech recognition. We then provide an overview of different adaptation approaches in the literature, and link each approach to the UDA problem formulation. Table I presents a summary of the key adaptation settings and applications that are explored in the literature. We see, that a relatively small amount of methods, and their variants, is used to address multiple real-world ASR problems, for example, cross-lingual, accent, speaker and noise adaptation. Furthermore, while the majority

of the works focus on the English language, there is an effort to explore other popular languages, e.g., Mandarin, and under-resourced languages, e.g., Ainu, Somali etc.

A. Problem Definition

Formally, the problem of UDA can be defined as follows. Let $X \subseteq \mathbb{R}^n$ be a real-valued space that consists of n-dimentional feature vectors $x \in X$, and Y a finite set of labels $y \in Y$, i.e., $Y = \{1, 2, \ldots, L\}$. Furthermore, assume two different distributions, i.e., the source domain distribution $\mathcal{S}(x, y)$ and the target domain distribution $\mathcal{T}(x, y)$, defined on the cartesian product $X \times Y$.

The goal is to train a model that learns a mapping between feature vectors $x_{\mathcal{T}}$ to their respective labels $y_{\mathcal{T}}$ for samples drawn from the target distribution $(x_{\mathcal{T}}, y_{\mathcal{T}}) \sim \mathcal{T}$.

At training time we have access to samples from the source distribution $\mathcal{S}(x, y)$ and the marginalized target distribution $\mathcal{T}(x)$, i.e., no target labels are provided. We define the training dataset D as the concatenation of the source and target training sets, $D = (D_S, D_T)$. D_S and D_T are defined as sequences of tuples, i.e.,

$$\begin{aligned} D_S &= \{(x_i, y_i) \,|\, (x_i, y_i) \sim \mathcal{S}(x, y),\, 1 \leq i \leq N\} \\ D_T &= \{(x_i, \emptyset) \,|\, x_i \sim \mathcal{T}(x),\, 1 \leq i \leq M\}, \end{aligned} \tag{1}$$

where we draw N samples from $\mathcal{S}(x, y)$ and M samples from $\mathcal{T}(x)$. Finally, we augment tuples in D with a domain indicator function:

$$\begin{aligned} D &= \{(x_i, y_i', \mathbb{1}_i) \,|\, 1 \leq i \leq N + M\} \\ \mathbb{1}_i &= \begin{cases} 0 & \text{if } \; x_i \sim \mathcal{S}(x), \\ 1 & \text{if } \; x_i \sim \mathcal{T}(x). \end{cases} \\ y_i' &= \begin{cases} y_i & \text{if } \; x_i \sim \mathcal{S}(x), \\ \emptyset & \text{if } \; x_i \sim \mathcal{T}(x). \end{cases} \end{aligned} \tag{2}$$

1) Unsupervised (Acoustic) Adaptation for ASR: The above definition can be directly extended in the case of speech recognition, with some modifications. In detail, we modify the feature space X, to be the set of (finite) sequences of real-valued feature vectors $(x_k)_{k \in \mathbb{N} \setminus \{\infty\}} \in X \subseteq (\mathbb{R}^n)^*$. Furthermore, the label space Y is modified to be the set of sequences $(y_m)_{m \in \mathbb{N} \setminus \{\infty\}}$, where $Y = (\{1, 2, \ldots, L\})^*$ contains finite-length sequences over a finite lexicon. For CTC training we make the assumption that $k > m$ for any sample (x_k, y_m), i.e., feature sequences are longer than their respective label sequences [46]. The rest of the definitions need no modifications.

2) Unsupervised (Language) Adaptation for ASR: Adaptation for ASR systems can also be performed at the language level, i.e., the label space. In this setting, we assume that the target domain samples are drawn from the marginalized target distribution $\mathcal{T}(y)$. The target dataset D_T now consists of tuples in the form $(\emptyset, y_i)$, where y_i is the label word sequence $(y_m)_{m \in \mathbb{N} \setminus \{\infty\}}$ for the i-th sample.

3) Weakly supervised Adaptation for ASR: The last setting we explore is the case were both audio and language in-domain samples are available, but the mapping between them is unknown. This situation can be encountered in real-world settings, e.g., in the case in-domain audio and text are collected independently. For example consider the case where audio clips from news casts are collected, along with contemporary newspaper articles. Another example is the case where long audio clips alongside with transcriptions are available, but no fine-grained time alignments[2]. In this case the target domain samples are drawn independently from the marginalized distributions $\mathcal{T}(x)$ and $\mathcal{T}(y)$, and the target dataset D_T consists of tuples in the form $(x_i, \emptyset)$ and $(\emptyset, y_i)$.

B. Teacher-Student Models

Teacher-Student learning or self-training, is one of the earliest methods in semi-supervised learning [47]–[49]. The key idea is to reduce the problem of unsupervised learning of the task at hand in the target domain to a supervised one. The general methodology is to train a teacher model g_S using the labeled data in the source domain D_S, and then use this for inference on the target domain to produce pseudolabels $\hat{y}_i = g_S(x_i)$, $x_i \sim \mathcal{T}(x)$. The target domain dataset D_T is augmented with these silver labels, to contain tuples $(x_i, \hat{y}_i)$. Finally, a student model g_T is trained in a supervised fashion, using the augmented D_T or a combination of D_S and D_T. This process is usually repeated, with the student model serving as the teacher model for the next iteration, until no further improvement is observed. More recently, soft target Teacher-Student learning has been explored for ASR [26], [31], [50], where the KL divergence between the teacher and student output label distributions is used as the loss function.

Being trained only on the source domain data the teacher model is susceptible to error propagation. Filtering is a commonly used technique to achieve the right balance between the size of the target domain used for training the student model and the noise in the pseudolabels. Confidence scoring based on the likelihood is usually applied, discarding those utterances for which the hypothesized labels are untrustworthy [51]. In [25] dropout is used to measure the model uncertainty. The agreement between model predictions with and without dropout are used for confidence scoring. In [23] a multi-task training objective with a confidence loss is applied to minimise the binary cross entropy between the estimated confidence and the binary target sequence. In order to learn more robust and generalizable features from the teacher model, Noisy Student Training (NST) has been proposed in [52]. The teacher models generates pseudolabels for D_T while the student models are trained on a heavily augmented version of D_T [52]. In [52], [53] the augmentation of the input target data is performed with SpecAugment [54], while in [29] a spectrum frequency augmentation is performed.

In [4] Teacher-Student learning with soft labels is introduced for ASR to tackle noisy, far-field, and children speech. In

[2]While a fully supervised in-domain dataset can be constructed in this case using long / forced alignment methods, this is not a focal point for the experimental part of this work.

[5], this approach is extended for LF-MMI based models and used for noisy, far-field and bandwidth adaptation. In [29] a weighted sum of hard and soft target cross entropy losses is used for Japanese dialects and children speech adaptation. Ramabhadran et al. [31] propose a self-adaptive distillation, and a method for distilling from multiple teachers that is applied across several multilingual ASR systems for different language groups. A comparison between soft and hard targets for RNN-T models [19] showed that soft targets perform better when both the teacher and student models have the same architecture. Otherwise, hard targets are superior [50].

C. Domain Adversarial Training

Domain Adversarial Training (DAT) was initially introduced for image classification [55]. The key idea is to train a model that learns deep features that solve the task at hand in the source domain, while being invariant with respect to the domain shift. Concretely, the model is trained end-to-end using a combination of the supervised task loss L_t, learned on D_S, and the domain discrimination loss L_a, i.e., $L = L_t - \alpha L_a$. The loss L_a is binary cross-entropy, trained for domain discrimination using the tuples $(x_i, \mathbb{1}_i)$. Notice the $-$ sign in the loss indicates adversarial learning, i.e., the model should learn features that cannot discriminate between domains, while solving the task.

In [6] DAT is employed for noise adaptation on a noise corrupted version of WSJ [56] as the target dataset. Using the Aurora-4 [57] dataset which has labels associated to the noise type, Serdyuk et al. [33] train an adversarial noise classifier. In [8] and [39] DAT is utilized for accent adaptation for Mandarin and English respectively. Anoop C.S. et al. [9] propose DAT, to address the scarcity of data in low-resource languages which share a common acoustic space with a high-resource language, namely Sanskrit and Hindi. They empirically demonstrate the effectiveness of adversarial training, presenting experiments with and without the reversal of the domain classification loss.

D. Leveraging In-domain Self-supervision

These lines of work have roots in Natural Language Processing tasks [45], [58], and explore domain adaptation by leveraging the in-domain data D_T for self-supervised learning. The core focus is domain adaptation of large pre-trained models, e.g., [59], and self-supervision is achieved by use of the pre-training self-supervised loss L_s. This process can either take part in stages, via continual pre-training [58], or by constructing a multitask objective $L = L_t + \alpha L_s$, as in [45].

Continual Pre-Training (CPT) has been explored for adaptation of ASR models. Robust wav2vec2 [24] explores the effectiveness of CPT for domain adaptation, indicating the importance of utilizing unlabeled in-domain data. In CASTLE [42], CPT is combined with an online pseudolabeling strategy for domain adaptation of wav2vec2. Cross-dataset evaluation for popular English speech corpora indicates that CPT helps to reduce the error rate in the target domain. In [43] and [11] CPT is utilized for cross-lingual adaptation of wav2vec2 for Korean and Ainu respectively. Notably for Ainu, which is an endagered language, CPT has resulted in significant system improvement. DeHaven and Jayadev [44] compare CPT and pseudolabeling for adapting XLSR-53 to four under-resourced languages, i.e., Georgian, Somali, Tagalog and Farsi. They find that both approaches yield similar improvements, with CPT being the more computationally efficient approach.

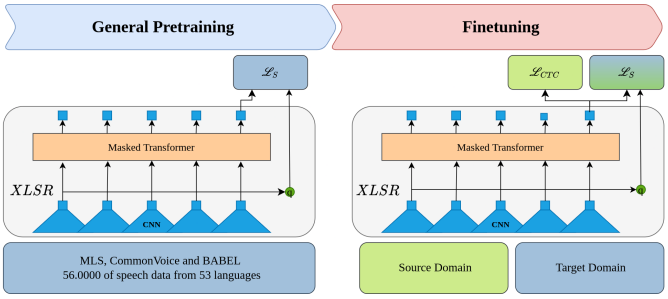


Fig. 1. Target-domain adaptation through self-supervision. In the left we see the general pre-training stage of XLSR-53 using the self-supervised loss L_s. General pre-training is performed on $56,000$ hours of audio in 53 languages. In the right, we see the proposed domain-adaptive finetuning stage, where the speech recognition task is learned using transcribed source domain data, while adaptation to the target domain is performed by including the self-supervised loss over (audio-only) source and target domain data

While CPT yields significant improvements in a variety of tasks, one common theme in these works is the assumption of hundreds or thousands of hours of available in-domain data, mostly from online resources, e.g., YouTube. This can be infeasible when we consider more niche adaptation settings, or possible privacy concerns, e.g., how would one collect 1000 hours of psychotherapy sessions in Greek? In this work, we explore domain adaptation methods in a more resource-constrained environment.

III. Domain Adaptation Through Multi-Domain Self-Supervision

The proposed approach is based on end-to-end adaptation of a large pre-trained speech model during the finetuning phase, by including in-domain self-supervision. We extend UDALM [45], that has shown promise for NLP tasks, for adaptation of wav2vec2 based acoustic models, and specifically XLSR. We focus on the problem of UDA in the context of a low-resource language, i.e., Greek. The key finding of our exploration is that straight-forward extension of UDALM, i.e., by using only target domain self-supervision, underperforms in this setting, and use of both source and target domain data is essential for successful adaptation. In this section, first, we will present a quick overview of the XLSR-53 training procedure, and then we are going to outline the proposed domain adaptation approach, which is shown in Fig. 1.

A. XLSR-53

XLSR-53 [21] is a massively pre-trained speech model, trained on $56,000$ hours of multilingual speech, covering 53 languages. The model is based on wav2vec2 [20], which is composed of a multi-layer convolutional feature encoder, that

TABLE II
THE GREC-MD CORPUS. WE CAN SEE THE DURATION OF EACH SPLIT IN HOURS:MINUTES:SECONDS FORMAT, AS WELL AS THE NUMBER OF SPEAKERS FOR EACH OF THE SUB-CORPORA.

Dataset	Domain	Speakers	Train	Dev	Test	Total Duration
HParl	Public (political) speech	387	99:31:41	9:03:33	11:12:28	119:47:42
CV	Crowd-sourced speech	325	12:16:17	1:57:44	1:59:19	16:13:20
Logotypografia	News casts	125	51:58:45	9:08:35	8:59:22	70:06:42
Total	-	713	163:46:43	20:09:52	22:11:44	206:08:19

extracts audio features z_t from the raw audio, and a transformer context encoder that maps the latent audio features to the output hidden states c_t. Each latent feature z_t corresponds to 25 ms of audio with stride 20 ms. A contrastive objective L_c is used for pre-training. For this, product quantization [60] is applied to the features z_t, and then a discrete approximation of z_t is obtained by sampling from a Gumbel-softmax distribution [61], to obtain discrete code vectors q_t, organized into $G = 2$ codebooks with $V = 320$ vocabulary entries each. The contrastive loss aims to identify the correct code vector for a given time step, among a set of distractors Q_t, obtained through negative sampling from other timesteps. To avoid mode collapse, a diversity loss L_d is included by maximizing the entropy over the averaged softmax distribution over the code vector entries $\bar{p}_g$. The total loss is:

$$L_s = \underbrace{-log\frac{e^{s(z_t,q_t)}}{\sum_{\tilde{q}\sim Q_t} e^{s(z_t,\tilde{q})}}}_{\text{Contrastive Loss}} \overbrace{-\frac{1}{GV}\sum_{g=1}^{G}\sum_{v=1}^{V}\bar{p}_{g,v}log(\bar{p}_{g,v})}^{\text{Diversity Loss}} \quad (3)$$

B. Domain Adaptive finetuning for Contrastive Learning of Speech Representations

Fig. 1 shows the proposed finetuning process. The key intuition is that we want the model to synergistically learn the task at hand (in our case ASR), while being adapted to the target domain by in-domain self-supervision. In the left we see the general pre-training stage of XLSR-53, which is pre-trained on 56K hours of multilingual audio corpora using the contrastive pre-training objective. In the right we see the proposed finetuning stage, which is inspired by [45].

During finetuning we form a mixed objective function:

$$L = L_{CTC}(x_s, y_s) + \alpha L_s(x_s) + \beta L_s(x_t), \quad (4)$$

where $(x_s, y_s) \sim \mathcal{S}(x, y)$, $x_t \sim \mathcal{T}(x)$, L_{CTC} is the CTC objective function, optimized using transcribed source domain data, and L_s is the contrastive loss from Eq. (3). We scale the contribution of each term using hyper-parameters α and β.

Note that contrary to [45], who use only in-domain self-supervision, we leverage both source and target domain samples for the mixed self-supervision. We find that this is essential in our case to avoid mode collapse, i.e., the model using only a few of the available discrete code vectors. Simultaneous self-supervision on both the source and target data alleviates mode collapse by anchoring the target code vector space to have a similar structure as the source code vectors.

Hence we refer to this approach as Mixed Multi-Domain Self-Supervision (M2DS2).

IV. THE GREC-MD CORPUS

For our experiments we compose a speech corpus for the Greek language, that is suitable for multi- and cross-domain evaluation. The GREC-MD corpus contains 206 hours of Greek speech. Audio is segmented into individual utterances and each utterance is paired with its corresponding transcription. Table II summarizes the included sub-corpora, as well as the train, development and test splits. The dataset is constructed with three core principles in mind:

1) **Data Volume**: We collect the largest publicly available speech recognition corpus for the Greek language, able to scale to hundreds of hours of transcribed audio.
2) **Temporal Relevance**: Language changes over time. We aim at an up-to-date corpus that encompasses the latest terms and topics that appear in daily speech.
3) **Multi-Domain Evaluation:** Single domain evaluation can lead to misleading estimations of the expected performance for ASR models. For example, state-of-the-art ASR models [27] achieve under 5% Word Error Rate (WER) on Librispeech [62] test sets, but this is an over-estimation of system performance in the field. This is extenuated when considering different acoustic conditions or terminology. We consider multi-domain evaluation essential when developing and deploying real-world ASR models.

To satisfy the first two points, we collect data from a public, continuously updated resource, i.e., the Hellenic Parliament Proceedings, where recordings of the parliamentary sessions are regularly uploaded. The benefit of using this resource is the straight-forward collection of a continuously growing, multi-speaker corpus of transcribed audio that is always up-to-date, as the parliamentary discussions revolve around current affairs. We refer to this corpus as HParl. For the multi-domain evaluation, we merge HParl with two publicly available corpora, that have different acoustic and language characteristics. We refer to the merged, multi-domain corpus as GREC-MD. In this Section, we will describe the collection and curation process of HParl, and present the relevant statistics for the experiments.

TABLE III
PLENARY SESSIONS INCLUDED IN HPARL. THE HOURS COLUMN REFERS TO THE RAW (UNSEGMENTED) HOURS OF COLLECTED AUDIO.

Start date	End date	#Sessions	Hours
15-02-2022	01-03-2022	10	55
18-01-2019	01-02-2019	10	52
28-03-2019	10-05-2019	20	108
10-12-2018	21-12-2018	10	88

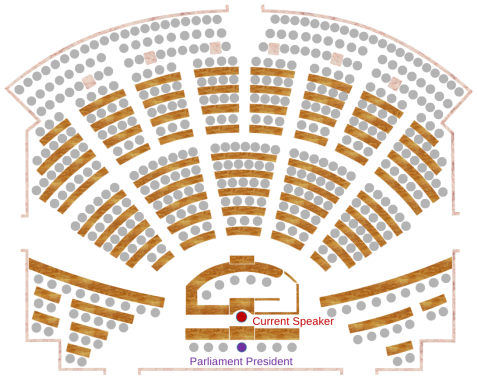


Fig. 2. Overview of the Hellenic Parliament Chamber. The chamber has an amphitheatrical shape and can accomodate approximately $400 - 450$ people. The positions of the key speakers, i.e., current speaker and the parliament president are annotated in the image.

A. Collection and Curation of HParl

Modern technological advances allow for more direct government transparency, through the commodification of storage and internet speeds. In this spirit, the records of plenary sessions of the Hellenic Parliament are made publicly available, for direct access through a webpage[3]. The available video recordings date back to 2015. For each plenary session, a video recording is uploaded, along with a full transcription that is recorded verbatim, and in real time by the parliament secretaries. For the creation of HParl, we build a web-crawler that can traverse and download the video recordings, along with the transcriptions from the official website. The collection process is parallelized over multiple threads, and parameterized by a range of dates and, optionally, a target corpus size in GB or in hours. For this version of HParl, we collect the plenary sessions in four date ranges, as described in Table III. The majority of the collected sessions are from 2019, but we also include sessions from 2018 and 2022 to include coverage of different topics. The individual components of the HParl curation pipeline are: Audio Pre-processing, Text Pre-processing, Alignment, Post-processing, and dataset Splitting.

1) Audio Pre-processing: Fig. 2 shows the layout of the Hellenic Parliament Chamber. Plenary sessions mainly take place in this room, or in the secondary House Chamber that has similar setup but is smaller in size. Because of the room and microphone characteristics, the captured audio in the video streams contains reverberation, due to sound reflections. We employ a light preprocessing pipeline, by passing the input video streams through FFmpeg, and converting them to monophonic, lossless audio format at 16000 Hz sampling rate. The resulting audio is not passed through any de-reverberation or speech enhancement software. The resulting audio files have a minimum, average and maximum duration of 6 minutes, 6 hours and 16 hours respectively.

2) Text Pre-processing: The text files contain full, word-by-word transcription of the speeches and questions asked by members of the audience, as well as extra annotations made by the parliament secretaries. Some annotations are relevant, i.e., the speaker name, while others are plain text descriptions of events happening during the session and need to be filtered out (e.g., "The session is interrupted for a 15 minute break"). We use a rule-based system, based on regular expressions, that filters the unnecessary information, keeping only the transcriptions and the speaker names. The speaker labels are created by transliterating their names and roles from Greek to Greeklish using the "All Greek to Me!" tool [63]. Text is lower-cased and normalized to remove multiple whitespaces. The result is a text file containing the raw transcriptions, and a mapping from speaker labels to their respective text parts.

3) Aligment and Segmentation: The primary challenge of exploiting the plenary sessions for ASR purposes is the length of the plenary recordings, as their durations vary from 6 minutes to 16 hours in length. However, data samples used to train ASR are generally less than 30 seconds long. Computational challenges have limited the length of training utterances for HMM-GMM models [64], and continue to do so in the contemporary neural network models. Therefore, we need to segment the sessions into smaller pieces more suitable for ASR training. A second challenge is posed by mismatches between audio and transcripts. Parliamentary proceedings do not fully capture everything that is said during the parliamentary sessions, and do not account for speech disfluencies.

In order to obtain smaller, clean segments, that are suitable for ASR training we follow the segmentation procedure proposed by [65]. Initially the raw recordings are segmented into 30 second segments and the transcriptions are split into smaller segments of approximately 1000 words called *documents*. Each segment is decoded using a seed acoustic model trained on the Logotypografia corpus [66] and a 4-gram biased LM trained on the corresponding transcription of each recording. The best path transcript of each segment is obtained and paired with the best matching *document* via TF-IDF similarity. Finally each hypothesis is aligned with the transcription using Smith-Waterman alignment [67] to select the best matching sub-sequence of words. The above method yields a list of text utterances, with their corresponding start and end times in the source audio files. The procedure yields 120 hours of useable segmented utterances out of the original 303 hours of raw audio, or a ratio of 39.6%.

4) Post-processing: After the segments are extracted, we filter out extremely short segments (less than 2 words). Moreover, the iterative alignment algorithm may replace some intermediate words with a `<spoken-noise>` tag. When this tag is inserted, we match the surrounding text with the raw transcriptions and re-insert the missing words. Furthermore, we match each segment to its corresponding speaker label. Segments without a speaker label are discarded. Lastly, speakers are associated to their gender based on name suffixes, using a simple, Greek language-specific, rule: Speaker names which end in a(α), h(η), w(ω) or is(ις) are classified as female, while the rest as male. We format the segments, speaker and gender mappings in the standard folder structure used by the Kaldi speech recognition toolkit [36].

5) Data Splitting: We provide an official train - development - test split. The development set contains 3 plenary sessions, one from 2018, one from 2019 and one from 2022,

[3]https://www.hellenicparliament.gr/en/

resulting to 9 hours of segmented speech. Similarly, the test set contains one session from each year, resulting to 11 hours of segmented speech. The rest 99 hours of segmented speech are assigned to the training set.

B. Including corpora from different domains

We merge HParl with two publicly available corpora to create GREC-MD for multi-domain evaluation.

1) Common Voice: Common Voice (CV) [68] is a crowd-sourced, multi-lingual corpus of dictated speech, created by Mozilla. The data collection is performed by use of a web app or an iPhone app. Contributors are presented with a prompt and are asked to read it. The prompts are taken from public domain sources, i.e., books, wikipedia, user submitted prompts and other public corpora. The maximum prompt length is 15 words. A rating system is built into the platform, where contributors can upvote or downvote submitted `<audio,transcript>` pairs. A pair is considered valid, if it receives two upvotes. Speaker independent train, development and test splits are provided. The dataset is open to the research community, released under a permisFsive Creative Commons license (CC0). In this work, we use version 9.0 of CV, accessed on April 27, 2022. We keep only the valid utterances, i.e., 16 hours of speech from 325 contributors ($19-49$ years old, 67% male / 23% female).

2) Logotypografia: Logotypografia [66] is one of the first corpora for Large Vocabulary Continuous Speech Recognition in Greek. The dataset contains $33,136$ newscast utterances, or 72 hours of speech. The utterances were collected from 125 speakers (55 male, 70 female), who were staff of the popular "Eleftherotypia" newspaper in Greece, under varied acoustic conditions. Approximately one third of the utterances were collected in a sound proof room, one third in a quiet room and the last third in an office room. The average utterance duration is 7.8 seconds. The transcriptions contain several speech and non-speech events (e.g., `<cough>`), lower-cased Greek words and stress marks. Numbers are expanded to full words. We use the whole dataset, and perform light preprocessing in the transcriptions, by discarding the annotated events and punctuation.

We hence refer to each dataset by the abbreviations: HParl: HP, CommonVoice: CV, Logotypografia: LG.

V. Experimental Settings

For our experiments we use the following hyper-parameter settings, unless explicitly stated otherwise. For model training, we use AdamW optimizer [69] with learning rate 0.0003. We apply warmup for the first 10% of the maximum training steps, and a linear learning rate decay after that. Models are finetuned for a maximum of 10000 steps. For speech recognition training, we make use of the Connectionist Temporal Classification (CTC) loss [70], optimized using the available transcribed data in each scenario. Validation runs every 500 steps on the development set, and early stopping is employed on the development CTC loss with patience 5. Batch size is set to 8 during finetuning for all scenarios, except for M2DS2. In the case of M2DS2 we create mixed batches of size 12, containing 4 transcribed source domain samples and 8 unlabeled target domain samples and train for $10,000$ CTC updates. For memory reasons we split the mixed batches in mini-batches of 4 and interleave them during model training. Gradients are accumulated over 3 interleaved batches. For the self-supervised objective, we create masks of maximum timestep length 10, with masking probability 0.4. We weigh the contributions of the source and target domain contrastive objectives, and bring them to the same order of magnitude as the CTC loss, by setting $\alpha = 0.01$ and $\beta = 0.02$. The convolutional feature encoder is kept frozen for all experiments. Our code is based on the huggingface [4] implementation of XLSR. For all experiments we resample the audio files to 16 kHz and downsample to single channel audio. We exclude utterances in the training set that are longer than 12 seconds. All experiments are run on a single NVIDIA RTX 3090 GPU, with mixed precision training.

For the Language model training, we create a large corpus for the Greek language using a subset of the Greek part of CC-Net [71] (approximately 11 billion tokens) and combine it with 1.5 billion tokens from the Greek version of Wikipedia and the Hellenic National Corpus (HNC) [72]. During preprocessing, we remove all punctuation and accents, deduplicate lines and convert all letters to lowercase. We will refer to this corpus as the Generic Greek Corpus (GGC). We train a 4-gram language model on GGC using KenLM [73] and prune bigrams, trigrams and four-grams with counts less than 3, 5 and 7 respectively. We incorporate the n-gram LMs at inference time using the pyctcdecode framework[5]. We use language model rescoring over a beam search decoder with 13 beams.

The evaluation metric is the Word Error Rate (WER) over the target test set. For assessing the adaptation effectiveness we also report the relative WER improvement over the unadapted baseline in appropriate scenarios, which is defined in Eq. (5). We refer to this metric as Relative Adaptation Improvement (RAI) for the rest of this paper:

$$RAI = -\frac{WER_{adapted} - WER_{unadapted}}{WER_{unadapted}} \times 100\% \quad (5)$$

The minus sign is included, so that RAI takes negative values when the adaptation fails, i.e., when $WER_{unadapted} < WER_{adapted}$.

TABLE IV
ASR performance of XLSR-53 over the three corpora for fully supervised in-domain finetuning (WER)

Dataset \ LM	No LM	4g GGC
HP	26.21	15.64
CV	29.33	9.52
LG	31.94	26.45

VI. Supervised In-Domain Training

In the first set of experiments, we explore the performance of supervised finetuning of XLSR-53 for each domain. This

[4]https://huggingface.co/docs/transformers/

[5]https://github.com/kensho-technologies/pyctcdecode

TABLE V
M2DS2 PERFORMANCE USING GREEDY DECODING FOR UDA BETWEEN HP, CV, AND LG. A → B INDICATES THAT A IS THE SOURCE DOMAIN AND B IS THE TARGET DOMAIN. (G) INDICATES GREEDY DECODING. (LM) INDICATES BEAM SEARCH WITH LM RESCORING. WE REPORT THE WER ON THE TARGET TEST SET, AS WELL AS THE RAI (%) OVER THE SO (UNADAPTED) BASELINE. WER: LOWER IS BETTER. RAI: HIGHER IS BETTER.

Method	SO (G)	CPT (G)		PSL (G)		M2DS2 (G)		SO (LM)	CPT (LM)		PSL (LM)		M2DS2 (LM)	
Setting	WER	WER	RAI	WER	RAI	WER	RAI	WER	WER	RAI	WER	RAI	WER	RAI
HP → CV	55.9	59.68	−6.8	55.3	1.2	**52.95**	**5.3**	25.26	26.44	−4.7	24.24	4.0	**18.35**	**27.4**
HP → LG	**48.65**	52.63	−8.2	57.68	−18.6	58.99	−21.3	**30.34**	32.27	−6.4	39.32	−29.6	32.58	−7.4
LG → CV	59.57	66.43	−13.4	81.90	−39.8	**51.31**	**12.4**	25.96	31.51	−21.4	52.05	−100.5	**17.30**	**33.4**
LG → HP	62.13	67.51	−8.7	71.46	−15.0	**60.09**	**3.3**	31.48	31.58	−0.3	45.36	−44.1	**31.36**	**0.4**
CV → LG	69.55	71.12	−2.3	71.34	−2.6	**63.40**	**8.8**	50.80	52.40	−3.2	48.68	4.2	**36.93**	**27.3**
CV → HP	70.72	73.83	−4.4	78.05	−10.4	**68.70**	**2.9**	52.09	52.18	−0.2	54.82	−5.2	**41.88**	**19.6**

will give an upper bound estimation for UDA performance. We finetune XLSR-53 on CV, HP and LG (separately) and perform in-domain evaluation on the respective test sets. Results are summarized in Table IV. The first row indicates the performance of greedy decoding, while in the second row we report the performance of the beam search decoder, rescored using the scores of the 4-gram GGC language model. We observe that the greedy decoding performance is under 30 WER for both HP and CV, while for LG we achieve ∼ 32 WER. This makes sense, as LG is the most diverse dataset, with respect to the included acoustic conditions. Furthermore, we observe that the incorporation of a language model results in an impressive WER reduction on CV, followed by HP and then LG. While CV includes relatively simple phrases with common vocabulary, HP and LG contain more specialized terminology.

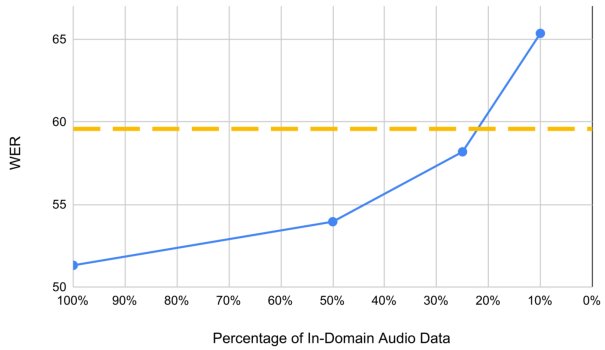


Fig. 3. Performance of M2DS2 (blue line) for the LG → CV setting, when reducing the amount of available target samples to 50%, 25%, and 10% of the original dataset (horizontal axis). SO performance is indicated with the orange line. Vertical axis: WER, Horizontal Axis: target audio percentage (100% → 0%)

VII. UNSUPERVISED DOMAIN ADAPTATION USING IN-DOMAIN AUDIO

Here, we evaluate the effectiveness of M2DS2 for UDA. We compare with three baselines:

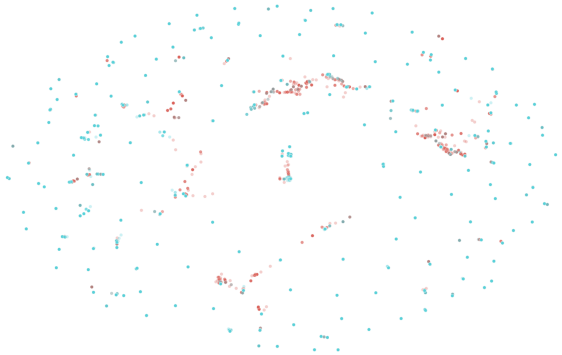
1) **Source Only Training (SO):** We perform supervised finetuning of XLSR-53 (CTC) using only the source-domain data, and run decoding on the target domain test set. No in-domain data are used for adaptation.
2) **Continual Pre-Training (CPT):** We perform a pre-training phase using the loss in Eq. (3) on the target domain train set, to create adapted versions of XLSR. Pre-training is run for 20000 steps with batch size 4. Only the audio is used, without transcriptions. The adapted checkpoints are then finetuned by use of CTC loss on the source domain transcribed data. Evaluation is performed on the target test set.
3) **Pseudolabeling (PSL):** We finetune XLSR-53 using the source domain data with CTC loss. Then we run inference on the source model, to extract silver transcriptions for the target domain training set. We use the silver transcriptions for supervised finetuning on the target domain.

In Table V we compare M2DS2 with the SO, CPT and PSL baselines for six adaptation scenarios, i.e., cross dataset evaluation between the three datasets in GREC-MD. The left half corresponds to greedy decoding, while for the right half we use the 4-gram LM trained on GGC. First, we observe the SO model performance. The SO models are the finetuned models from Table IV, evaluated in out-of-domain settings. We see that out-of-domain evaluation results in a large performance hit, e.g., while in the CV9 → CV9 in-domain setting we achieve 29.33 WER, in the CV9 → HP out-of-domain setting we get 69.55 WER. This confirms that for real-world ASR tasks, multi-domain evaluation is of essence. Second, we observe that in most adaptation scenarios both CPT and PSL fail to surpass the SO (unadapted) baseline. In the case of CPT, we hypothesize that is due to the relatively data constrained version of our setting. In the best-case scenario, we have 99 hours of available target domain audio, which is not enough to perform a discrete CPT stage. Note that most of works in the literature use ∼ 1000 hours of target audio for CPT. In the case of PSL, the poor performance is due to the quality of the silver labels created by the seed model. While the performance would improve with more elaborate approaches (e.g., confidence filtering), in challenging adaptation scenarios PSL approaches are limited by the SO model's performance. Lastly, we observe that M2DS2 is the only approach among our baselines that manages to achieve a positive RAI in most adaptation scenarios, by consistently outperforming the SO baseline by significant margins. This is exaggerated when we include a LM during inference. One exception in this pattern is the HP → LG scenario, where the SO baseline achieves the best performance. We attribute this to the fact that we performed minimal hyper-parameter tuning during model development.

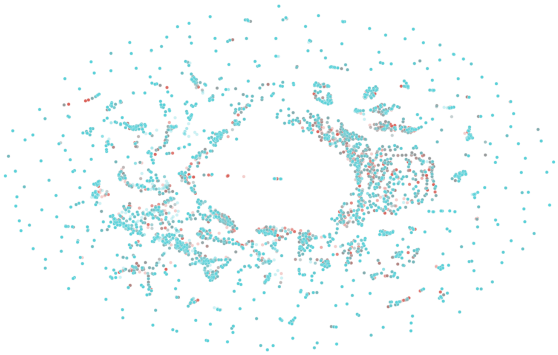
A. The sample efficiency of M2DS2

One key observation in the literature, and in our experiments is that CPT requires a large amount of un-transcribed target domain audio. This raises the question, can we leverage self-supervision for domain adaptation in data constrained settings?

In Fig. 3 we evaluate the performance of M2DS2, when we reduce the amount of target domain audio. Specifically we focus on the scenario of LG → CV. The full training corpus of CV contains 12 hours of audio. We train M2DS2 with 50%, 25% and 10% of the available samples, or 6, 3 and 1.2 hours of audio respectively, and plot the resulting WER on the target (CV) test set. In all cases, the full source (LG) training corpus is used. We observe that M2DS2 achieves lower WER than the SO baseline, even with only 3 hours of target domain audio. While CPT can suffer from catastrophic forgetting, as most multi-stage training approaches, M2DS2 avoids this issue, being a single-stage approach with a mixed task-specific and self-supervised objective. This provides a promising avenue for adaptation, when collection of in-domain recordings is expensive, or infeasible.



(a) Only target domain self-supervision



(b) Target and source domain self-supervision

Fig. 4. T-SNE scatter plots of code vectors extracted from M2DS2 without source domain self-supervision (top) and with source domain self-supervision (bottom) for LG (red) and CV (teal)

B. The importance of Multi-Domain Self-Supervision

In Section III-B we argue that it is essential to include both source and target domain data for the self-supervised objective of M2DS2. To illustrate the effect of this approach, we train two versions of M2DS2 for the LG → CV scenario. For the first version we set $\alpha = 0.01$, while for the second we set $\alpha = 0$, removing the second term of Eq. (4). We extract the code vectors for the first 100 samples of both LG and CV, and flatten them across the time steps , resulting to 60000×768 code vectors corresponding to individual timesteps. We plot these code vectors using T-SNE [74] in Fig. 4 for both models. We see that when we do not include the source domain self-supervision, the code vector space collapses in a few tight clusters, and most audio segments correspond to just a few code vectors. This is a visual clue that indicates the mode collapse problem. When we include the source domain term, we see that the that the code vector space has more structure, and coverage of the space is more complete, both for CV (target domain) and LG (source domain). Experimentally we train M2DS2 with $\alpha = 0$ for all source / target domain pairs and we find that the mode collapse is destructive for target domain performance. During our experiments we got WER in the range $80 - 99$, indicating failure to converge to acceptable solutions across all scenarios. The simple inclusion of both source and target domain self supervision stabilizes training, avoids mode collapse and leads to successful unsupervised adaptation between domains.

TABLE VI
LANGUAGE ADAPTATION OF THE M2DS2 LG → CV MODEL, USING BIASED AND AUGMENTED LMS. WE USE THE VARIANT OF THE MODEL TRAINED WITH 3 HOURS OF IN-DOMAIN AUDIO. WE VARY THE AMOUNT OF IN-DOMAIN TEXT DATA FROM 752K TOKENS TO 38K TOKENS.

	Biased LM	Augmented LM
100%	11.22	12.84
50%	15.13	15.05
25%	20.84	16.64
10%	27.75	18.47
5%	33.04	19.31
Baseline (M2DS2 + Generic LM)	20.7	

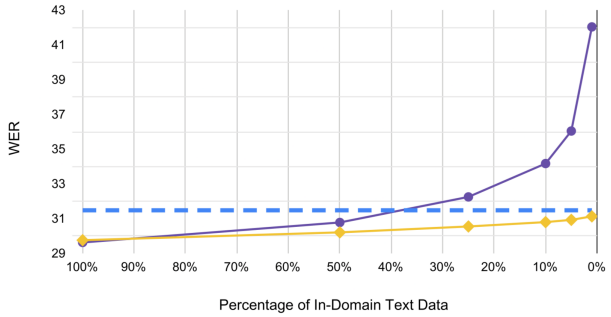


Fig. 5. Language-only adaptation for LG → HP using the SO model finetuned on LG. In-domain text data range from 11M tokens (left) to 110K tokens (right). Blue/dashed: Baseline with generic LM. Purple/circles: Biased LM. Orange/diamonds: Augmented LM.

VIII. UNSUPERVISED AND WEAKLY SUPERVISED LANGUAGE ADAPTATION

When small amounts of in-domain textual data are available, simple N-gram LM adaptation techniques can be very effective. In this brief set of experiments, we first explore the unsupervised language adaptation setting, where no in-

TABLE VII
CLOSING THE GAP BETWEEN SO TRAINING AND FULLY SUPERVISED TRAINING FOR THE LG → CV ADAPTATION SCENARIO USING M2DS2, WITH VARYING AMOUNTS OF AVAILABLE UNPAIRED IN-DOMAIN AUDIO AND TEXT. (U): UNSUPERVISED ACOUSTIC OR LANGUAGE ADAPTATION. (W): WEAKLY SUPERVISED ADAPTATION.

Method	#Audio (h)	#Tokens	LM	WER
SO (U)	-	-	N/A	59.57
M2DS2 (U)	3	-	N/A	57.31
M2DS2 (U)	12	-	N/A	51.31
SO (U)	-	-	Generic	25.96
SO (U)	-	38,632	Augmented	24.67
SO (U)	-	751,953	Augmented	20.46
M2DS2 (U)	3	-	Generic	20.7
M2DS2 (U)	12	-	Generic	17.3
M2DS2 (W)	3	38,632	Augmented	19.31
M2DS2 (W)	12	38,632	Augmented	16.29
M2DS2 (W)	3	751,953	Augmented	12.84
M2DS2 (W)	12	751,953	Augmented	10.61
Supervised	12	751,953	Generic	9.52
Supervised	12	751,953	Augmented	7.94

domain audio is used, and then we relax the problem to the weakly supervised setting, where M2DS2 is combined with the adapted N-Gram LMs. These settings are described in Sections II-A2 and II-A3 respectively. We explore two approaches for LM adaptation: biased LMs, and in-domain data augmentation. To create biased LMs, we train a 4-gram LM on the available in-domain data. Then we replace the generic LM trained on GGC. For LM data augmentation we follow a perplexity filtering approach similar to [71]. We first train a biased LM using available target domain text, and then use it to calculate the perplexity of each line in the GGC corpus. We keep the 10% of the lines with the lowest perplexity. Then we train a 4-gram LM on the augmented "in-domain" corpus and use it for inference.

Fig. 5 shows the performance of the SO LG → HP model with biased and augmented LMs, as we reduce the amount of available in-domain text data from 100% to 1% of the in-domain transcriptions (11B tokens to 110K tokens respectively). As a baseline we include the LG → HP SO model in combination with the generic LM trained on GGC. We observe that the use of biased LMs can lead to successful adaptation, when an adequate amount of in-domain text data is available. On the other hand the LM augmentation approach results to successful augmentation, even with very small amounts of in-domain text.

In Table VI we see the results of LM adaptation, combined with the M2DS2 LG → CV model. To demonstrate the sample efficiency of the approach, we use the variant that was trained using only 25% of the target domain audio (3 hours). We compare with M2DS2 combined with the 4-gram GGC LM for inference. We draw similar conclusions, i.e., use of biased LMs performs well for sufficient text data. When we use augmented LMs we can leverage very small amounts of in-domain text.

IX. DISCUSSION & CONCLUSIONS

In this work, we have explored Unsupervised and Weakly Supervised Domain Adaptation of ASR systems in the context of an under-resourced language, i.e., Greek. We focus on domain adaptation through in-domain self-supervision for XLSR-53, a state-of-the-art multilingual ASR model. Specifically, we adopt a mixed task and self-supervised objective, inspired from NLP, and show that using only in-domain self-supervision can lead to mode collapse of the representations created by the contrastive loss of XLSR-53. Therefore, we propose the use of mixed task and multi-domain self-supervision, M2DS2, where the contrastive loss leverages both the source and target domain audio data. For evaluation we create and release HParl, the largest to-date public corpus of transcribed Greek speech (120 hours), collected from the Greek Parliamentary Proceedings. HParl is combined with two other popular Greek speech corpora, i.e., Logotypografia and CommonVoice, for multi-domain evaluation.

In our experiments, we find that while most UDA baselines fail in our low-resource setting, the proposed mixed task and multi-domain self-supervised finetuning strategy yields significant improvements for the majority of adaptation scenarios. Furthermore, we focus our ablations on showcasing the sample efficiency of the proposed finetuning strategy, and demonstrating the necessity of including both source and target domain data for self-supervision. Finally, we show that M2DS2 can be combined with simple language model adaptation techniques in a relaxed weakly supervised setting, where we achieve significant performance improvements with a few hours of in-domain audio and a small, unpaired in-domain text corpus.

More concretely, in Table VII we present a summary of the discussed unsupervised and weakly supervised adaptation combinations, for different amounts of available in-domain audio and text. Note that for the weakly supervised scenarios, the in-domain audio and text are unpaired. We see, that when no in-domain data are available, including an n-gram LM trained on large corpora is recommended. Furthermore, when in-domain audio is available, following a mixed multi-domain finetuning strategy using M2DS2 can yield significant WER reductions, even for a few hours of audio. When small amounts of in-domain text is available, using a corpus augmentation strategy, e.g., perplexity filtering, can produce adapted LMs and yield small improvements to the final WER. In the case of sufficient amounts of unpaired in-domain text and audio, independent adaptation of XLSR-53 using the audio data and the n-gram LM using the text data can yield comparable performance with a fully supervised finetuning pipeline.

X. FUTURE WORK

In the future we plan to explore the effectiveness of the proposed adaptation strategy for other languages, and different adaptation settings, e.g., accent or cross-lingual adaptation. Of special interest is the investigation of the effectiveness of our approach for endagered languages, e.g., Pomak. Furthermore, we plan to explore the combination of in-domain self-supervision, when combined with other popular UDA techniques, e.g., teacher student models, adversarial learning, and data augmentation approaches. On the language adaptation side, we plan to explore multi-resolution learning, which has

shown promise for ASR [75], and investigate more elaborate end-to-end weakly supervised adaptation methods. Finally, we plan to expand our study in a multimodal setting, where both audio and video are available, e.g., lip reading.

REFERENCES

[1] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan, "Learning transferable features with deep adaptation networks," in *Proc. ICML*. PMLR, 2015, pp. 97–105.

[2] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky, "Domain-adversarial training of neural networks," *J. Mach. Learn. Res.*, vol. 17, no. 1, pp. 2096–2030, jan 2016.

[3] Peter Bell, Joachim Fainberg, Ondrej Klejch, Jinyu Li, Steve Renals, and Pawel Swietojanski, "Adaptation algorithms for neural network-based speech recognition: An overview," *IEEE Open Journal of Signal Processing*, vol. 2, pp. 33–66, 2021.

[4] Jinyu Li, Michael L. Seltzer, Xi Wang, Rui Zhao, and Yifan Gong, "Large-scale domain adaptation via teacher-student learning," in *Proc Interspeech*, 2017.

[5] Vimal Manohar, Pegah Ghahremani, Daniel Povey, and Sanjeev Khudanpur, "A teacher-student learning approach for unsupervised domain adaptation of sequence-trained asr models," in *Proc. Spoken Language Technology Workshop (SLT)*, 2018.

[6] Yusuke Shinohara, "Adversarial Multi-Task Learning of Deep Neural Networks for Robust Speech Recognition," in *Proc. Interspeech 2016*, 2016, pp. 2369–2372.

[7] Zhong Meng, Jinyu Li, Zhuo Chen, Yang Zhao, Vadim Mazalov, Yifan Gong, and Biing-Hwang Juang, "Speaker-invariant training via adversarial learning," in *Proc. ICASSP*. 2018, IEEE.

[8] Sining Sun, Ching-Feng Yeh, Mei-Yuh Hwang, Mari Ostendorf, and Lei Xie, "Domain adversarial training for accented speech recognition," in *Proc. ICASSP*, 2018, pp. 4854–4858.

[9] Anoop C S, Prathosh A P, and A G Ramakrishnan, "Unsupervised domain adaptation schemes for building asr in low-resource languages," in *Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2021, pp. 342–349.

[10] Taichi Asami et al., "Domain adaptation of dnn acoustic models using knowledge distillation," in *Proc. ICASSP*. IEEE, 2017, pp. 5185–5189.

[11] Karol Nowakowski, Michal Ptaszynski, Kyoko Murasaki, and Jagna Nieuważny, "Adapting multilingual speech representation model for a new, underresourced language through multilingual fine-tuning and continued pretraining," *Information Processing & Management*, vol. 60, no. 2, pp. 103148, 2023.

[12] Sadaoki Furui, "A training procedure for isolated word recognition systems," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 2, pp. 129–136, 1980.

[13] Yajie Miao, Hao Zhang, and Florian Metze, "Towards speaker adaptive training of deep neural network acoustic models," in *Proc. Interspeech*, 2014, pp. 2189–2193.

[14] Sree HK Parthasarathi e al., "fmllr based feature-space speaker adaptation of dnn acoustic models," in *Proc. Interspeech*, 2015, pp. 3630–3634.

[15] Vishwa Gupta, Patrick Kenny, Pierre Ouellet, and Themos Stafylakis, "I-vector-based speaker adaptation of deep neural networks for french broadcast audio transcription," in *Proc. ICASSP*, 2014.

[16] Hans-Günter Hirsch and David Pearce, "The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *ASR2000-Automatic speech recognition: challenges for the new Millenium ISCA tutorial and research workshop (ITRW)*, 2000.

[17] Yanmin Qian, Tian Tan, and Dong Yu, "An investigation into using parallel data for far-field speech recognition," in *Proc. ICASSP*, 2016.

[18] William Chan, Navdeep Jaitly, Quoc Le, and Oriol Vinyals, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," in *Proc. ICASSP*, 2016, pp. 4960–4964.

[19] Alex Graves, "Sequence transduction with recurrent neural networks," *CoRR*, vol. abs/1211.3711, 2012.

[20] 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.

[21] Alexis Conneau et al., "Unsupervised Cross-Lingual Representation Learning for Speech Recognition," in *Proc. Interspeech*, 2021, pp. 2426–2430.

[22] Pavel Denisov, Ngoc Thang Vu, and Marc Ferras Font, "Unsupervised domain adaptation by adversarial learning for robust speech recognition," in *Speech Communication*, 2018, pp. 1–5.

[23] Dongseong Hwang, Ananya Misra, Zhouyuan Huo, Nikhil Siddhartha, Shefali Garg, David Qiu, Khe Chai Sim, Trevor Strohman, Françoise Beaufays, and Yanzhang He, "Large-scale asr domain adaptation using self- and semi-supervised learning," in *NeurIPS*, 2022, pp. 6627–6631.

[24] Wei-Ning Hsu, Anuroop Sriram, Alexei Baevski, Tatiana Likhomanenko, Qiantong Xu, Vineel Pratap, Jacob Kahn, Ann Lee, Ronan Collobert, Gabriel Synnaeve, and Michael Auli, "Robust wav2vec 2.0: Analyzing Domain Shift in Self-Supervised Pre-Training," in *Proc. Interspeech*, 2021, pp. 721–725.

[25] Sameer Khurana, Niko Moritz, Takaaki Hori, and Jonathan Le Roux, "Unsupervised domain adaptation for speech recognition via uncertainty driven self-training," *Proc. ICASSP*, pp. 6553–6557, 2020.

[26] Sankaran Panchapagesan, Daniel S. Park, Chung-Cheng Chiu, Yuan Shangguan, Qiao Liang, and Alexander Gruenstein, "Efficient knowledge distillation for rnn-transducer models," in *Proc ICASSP*, 2021, pp. 5639–5643.

[27] Anmol Gulati et al, "Conformer: Convolution-augmented transformer for speech recognition," *Proc. Interspeech*, pp. 5036–5040, 2020.

[28] Hasim Sak, Andrew W. Senior, and Françoise Beaufays, "Long short-term memory recurrent neural network architectures for large scale acoustic modeling," in *Proc. Interspeech*, 2014, pp. 338–342.

[29] Taichi Asami, Ryo Masumura, Yoshikazu Yamaguchi, Hirokazu Masataki, and Yushi Aono, "Domain adaptation of dnn acoustic models using knowledge distillation," in *Proc. ICASSP*, 2017, pp. 5185–5189.

[30] Takuya Yoshioka, Nobutaka Ito, Marc Delcroix, Atsunori Ogawa, Keisuke Kinoshita, Masakiyo Fujimoto, Chengzhu Yu, Wojciech J. Fabian, Miquel Espi, Takuya Higuchi, Shoko Araki, and Tomohiro Nakatani, "The ntt chime-3 system: Advances in speech enhancement and recognition for mobile multi-microphone devices," in *Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2015, pp. 436–443.

[31] Bhuvana Ramabhadran, Brian Farris, Isabel Leal, Manasa Prasad, Neeraj Gaur, Parisa Haghani, Pedro Jose Moreno Mengibar, and Yun Zhu, "Self-adaptive distillation for multilingual speech recognition: Leveraging student independence," in *Interspeech 2021*, 2021.

[32] Yanzhang He, Tara N. Sainath, Rohit Prabhavalkar, Ian McGraw, Raziel Alvarez, Ding Zhao, David Rybach, Anjuli Kannan, Yonghui Wu, Ruoming Pang, Qiao Liang, Deepti Bhatia, Yuan Shangguan, Bo Li, Golan Pundak, Khe Chai Sim, Tom Bagby, Shuo-yiin Chang, Kanishka Rao, and Alexander Gruenstein, "Streaming end-to-end speech recognition for mobile devices," in *Proc. ICASSP*, 2019, pp. 6381–6385.

[33] Dmitriy Serdyuk, Kartik Audhkhasi, Philemon Brakel, Bhuvana Ramabhadran, Samuel Thomas, and Yoshua Bengio, "Invariant representations for noisy speech recognition," *CoRR*, 2016.

[34] Sining Sun, Binbin Zhang, Lei Xie, and Yanning Zhang, "An unsupervised deep domain adaptation approach for robust speech recognition," *Neurocomputing*, vol. 257, pp. 79–87, 2017, Machine Learning and Signal Processing for Big Multimedia Analysis.

[35] Vijayaditya Peddinti, Daniel Povey, and Sanjeev Khudanpur, "A time delay neural network architecture for efficient modeling of long temporal contexts," in *Proc. Interspeech*, 2015, pp. 3214–3218.

[36] Daniel Povey et al., "The kaldi speech recognition toolkit," in *Proc. ASRU Workshop*. 2011, IEEE.

[37] Seyedmahdad Mirsamadi and John H.L. Hansen, "Multi-domain adversarial training of neural network acoustic models for distant speech recognition," *Speech Communication*, vol. 106, pp. 21–30, 2019.

[38] Jan K Chorowski, Dzmitry Bahdanau, Dmitriy Serdyuk, Kyunghyun Cho, and Yoshua Bengio, "Attention-based models for speech recognition," *Advances in neural information processing systems*, vol. 28, 2015.

[39] Hu Hu, Xuesong Yang, Zeynab Raeesy, Jinxi Guo, Gokce Keskin, Harish Arsikere, Ariya Rastrow, Andreas Stolcke, and Roland Maas, "redat: Accent-invariant representation for end-to-end asr by domain adversarial training with relabeling," in *Proc. ICASSP*, 2021.

[40] Aditay Tripathi, Aanchan Mohan, Saket Anand, and Maneesh Singh, "Adversarial learning of raw speech features for domain invariant speech recognition," in *Proc. ICASSP*, 2018, pp. 5959–5963.

[41] Konstantinos Bousmalis, George Trigeorgis, Nathan Silberman, Dilip Krishnan, and Dumitru Erhan, "Domain separation networks," in *Proc. NIPS*, Red Hook, NY, USA, 2016, NIPS'16, p. 343–351, Curran Associates Inc.

[42] Han Zhu, Gaofeng Cheng, Jindong Wang, Wenxin Hou, Pengyuan Zhang, and Yonghong Yan, "Boosting cross-domain speech recognition with self-supervision," *arXiv preprint arXiv:2206.09783*, 2022.

[43] Jounghee Kim and Pilsung Kang, "K-Wav2vec 2.0: Automatic Speech Recognition based on Joint Decoding of Graphemes and Syllables," in *Proc. Interspeech*, 2022, pp. 4945–4949.

[44] Mitchell DeHaven and Jayadev Billa, "Improving low-resource speech recognition with pretrained speech models: Continued pretraining vs. semi-supervised training," *arXiv preprint arXiv:2207.00659*, 2022.

[45] Constantinos Karouzos, Georgios Paraskevopoulos, and Alexandros Potamianos, "UDALM: Unsupervised domain adaptation through language modeling," in *Proc. Association for Computational Linguistics: Human Language Technologies*, Online, June 2021, pp. 2579–2590, Association for Computational Linguistics.

[46] Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber, "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks," in *Proc. of the 23rd Int. Conf. on Machine Learning*, New York, NY, USA, 2006, Proc. ICML, p. 369–376, Association for Computing Machinery.

[47] H. Scudder, "Probability of error of some adaptive pattern-recognition machines," *IEEE Transactions on Information Theory*, vol. 11, no. 3, pp. 363–371, 1965.

[48] David Yarowsky, "Unsupervised word sense disambiguation rivaling supervised methods," in *Annu. Meeting of the Association for Computational Linguistics*, 1995.

[49] Ellen Riloff and Janyce Wiebe, "Learning extraction patterns for subjective expressions," in *Conf. Empirical Methods in Natural Language Processing*, 2003, pp. 105–112.

[50] Dongseong Hwang, Khe Chai Sim, Yu Zhang, and Trevor Strohman, "Comparison of soft and hard target rnn-t distillation for large-scale asr," 2022.

[51] Jacob Kahn, Ann Lee, and Awni Hannun, "Self-training for end-to-end speech recognition," in *Proc. ICASSP*, 2020, pp. 7084–7088.

[52] Daniel S. Park et al, "Improved noisy student training for automatic speech recognition," in *Proc. Interspeech*, 2020.

[53] Yu Zhang, James Qin, Daniel S. Park, Wei Han, Chung-Cheng Chiu, Ruoming Pang, Quoc V. Le, and Yonghui Wu, "Pushing the limits of semi-supervised learning for automatic speech recognition," 2020.

[54] Daniel S. Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D. Cubuk, and Quoc V. Le, "SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition," in *Proc. Interspeech*, 2019, pp. 2613–2617.

[55] Yaroslav Ganin and Victor Lempitsky, "Unsupervised domain adaptation by backpropagation," in *Proc. ICML*. 2015, ICML'15, p. 1180–1189, JMLR.org.

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

[57] Siu-Kei Au Yeung and Man-Hung Siu, "Improved performance of aurora 4 using htk and unsupervised mllr adaptation," in *Conf. Spoken Language Processing*, 2004.

[58] Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith, "Don't stop pretraining: Adapt language models to domains and tasks," in *Proc. of the 58th Annu. Meeting of the Association for Computational Linguistics*, Online, July 2020, pp. 8342–8360, Association for Computational Linguistics.

[59] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," *CoRR*, vol. abs/1810.04805, 2018.

[60] Herve Jegou, Matthijs Douze, and Cordelia Schmid, "Product quantization for nearest neighbor search," *IEEE transactions on pattern analysis and machine intelligence*, vol. 33, no. 1, pp. 117–128, 2010.

[61] Eric Jang, Shixiang Gu, and Ben Poole, "Categorical reparametrization with gumbel-softmax," in *Proc. ICLR*, Apr. 2017.

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

[63] Aimilios Chalamandaris et al., "All greek to me! an automatic greeklish to greek transliteration system," in *Proc. LREC*, 2006.

[64] Carsten Meyer and Hauke Schramm, "Boosting hmm acoustic models in large vocabulary speech recognition," *Speech Communication*, vol. 48, no. 5, pp. 532–548, 2006.

[65] Vimal Manohar, Daniel Povey, and Sanjeev Khudanpur, "Jhu kaldi system for arabic mgb-3 asr challenge using diarization, audio-transcript alignment and transfer learning," in *Proc. ASRU Workshop*, 2017, pp. 346–352.

[66] Vassilios Digalakis et al., "Large vocabulary continuous speech recognition in greek: corpus and an automatic dictation system," in *Proc. Eurospeech*, 2003, pp. 1565–1568.

[67] T.F. Smith and M.S. Waterman, "Identification of common molecular subsequences," *Journal of Molecular Biology*, vol. 147, no. 1, pp. 195–197, 1981.

[68] Rosana Ardila et al., "Common voice: A massively-multilingual speech corpus," in *Proc. LREC*, 2020, pp. 4218–4222.

[69] Ilya Loshchilov and Frank Hutter, "Decoupled weight decay regularization," in *Proc. ICLR*, 2018.

[70] Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *Proc. ICML*, 2006, pp. 369–376.

[71] Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Édouard Grave, "Ccnet: Extracting high quality monolingual datasets from web crawl data," in *Proc. of the 12th Language Resources and Evaluation Conf.*, 2020, pp. 4003–4012.

[72] Nick Hatzigeorgiu, Maria Gavrilidou, Stelios Piperidis, George Carayannis, Anastasia Papakostopoulou, Athanassia Spiliotopoulou, Anna Vacalopoulou, Penny Labropoulou, Elena Mantzari, Harris Papageorgiou, et al., "Design and implementation of the online ilsp greek corpus.," in *LREC*, 2000.

[73] Kenneth Heafield, "Kenlm: Faster and smaller language model queries," in *Proc. of the 6th workshop on statistical machine translation*, 2011, pp. 187–197.

[74] Laurens Van der Maaten and Geoffrey Hinton, "Visualizing data using t-sne.," *Journal of machine learning research*, vol. 9, no. 11, 2008.

[75] Georgios Paraskevopoulos, Srinivas Parthasarathy, Aparna Khare, and Shiva Sundaram, "Multimodal and multiresolution speech recognition with transformers," in *Proc. of the 58th Annu. Meeting of the Association for Computational Linguistics*, 2020, pp. 2381–2387.