Domain Expansion for End-to-End Speech Recognition: Applications for Accent/Dialect Speech

Shahram Ghorbani, *Student Member, IEEE*, and John H.L. Hansen, *Fellow, IEEE*

***Abstract*—Training Automatic Speech Recognition (ASR) systems with sequentially incoming data from alternate domains is an essential milestone in order to reach human intelligibility level in speech recognition. The main challenge of sequential learning is that current adaptation techniques result in significant performance degradation for previously-seen domains. To mitigate the catastrophic forgetting problem, this study proposes effective domain expansion techniques for two scenarios: 1) where only new domain data is available, and 2) where both prior and new domain data are available. We examine the efficacy of the approaches through experiments on adapting a model trained with native English to different English accents. For the first scenario, we study several existing and proposed regularization-based approaches to mitigate performance loss of initial data. The experiments demonstrate the superior performance of our proposed Soft KL-Divergence (SKLD)-Model Averaging (MA) approach. In this approach, SKLD first alleviates the forgetting problem during adaptation; next, MA makes the final efficient compromise between the two domains by averaging parameters of the initial and adapted models. For the second scenario, we explore several rehearsal-based approaches, which leverage initial data to maintain the original model performance. We propose Gradient Averaging (GA) as well as an approach which operates by averaging gradients computed for both initial and new domains. Experiments demonstrate that GA outperforms retraining and specifically designed continual learning approaches, such as Averaged Gradient Episodic Memory (AGEM). Moreover, GA significantly improves computational costs over the complete retraining approach.**

***Index Terms*—Accented speech, continual learning, domain expansion, end-to-end systems, model adaptation, speech recognition.**

I. Introduction

CURRENT state-of-the-art machine learning-based Automatic Speech Recognition (ASR) systems have advanced to near human performance in several evaluation settings [1], [2], [3]. However, there remain technological barriers in order to achieve flexible solutions and user satisfaction under naturalistic field scenarios. A major issue for current ASR systems is sustained performance for alternate accents/dialects of that language, which precludes the use of these systems by/for specific populations [4], [5]. For example, many ASR systems aim to achieve the best performance for English, since it is the most spoken second language worldwide. However, English as a language comes with many alternate accents/dialects; native English-speaking countries (e.g., U.K., US, Canada, Australia) have developed a diverse and expansive set of distinct dialects. In addition, increasing worldwide communication has expanded bilingual speakers with English as a second language, which due to the impact of their first L1 language, speak English with varying degrees of L1 dependent accent. The diverse number of possible accents/dialects pose a major challenge for robust ASR since DNN-based ASR systems are known to generalize poorly to unseen domains [6], [7], [8].

A straightforward approach for an advanced multi-domain ASR system accumulates the training data, including initial and unseen domains, then re-trains the entire model using various multi-condition training approaches [9], [10], [11]. Nonetheless, for many realistic settings, the re-training approach is not a feasible solution. First, as training data size grows, storing and re-training on the accumulated large-scale dataset becomes practically impossible. Second, data for new domains (e.g., data for accented English) is usually smaller than initial domains (e.g., data for native English). The resulting domain-imbalanced dataset poses added difficulty for multi-condition training approaches. Moreover, accessing the initial train dataset is not always feasible, for example, saving user audio due to privacy or safety concerns. In this scenario, an alternative approach is to leverage the pre-trained model and new datasets to perform model adaptation. However, adaptation techniques would lead to *catastrophic forgetting:* where previously learned information is lost by learning the new domain information. As illustrated in Fig. 1, domain expansion approaches try to address these problems by building an ASR system that not only performs well for new domains, it retains performance for previously seen domains.

Domain Expansion – In our domain expansion scenario, we have a model trained on an initial domain, a dataset for an unseen domain, and in some scenarios, the initial dataset as well. The goal of a domain expansion approach is to adapt the pre-trained model such that it performs well for both initial and unseen domains. The main challenges for domain expansion approaches are maintaining the initial model's functionality (input-output mapping) and simultaneously reducing the computational cost. Past continual learning approaches have been proposed to deal

Manuscript received 28 January 2022; revised 23 October 2022; accepted 21 December 2022. Date of publication 30 December 2022; date of current version 13 January 2023. This work was supported in part by NSF EAGER Project under Grant 2140415 and in part by the University of Texas at Dallas from the Distinguished University Chair in Telecommunications Engineering held by John H.L. Hansen. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Lei Xie. *(Corresponding author: John H. L. Hansen.)*

The authors are with the Center for Robust Speech Systems, University of Texas at Dallas, Richardson, TX 75080 USA (e-mail: shahram.ghorbani@utdallas.edu; john.hansen@utdallas.edu).

Digital Object Identifier 10.1109/TASLP.2022.3233238

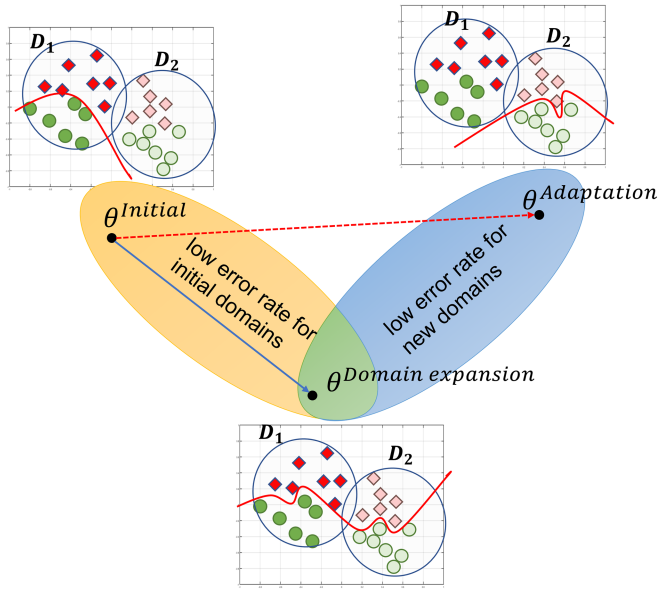

Fig. 1. Domain expansion vs. domain adaptation. Domain expansion enables the model to adapt to new domains while maintaining its performance for initial data.

with these problems, which can be divided into three categories: architectural, regularization, and rehearsal strategies.

Architectural strategies – These methods leverage the architectures of neural networks to mitigate the forgetting problem. Progressive neural network (PNN) [12] is a popular architectural strategy which leverages a previously frozen trained network to obtain an intermediate representation used as inputs into a new smaller network. A major drawback for a PNN solution is that if a large number of domains is required to be included, the size of the resulting model increases linearly with the number of domains [12]. In a recent study, Sadhu et al. [13] proposed a domain expansion approach by leveraging multiple domain-specific models trained sequentially for a DNN-HMM setting. However, this approach requires storing and using domain-specific models; therefore, the computational cost of this approach increases linearly as well.

Regularization strategies – Regularization approaches try to alleviate forgetting the previously trained information by imposing additional constraints on updating parameters. A straightforward approach is to employ *weight constrained adaptation (WCA)* which penalizes deviation of the model parameters from the initial model implemented by imposing an L_2 distance between the initial and adapted weights [14]. Kirkpatrick et al. [15] suggested that all model weights are not equally important to maintain performance for the initial domains. Therefore, to advance WCA, they introduced *elastic weight consolidation (EWC)* approach, which selectively slows down the training for weights that are important for previously seen domains. Another approach for computing the importance score of model parameters is Synaptic Intelligence (SI) [16]. SI computes the importance of each weight based on its contribution to the objective loss over the entire training steps. Alternative approaches constrain the model's outputs to maintain functionality (input-output mapping) of the initial model [17], [18], [19]. *Learning without forgetting (LWF)*, as an effective approach of this class, proposed to sustain output stability of the previously seen tasks in order to learn a sequence of tasks while maintaining performance for previously seen ones [18]. In similar work, Jung et al. [19] investigated the domain expansion problem for image classification tasks. They leveraged an L_2 distance between the final hidden representations of initial and adapted models to alleviate forgetting of the initial model. In our previous work [17], we examined the efficacy of WCA, EWC, Soft KL-Divergence (SKLD), and our proposed hybrid SKLD-EWC approach for an advanced domain expansion solution in a DNN-HMM ASR setting. We demonstrated the effectiveness of the proposed SKLD-EWC approach in adapting a model trained with native English to unseen accented datasets while sustaining initial performance. In a similar work, the EWC and LWF approaches were leveraged to train a multi-dialect acoustic model in a sequential transfer learning framework [20].

Rehearsal strategies – These approaches store (a subset of) the prior training data and periodically replay them for future training. A re-training strategy can alleviate the forgetting problem, but processing the entire initial data can be resource-intensive. To alleviate the memory problem of full-rehearsal strategies, Hayes et al. proposed EXSTREAM, a new partitioning-based approach to select representative samples of the initial data [21]. Generative models have also been leveraged to alleviate the resource costs of rehearsal approaches [22]. To advance the efficiency of rehearsal-based approaches, some studies investigated the relative directions between gradient vectors computed for new and initial data to solve the problem of learning the new data without interfering in the previously learned information [23], [24], [25]. In [24], Averaged Gradient Episodic Memory (AGEM) was introduced, which alleviates the interference problem by projecting gradients from new datasets onto a subspace in which they have no information interference with gradients from initial datasets. Mehrdad et al. proposed Orthogonal Gradient Descent (OGD), which operates similarly to AGEM; however, instead of storing and reusing the initial dataset samples, OGD stores the gradients from the initial dataset [25]. With the recent progress in the meta-learning area [26], the potential of leveraging this framework to advance continual learning is investigated in several studies [26], [27]. [27] advances Model-Agnostic Meta-Learning (MAML) framework, introduced in [28], by leveraging a replay buffer and optimizing a meta-objective that alleviates forgetting.

This current study proposes novel advanced regularization-based and rehearsal-based approaches to address the domain expansion problem for an End-to-End ASR model. An overview of domain expansion approaches developed in this study is shown in Fig. 2. We examine the efficacy of the approaches through experiments on adapting a model trained with native English to two English accents/dialects: Australian English and Indian accent. For regularization-based approaches, we investigate the efficacy of existing weight-constrained approaches WCA, EWC, and SI in alleviating the forgetting problem in domain expansion. Furthermore, we examine performance of SKLD and hybrid SKLD-EWC introduced in our previous study [17] for an End-to-End ASR model. SKLD mitigates the forgetting problem by imposing a constraint that penalizes

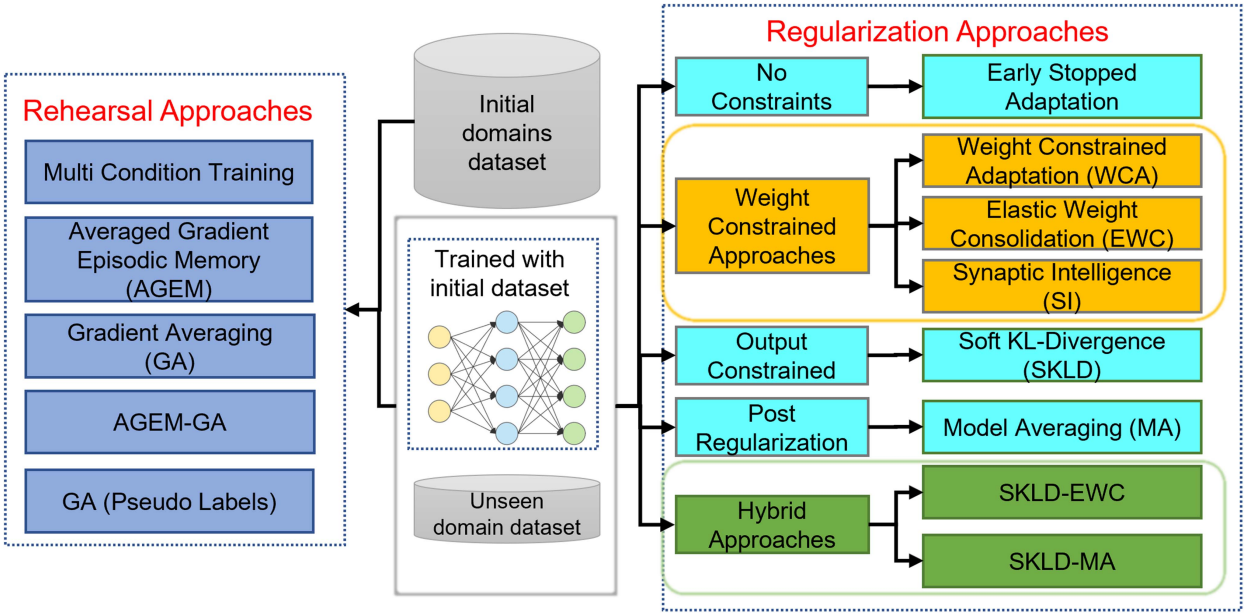

Fig. 2. Overview of regularization and rehearsal based domain expansion approaches developed in this study.

the KL-divergence between the initial model's output and the adapted model's output. We also propose leveraging model averaging (MA) as a domain expansion solution, which operates by post-regularizing an adapted model. MA operates by first adapting the initial model to reach an optimum loss for the new data. Next, it averages the parameters of the initial and adapted models to ensure compromise between the initial and new domains. Alternatively, to improve the adaptation step, we propose a hybrid SKLD-MA approach; where first, SKLD regularizes the adaption and alleviates forgetting initial data. Next, MA makes the final efficient trade-off between the two datasets. Our experimental evaluation will demonstrate the superior performance of MA and SKLD-MA compared to other regularization-based approaches.

For rehearsal-based approaches, we first investigate the efficacy of AGEM and multi-condition training. Multi-condition training offers reasonable performance for domain expansion, but since it is highly resource-intensive, this approach would be impractical for many large-scale ASR scenarios. Our experiments demonstrate that the AGEM approach performs poorly compared to other rehearsal-based approaches. By analyzing the results, we show that the AGEM's poor performance is mainly due to leveraging the unreliable angle between gradients of mini-batches and, consequently, forgetting the initial data. To improve AGEM, we propose to increase the contribution of the initial gradients regardless of the angle value. This resulting approach will be shown to address the AGEM drawback and improve domain expansion performance. We also propose gradient averaging (GA), which operates by starting from the initial model; then, in each adaptation step, it averages the gradient vectors of initial and new data with a flexible weight. The experiments will demonstrate superior performance of GA compared to other domain expansion approaches, including multi-condition training. The GA approach narrows the performance gap between a domain expansion solution to domain-specific models to 1.6%–2.12%. Finally, we study the impact of the initial data buffer size on overall domain expansion performance.

This study extends our preliminary research [17] and investigates the domain expansion problem comprehensively. The core contributions of this study are summarized as follows:

1) Propose an MA-based approach and SKLD-MA as novel regularization-based domain expansion approaches; the proposed approaches are shown to outperform existing approaches in an ASR setting.
2) Adapt advanced existing continual learning approaches to an ASR setting; evaluate and analyze their efficacy for domain expansion.
3) Propose GA as an advanced rehearsal-based approach that outperforms existing approaches, including multi-condition training. A systematic investigation demonstrates why existing approaches such as AGEM perform poorly for ASR.
4) Investigate the impact of initial data buffer size on domain expansion performance, providing new insights into the problem.

The remainder of this paper is organized as follows. Details of regularization-based and rehearsal-based approaches are presented in Sections II & III, respectively. The description of the end-to-end model used in this study, as well as training and evaluation settings, are provided in Section IV. Section V details and analyzes experimental results. Finally, conclusions are presented in Section VI.

II. Regularization-Based Domain Expansion Approaches

This section presents details of existing and proposed regularization-based domain expansion approaches investigated

in this study. These approaches seek to advance domain expansion performance for scenarios where the initial training dataset is not available. However, we assume that the initial development set is available to tune the hyperparameters of models and algorithms.

Problem Setup – In the domain expansion task, an initial model $\mathcal{M}^{init}$, trained on an initial domain $\mathcal{D}^{init}$ is given along with a dataset for an unseen domain $\mathcal{D}^{n}$. The goal is to find a new model $\mathcal{M}^{n}$ that performs well for both $\mathcal{D}^{init}$ and $\mathcal{D}^{n}$.

A. Weight Constrained Adaptation (WCA)

WCA was first proposed in [14] for regularized adaptation of discriminative classifiers. In a related study [18], WCA was employed for continual learning in a sequence of disjoint tasks. This technique addresses the domain expansion problem by finding a model adapted to the new domain, $\mathcal{D}^{n}$, which is also close to the initial model, $\mathcal{M}^{init}$.

For a neural network architecture, there are many configurations of model parameters with comparable performance [15]. Therefore, we also assume that many configurations of the model can efficiently model our new domain $\mathcal{D}^{n}$. Among such configurations, an effective domain expansion solution is one that stands closer to the initial model $\mathcal{M}^{init}$. WCA leverages the Euclidean distance between the learnable parameters of $\mathcal{D}^{init}$ and $\mathcal{D}^{n}$ to measure the similarity between these models. This idea can be implemented by imposing an additional L_2 constraint which penalizes parameter changes as follows:

$$J_{WCA}(\theta^n) = J_{CTC}(\theta^n) + \frac{\lambda_w}{2}||\theta^n - \theta^{init}||_2, \qquad (1)$$

where θ^{init} and θ^n are the learnable parameters of $\mathcal{M}^{init}$ and $\mathcal{M}^{n}$, respectively; $J_{CTC}(\theta^n)$ is the main optimization loss (i.e., Connectionist Temporal Classification (CTC) loss function [29]); $||.||_2$ is the L_2 norm; and λ_w is a regularization hyper-parameter that determines the felixibilty of the parameters to learn the desired new domain.

B. Elastic Weight Consolidation (EWC)

The WCA technique considers all weights within the model equally. However, model weights might not be equally important to maintain initial model performance. Therefore, WCA can potentially result in a suboptimal solution for maintaining model performance for the initial domain $\mathcal{D}^{init}$ while learning the new domain $\mathcal{D}^{n}$. EWC [15] was proposed to compute the importance of each weight and leverage the weight importance for an advanced domain expansion solution that can better balance initial and desired data domains.

Intuitively, after a DNN model is trained for a sufficient number of iterations, the model converges to a local minimum point of the optimization landscape. At this point, one can estimate the sensitivity of the loss function w.r.t. the i-th learnable weight, θ_i^n, by studying the curvature of that loss function along the direction of θ_i^n changes. A weight with high curvature means small changes to that weight would significantly change the overall loss function. EWC leverages the curvature information in order to preserve performance of the network for previously seen domains, while penalizing modifications to the parameters with high curvature. Alternatively, parameters with low curvature values are proper choices to be tuned with new data without significantly affecting the model performance for the initial data domain.

The curvature of the loss function is equivalent to the diagonal of the Fisher information matrix F [15]. EWC therefore incorporates the importance of the learnable weights (curvature of the loss function w.r.t. the weights) by imposing a constraint on the adaptation process. Following the previous WCA approach, EWC employs a weighted L_2 norm to restrict changing of each learnable weight proportional to its importance:

$$J_{EWC}(\theta^n) = J_{CTC}(\theta^n) + \frac{\lambda_e}{2}\sum_i diag\{F\}_i(\theta_i^n - \theta_i^{init})^2, \qquad (2)$$

where $diag\{F\}_i$ is the i-th element of the diagonal of the Fisher information matrix F (representing the importance of the i-th learnable weight); θ_i^n and θ_i^{init} are the i-th weight of the new and initial models, respectively; and summation is taken over all learnable weights of the network. Here, $diag\{F\}$ can be easily calculated using the variance of the first order derivatives of the loss function w.r.t. the learnable weights (i.e., $Var\{\partial J(\theta)/\partial\theta_i\}$) [15], [30].

C. Synaptic Intelligence (SI)

EWC computes the importance of each weight based on the curvature of loss function around that weight regardless of its past changes and contributions to address the task objective. However, another class of algorithms, namely synaptic intelligence introduced in [16], computes the weight importance by monitoring the changes of each weight during training. The importance score of each weight is computed over the entire training steps. At each training step, the contribution of each weight θ_i to the objective loss is proportional to the update size $\Delta\theta_i$ and the gradient $\partial J(\theta)/\partial\theta_i$ as follows;

$$J(\theta + \Delta\theta) - J(\theta) \approx \sum_i \frac{J(\theta)}{\theta_i}\Delta\theta_i. \qquad (3)$$

Therefore, the contribution of parameter θ_i to the total loss over the entire training steps is computed as,

$$W_{\theta_i} = \sum_{mini-batches} -\frac{J(\theta)}{\theta_i}\Delta\theta_i. \qquad (4)$$

Since we are minimizing the total loss, a minus sign is added to the (4). We estimate the per-parameter contribution W_{θ_i} using mini-batches which would introduce noise to the computation. These added noises are expected to result in an overestimation of the true value of W_{θ_i}. Following [16], the per-parameter importances are therefore normalized before leveraging them to regularize the adaptation process;

$$S_{\theta_i} = \frac{W_{\theta_i}}{(\theta_i^{init} - \theta_i^{start})^2 + \epsilon}, \qquad (5)$$

where θ_i^{start} is the i-th parameter of the starting point model for initial training. Here, ϵ is added for practical reasons to

provide an upper bound on the division since $\theta_i^{init} - \theta_i^{start}$ can be very small. To leverage the resulting weight importances in addressing domain expansion, we follow both WCA and EWC frameworks to regularize the training as:

$$J_{SI}(\theta^n) = J_{CTC}(\theta^n) + \lambda_{si} \sum_i S_{\theta_i}(\theta_i^n - \theta_i^{init})^2, \quad (6)$$

where λ_{si} determines the regularization component weight. In this study, we tune the values of ϵ and λ_{si} based on domain expansion performance based on held-out development sets.

D. Soft KL-Divergence (SKLD)

A significant difficulty in domain expansion is to preserve the functionality (input-output mapping) of the initial model $\mathcal{M}^{init}$. Weight constrained approaches (e.g., WCA, EWC, and SI) attempt to accomplish this by preserving the previously learned information stored in the model weights. However, according to experiments performed in [19], generally, restricting learnable parameters is not an efficient way of preserving the functionality of a model since applying slight changes to the parameters can propagate through the entire network and result in significant changes in model functionality and performance.

In another class of methods, preserving the functionality of $\mathcal{M}^{init}$ can be accomplished by imposing constraints on the outputs of the model [31]. By constraining the outputs of $\mathcal{M}^n$ to mimic the outputs of $\mathcal{M}^{init}$ for the same inputs, it is possible to assure that these two models are similar to each other. SKLD implements this through two steps: (1) it takes the initial model $\mathcal{M}^{init}$ and the data for the new domain $\mathcal{D}^n$, and generates an output of $\mathcal{M}^{init}$ (pseudo-labels) for all samples of the dataset; (2) next, SKLD trains the new model $\mathcal{M}^n$ by initializing it from $\mathcal{M}^{init}$ and leveraging the pseudo-labels to regularize the model adaptation as follows:

$$\begin{aligned} J_{SKLD}(\theta^n) &= (1-\lambda_s) J_{CTC}(\theta^n) \\ &\quad + \lambda_s \sum_{i \in I} J_{CE}(\mathcal{M}^{init}(x_i), \mathcal{M}^n(x_i)), \\ J_{CE}(\mathcal{M}^{init}(x_i), \mathcal{M}^n(x_i)) &= \sum_{c \in C} \mathcal{M}_c^{init}(x_i) \log(\mathcal{M}_c^n(x_i)), \end{aligned} \quad (7)$$

where I is the total number of utterances; x_i is the i-th input feature vector; $\mathcal{M}_c^{init}(x_i)$ and $\mathcal{M}_c^n(x_i)$ are the probability of the c-th class generated by initial and new models for input vector x_i. Here, $0 \leq \lambda_s \leq 1$ is a regularization hyper-parameter that provides a compromise between learning the new domain (by optimizing J_{CTC}) and preserving the input-output mapping of the initial model (by optimizing the cross-entropy loss J_{CE}). Here, $\lambda_s = 0$ results in the conventional transfer learning (pre-training/fine-tuning adaptation). By increasing the value of λ_s, the adapted model is more similar to the initial model at a cost of restricting the model's flexibility to learn new data. We tune λ_s to ensure a balanced trade-off between learning a new domain versus mitigating the forgetting effect problem. For experiments, we tune λ_s to achieve the best performance for development sets from both domains.

E. Hybrid SKLD-EWC

In previous sections, we described the SKLD and EWC approaches. In our previous work [17], we suggested that these approaches are complementary and can be combined to form an advanced hybrid approach, namely SKLD-EWC. The advantage of EWC is to leverage the initial data in the first training phase to compute the Fisher information matrix (that quantifies the importance of the weights). However, after the model is updated during adaptation, the curvature of the loss function (consequently the Fisher matrix) changes; therefore, a fixed initial Fisher matrix cannot reliably preserve the original model performance. Alternatively, the advantage of SKLD is that it is more efficient in preserving model functionality since the efficacy of SKLD does not change during the adaptation steps. Our proposed technique can be implemented by imposing both SKLD and EWC constraints on the tuning loss as follows:

$$\begin{aligned} J_{SKLD-EWC}(\theta^n) &= (1-\lambda_s) J_{CTC}(\theta^n) \\ &\quad + \lambda_s \sum_{i \in I} J_{cross}(\mathcal{M}^{init}(x_i), \mathcal{M}^n(x_i)) \\ &\quad + \lambda_e \sum_i diag\{F\}_i (\theta_i^n - \theta_i^{init})^2. \end{aligned} \quad (8)$$

This hybrid method requires two regularization parameters: λ_s and λ_e defined for regularizing the outputs and weights, respectively. These two parameters provide a more flexible domain expansion technique at the expense of more difficult hyper-parameter tuning.

F. Model Averaging (MA)

All regularization approaches aim to reach a balanced model that performs well for both initial and new domain data. However, to achieve the best balanced performance, they have to compromise between characteristics of both datasets. Therefore, it is possible the model never reaches the best possible performance for the new domain data. Alternatively, to reach the best performance for new data, we need to tune all (a large subset of) model parameters. However, the adapted model would then quickly deviate from its initial point and consequently forget the initial data conditions. In our proposed model averaging approach, we address the forgetting problem by post-regularizing an adapted model. Model averaging is performed in two steps; first, it adapts the initial model to reach the minimum loss for the new data. Next, by leveraging the initial optimum model $\mathcal{M}^{init}$ and adapted optimum model $\mathcal{M}^{adp}$, along with held-out development datasets for both domains, it averages parameters of the models to make a compromise between the two ends;

$$\theta_i^n = (1-\lambda_{ma})\theta_i^{init} + \lambda_{ma}\theta_i^{adp}, \quad (9)$$

where λ_{ma} is tuned to achieve the best average performance on the development sets. The intuition behind model averaging is to achieve an advanced solution for the domain expansion problem on the direct path between initial and adapted models. These models would potentially perform superior to models achieved in adaptation steps because; 1) the loss function between initial and adapted models can be smooth and convex-like [32]. As

a result, moving from the initial model towards the adapted model on a direct path would gradually drop performance for the initial data and gain performance for the new domain data. Therefore, a model in the middle of the direct path would have a balanced performance for both datasets. 2) the models on the direct path would now possess the minimum average Euclidean distance from the initial and adapted models. As a result, the middle model overall performance can be better than models with larger distances from both ends (e.g., model checkpoints achieved during adaptation iterations). This approach is referred to as Adaptation Model Averaging (Adaptation-MA).

Hybrid SKLD Model Averaging – The proposed Adaptation-MA tunes all model parameters to reach an optimum point for the new data. However, it achieves an optimum point at a potential cost of catastrophically forgetting of the initial data. Alternatively, regularizing the adaptation using an approach that alleviates the forgetting effects without hurting new data performance would potentially result in an adapted model with higher overall average performance for both domains. The resulting adapted model can be leveraged to improve performance of the model averaging approach. In [31], it was demonstrated that SKLD can improve adaptation performance when the new domain data size is small. Additionally, in our previous work [17], it was shown that SKLD could effectively alleviate forgetting the initial data. Therefore, we propose to leverage SKLD with a small contribution of the pseudo-labels to regularize the adaptation before applying model averaging. This approach is referred to as hybrid SKLD model averaging (SKLD-MA). We hypothesize that SKLD-MA would potentially perform better than Adaptation-MA.

III. Rehearsal-Based Domain Expansion Approaches

This section presents details for several existing and proposed rehearsal-based domain expansion approaches. These methods aim to leverage initial data effectively to sustain model performance for initial data while adapting the model to new domains. In addition, advanced rehearsal-based domain expansion approaches try to address re-training drawbacks by; 1) reducing the computational cost of domain expansion, 2) training an advanced multi-domain model by leveraging domain-imbalanced training data, and 3) generally requiring having access to a subset of the initial data.

A. Averaged Gradient Episodic Memory (AGEM)

For a given model represented by θ and two random samples of the data $x = (x_i, x_o)$ and $y = (y_i, y_o)$, there is an information transfer between these samples if learning each of them reduces the objective loss for the other. In contrast, there is an information interference if learning one increases the loss for the other. Therefore the goal of an ideal domain expansion approach would be to learn the samples of the new data while ensuring minimum (e.g., no) interference with the initial samples. The core idea for many rehearsal-based approaches is based on the angle between gradient vectors for initial and new data [23], [24], [25]. The angle determines whether there is an information transfer or interference between data samples. For example, for two random samples x and y, information transfer occurs if:

$$\frac{\partial J(\theta, x)}{\partial \theta} . \frac{\partial J(\theta, y)}{\partial \theta} > 0, \tag{10}$$

while information interference occurs if:

$$\frac{\partial J(\theta, x)}{\partial \theta} . \frac{\partial J(\theta, y)}{\partial \theta} < 0, \tag{11}$$

where "·" is the dot product operator.

GEM [23] and AGEM [24] are effective continual learning approaches proposed to minimize information inference while learning a sequence of tasks. These algorithms operate by constraining the relative directions between gradient vectors computed for current and past tasks. Adapting the original GEM to our domain expansion setting leads to solving the following problem:

$$\begin{aligned} &\text{minimize}_{\theta} J(\theta, D^n) \\ &\text{s.t. } < g_b^n, g^{init} >= \frac{\partial J(\theta, B^n)}{\partial \theta} . \frac{\partial J(\theta, D^{init})}{\partial \theta} \geq 0 \end{aligned} \tag{12}$$

where for each training step, g_b^n is the model gradient computed with a mini-batch of the new data B^n, and g^{init} is the model gradient for the entire (a subset of) initial data. Solving this problem would ensure that every step of learning for new data does not increase the loss function for the initial domain data. The main drawback of GEM for scenarios where the initial data size is enormous (e.g., domain expansion for ASR) is the burden of computing gradients for the entire (a large subset of) initial data at every step. The high training cost of this approach makes it impractical for ASR settings where the data size could be thousands of hours of data.

AGEM was proposed to mitigate the computational burden of the original GEM. For a continual multi-task learning scenario, AGEM leverages a batch of past task samples to represent an average of the entire previously seen data. In [24], it was shown that for a continual multi-task setting, AGEM performs comparably to original GEM in terms of average accuracy on all tasks; however, AGEM is about 100 times faster. Therefore, in our study, we investigate the efficacy of AGEM as a feasible domain expansion solution for ASR systems. The optimization problem of AGEM is the same as (12), but for AGEM, the gradients of the initial domain data are computed for a mini-batch. Therefore, the condition becomes $< g_b^n, g_b^{init} >= \frac{\partial J(\theta, B^n)}{\partial \theta} . \frac{\partial J(\theta, B^{init})}{\partial \theta} \geq 0$; replacing g^{init} with g_b^{init}. For the remainder of this study, we use AGEM-Criterion to refer to $g_b^n . g_b^{init}$.

The AGEM algorithm addresses the interference problem by computing the angle between g_b^n and g_b^{init}. If the angle is less than 90° (i.e., AGEM-Criterion is positive), then only g_b^n is used to update the model's parameters. However, if the angle is larger than 90° (i.e., AGEM-Criterion is negative), then g_b^n is projected to the nearest L2-distance vector, which maintains the angle within the bound. Mathematically, to find the projected vector, we need to solve the following optimization problem:

$$\text{minimize}_{g_p^n} \quad \frac{1}{2} ||g_b^n - g_p^n||_2^2 \text{ s.t. } g_p^n . g_b^{init} \geq 0, \tag{13}$$

where g_p^n is the projected vector of g_b^n. In [24], it was shown that the solution to this optimization problem is:

$$g_p^n = g_b^n - \frac{g_b^n . g_b^{init}}{g_b^{init} . g_b^{init}} g_b^{init}. \quad (14)$$

Therefore, starting from the initial model $\mathcal{M}^{init}$, AGEM leverages both initial and new domain data. In each training step, based on the AGEM-Criterion, the model's parameters are updated using either g_b^n or g_p^n. To analyze the computational complexity of AGEM, if we only consider forward-backward propagation for training, each epoch of learning using the new data requires AGEM to process the same number of utterances from the initial data. Therefore, the run-time complexity is almost two times larger than an adaptation approach that only leverages the new data. For many practical ASR settings, the initial data size is enormous. However, unseen data is usually significantly smaller. Therefore, the computational burden of processing the new domain dataset twice (i.e., AGEM computational cost) is much smaller than reprocessing the entire initial data plus new data (i.e., re-training computational cost).

B. AGEM-Gradient Averaging (AGEM-GA)

When the angle between g_b^n and g_b^{init} is larger than $90°$, AGEM adds g_b^{init} to the final gradient with a contribution factor of $\lambda_{agem} = -\frac{g_b^n . g_b^{init}}{g_b^{init} . g_b^{init}}$. AGEM considers the minimum contribution of g_b^{init} is needed to make the angle between the projected vector and g_b^{init} smaller than $90°$. However, since both g_b^n and g_b^{init} are computed using small mini-batches, their directions do not accurately represent the true gradient for the entire initial and new datasets [33]. Therefore, targeting the angle between g_p^n and g_b^{init} to be less than $90°$ can be insufficient to guarantee sustained performance for the initial data after training the model with g_p^n. To examine this hypothesis, we propose to introduce a safety margin by adding the g_b^{init} vector to the final gradient vector with a contributing factor of λ_{base}, regardless of the AGEM-Criterion condition. However, to leverage the angle information, we increase the contribution of g_b^{init} by $c\lambda_{agem}$ when the AGEM-Criterion is negative as follows:

$$g_p^n = \begin{cases} g_b^n + \lambda_{base} g_b^{init}, & g_b^n . g_b^{init} \geq 0 \\ g_b^n + (\lambda_{base} + c\lambda_{agem}) g_b^{init}, & g_b^n . g_b^{init} < 0 \end{cases} \quad (15)$$

where c controls the contribution of λ_{agem}.

C. Gradient Averaging (GA)

To study whether angle information provides sufficient beneficial information for AGEM and AGEM-GA, we investigate leveraging gradients for initial and new domain data, and entirely disregard the AGEM-Criterion. In this approach, in every training step, the final projected vector results from averaging the initial gradient vector with the new gradient vector; $g_p^n = g_b^n + \lambda_{base} g_b^{init}$. Setting $\lambda_{base} = 1$ makes this approach equivalent to tuning the initial model $\mathcal{M}^{init}$ with samples drawn from initial and new datasets with equal probability for each set. However, $\lambda_{base} = 1.0$ can be suboptimum. For example, since $\mathcal{M}^{init}$ is already fitted to the initial data, setting $\lambda_{base} < 1.0$ would give sufficient flexibility to the model to learn new data.

GA with Pseudo Labels – In our GA approach, in each training step, for a initial data mini-batch B^{init}, we leverage target labels to compute $g_b^{init} = \frac{\partial J(\theta, B^{init})}{\partial \theta}$. Alternatively, since $\mathcal{M}^{init}$ is already fitted to the initial data, the model can be leveraged to generate pseudo labels for initial utterances. As demonstrated in [17], leveraging pseudo labels computed by a pre-trained initial model mitigates forgetting previously learned information. Therefore, since the purpose of leveraging initial data is to maintain model functionality, one can use pseudo labels as target outputs instead of the true labels (i.e., target transcriptions). To study the efficacy of pseudo labels, we examine performance of GA where we compute initial gradients using pseudo target labels, which is similar to the cross-entropy loss used in the SKLD approach (7).

IV. Experimental Setup

End-to-End LSTM-CTC Model – We evaluate the efficacy of domain expansion techniques for a Long Short-Term Memory (LSTM)-CTC ASR system. The acoustic model consists of two fully connected layers to process the input features. Next, four bidirectional LSTM (BLSTM) [34] layers are used to model temporal relations. Finally, two fully connected layers map the BLSTMs' outputs to target units. The fully connected layers are regularized by drop out. For the BLSTM layers, each layer has 512 cells for each direction. For the intermediate fully connected layers, we use the LeakyReLU [35] activation function. The last layer leverages a softmax activation to map the model's logits to a probability distribution over an output set. Overall, the model contains about 25 million trainable parameters. The last layer of the model consists of $|\mathbf{S}|$ outputs, where $\mathbf{S} = \{English\ characters,\ blank, space, \text{“}, \text{”}, ?, !, ., '\}$. The *blank* unit is a special character used by CTC for computing the loss [29].

For the inference step, we leverage two decoding approaches to map the sequence of probabilities generated by a trained acoustic model to a transcription; Greedy algorithm (best path decoding) [29], and Weighted Finite-State Transducer (WFST)-based decoding [36]. The Greedy algorithm performs decoding through two straightforward steps. First, for each time instance, it chooses the most probable output units. Next, it applies a Squash function to remove repeated units. We leverage the Greedy algorithm to compute a Character-Error-Rate (CER) to measure the quality of the generated output for each of the trained acoustic models. In this study, we tune the hyperparameters of each approach and model (e.g., learning rate or drop out rate) to improve CER performance on the development sets. The performance of the final tuned model for each approach is measured by computing Word-Error-Rate (WER) on the test sets. To compute WER, we leverage the WFST approach to output high-quality word sequences by integrating the lexicon and language model efficiently.

Dataset & Features – To train the initial model, we leverage the LIBRISPEECH corpus (Libri) [37]. This dataset contains about 1000 hr of English read speech mainly recorded from

native US English speakers. Here, the entire Libri training set is leveraged to train the models, and the test-clean set is used for evaluation. For unseen new data, we use Indian (IND), and Australian (AUS) parts of the UT-CRSS-4EnglishAccent corpus [6]. The data for each accent is collected from 100 speakers, with sessions consisting of read and spontaneous speech. For each accent, there is more than 28 h of training data, 5 h of development, and 5 h of evaluation speech data. The relative size of the Libri dataset (i.e., initial training data) compared to new accented datasets (i.e., new domains) resembles real large-scale scenarios for ASR systems. In this scenario, re-training would be highly computational expensive compared to adapting a pre-trained model to unseen domains.

The input features are 40-dimensional filterbank features together with their first and second-order derivatives. The features are extracted from time windows of 25 ms with a frameshift of 10 ms. We expand each frame by stacking four frames from each side to form 1080-dimensional feature vectors. Next, to reduce training time, the resulting frames are downsampled by a skip rate of two frames (i.e., for every three successive frames, one frame will be processed).

We use the standard language model (LM) provided for Libri to decode the initial data. Since accented datasets contain spontaneous utterances, we train n-gram LMs by pooling transcriptions from Fisher, Switchboard, and UT-CRSS-4EnglishAccent corpora.

Training and Evaluation Settings – Model implementation and training are performed using PyTorch [38]. We use an Adam optimizer [39] to train or adapt the models. To train the initial model, the starting learning rate is 3.0×10^{-4}, which is halved after every 90k iterations. For domain expansion experiments, we use a fixed learning rate of 1.0×10^{-4}. Gradients are computed from mini-batches of 16 utterances. For rehearsal-based approaches, in each iteration, the model gradients are computed for two mini-batches of the new and initial datasets. We apply the same evaluation settings to examine regularization-based and rehearsal-based approaches. For each domain expansion approach, the final model (for evaluation) is selected based on average CER performance for initial and new development sets.

We also examine performance of multi-condition training (i.e., re-training) [40]. Here, we first pool all available training data (i.e., initial and new datasets) and train the ASR model using the resulting multi-condition dataset. For many continual learning studies, re-training performance is considered a milestone for domain-expansion methods [15], [16], [24]. However, for our training scenario where initial data is much larger than new datasets, naive multi-condition training could produce a suboptimal performance for both datasets. We leverage domain-specific models to set a target performance for domain expansion approaches. We train domain-specific ASR models for each dataset by adapting the initial model to that dataset. To improve adaptation performance, we regularize training by imposing the SKLD constraints [31]. Note that we leverage SKLD for domain expansion and domain adaptation. However, for domain adaptation, the contribution of the pseudo-labels (e.g., the value of λ_s) is much smaller. The domain-specific performance represents an ideal domain expansion approach that maintains model performance for both the initial data while achieving optimum performance for new datasets. The average domain-specific performance is reported in Tables I & II. Additionally, the relative gap to domain-specific models is reported, referred to as gap to domain-specifics (Gap-DS).

V. Results

We conduct a series of experiments to evaluate performance of the domain expansion methods developed in this study. In the first set, we compare regularization-based approaches with results summarized in Table I. In the second set of experiments, we examine performance of rehearsal-based approaches with results summarized in Table II.

A. Regularization-Based Domain Expansion Approaches

Catastrophic Forgetting – As shown in Table I, the initial model performs well for the initial domain test set. However, due to domain-mismatch, it performs poorly for unseen new domains. Adapting the initial model to new datasets (referred to as Adaptation in Table I) addresses mismatch, but results in categorically forgetting of the initial dataset (i.e., the adapted model performance drops dramatically for the initial dataset). For the Adaptation approach, we only consider model performance on new data. However, to perform domain expansion, we need to determine an early-stop point for the training where average performance on the initial and new domain development sets starts to rise. This approach is referred to as Early-Stopped Adaptation (ES-Adaptation) in Table I. The average performance of ES-Adaptation is significantly better than Adaptation. However, the relative gap to the domain-specific performance is still significant. These experiments demonstrate that an effective domain expansion solution can not be achieved by simply adapting the model and monitoring average performance.

Weight Constrained Approaches Results – The methods WCA, EWC, and SI impose a constraint on each model parameter to limit them from deviating much from the initial solution. First, we compare performance of WCA with ES-Adaptation to examine whether limiting the weight changes can result in a better solution for domain expansion. Results show that WCA outperforms ES-Adaptation for both datasets. For AUS accent, WCA improves WER from $8.52\% \rightarrow 7.93\%$, and for IND accent, WER improves from $10.91\% \rightarrow 10.5\%$. Next, we compare performance of SI and EWC with WCA to see whether leveraging the importance weights computed for SI (i.e., S_θ in (5)) and EWC (i.e., $diag\{F\}$ in (2)) can result in improved domain expansion. Compared to WCA, EWC offers a slight improvement for both AUS and IND accents. However, SI fails to improve performance for AUS accent. One explanation for SI's poor performance is that this approach computes the contribution of each parameter based on gradients computed from mini-batches. However, as we will also demonstrate for the AGEM approach in the next section, the direction of these gradients can be significantly different from the true gradient computed from the entire dataset. Therefore, the resulting weight importance S_θ does not always reliably represent the contribution of each parameter. Although these approaches have performed well for small models and

TABLE I
WERs OF REGULARIZATION-BASED DOMAIN EXPANSION METHODS ON THE INITIAL AND NEW DOMAINS

Regularization Approaches								
	AUS (WER%)				IND (WER%)			
Method	Initial	New	Average	Gap-DS(%)	Initial	New	Average	Gap-DS(%)
Domain Specific Models	5.06	7.32	6.2	-	5.06	11.9	8.48	-
Multi-Condition	5.16	8.12	6.64	7.1	5.43	13.2	9.32	9.90
Initial	5.06	13.87	9.46	55.5	5.06	19.3	12.18	43.63
Adaptation	16.76	8.06	12.41	100.2	17.69	13.11	15.4	81.6
Early-Stopped Adaptation	8.08	8.96	8.52	37.42	8.35	13.46	10.91	28.6
WCA	7.74	8.12	7.93	27.9	7.61	13.4	10.5	23.8
EWC	6.89	8.74	7.82	26.1	7.15	13.47	10.31	21.6
SI	7.26	8.68	7.97	28.5	7.57	12.95	10.26	21.0
SKLD	7.42	7.92	7.67	23.7	7.83	12.67	10.25	20.9
SKLD-EWC	6.88	8.08	7.48	20.6	6.85	13.13	10.0	17.9
Adaptation-Model Averaging (MA)	6.96	7.86	7.41	19.5	7.48	12.33	9.9	16.74
SKLD-Model Averaging (MA)	**6.79**	**7.48**	**7.14**	**15.16**	**7.18**	**12.07**	**9.62**	**13.44**

The results are presented for two scenarios, where the unseen domain is the Indian Accent (IND) or the Australian Accent (AUS). For each approach, relative WER GAP compared to average performance of domain-specific models is presented in GAP-DS column.

TABLE II
WERs OF REHEARSAL-BASED DOMAIN EXPANSION METHODS ON THE INITIAL AND NEW DOMAINS

Rehearsal Approaches								
	AUS (WER%)				IND (WER%)			
Method	Initial	New	Average	Gap-DS(%)	Initial	New	Average	Gap-DS(%)
Domain Specific Models	5.06	7.32	6.2	-	5.06	11.9	8.48	-
Multi-Condition	5.16	8.12	6.64	7.1	5.43	13.2	9.32	9.90
AGEM	8.74	7.82	8.28	33.54	8.85	13.15	11.0	29.7
AGEM-GA	5.47	7.47	6.47	4.35	5.41	12.14	8.77	3.42
GA-Pseudo Labels	5.6	7.33	6.46	4.2	5.63	12.4	9.01	6.25
GA (λ_{base} = 1.0)	5.35	7.4	6.38	2.9	5.39	12.25	8.82	4.0
GA (λ_{base} = 0.5)	**5.37**	**7.24**	**6.3**	**1.6**	**5.41**	**11.92**	**8.66**	**2.12**

The results are presented for two scenarios, where the unseen domain is the Indian Accent (IND) or the Australian Accent (AUS). The gradient averaging approach is examined for two settings, where λ_{base} = 0.5 or λ_{base} = 1.0. For each approach, relative WER GAP compared to average performance of domain-specific models is presented in GAP-DS column.

simple image recognition tasks [15], [16], this study as well previous studies for speech recognition [17], [20] show that these approaches do not offer the same performance benefits where tasks and models become complicated (e.g., RNN models for ASR).

SKLD Results – Compared to the weight constrained-based approaches (i.e., WCA, EWC, and SI), SKLD offers the same or better performance for new datasets. Experiments demonstrate superior performance of hybrid SKLD-EWC, which benefits from curvature information of the model and maintains model functionality using SKLD. Compared to ES-Adaptation (the baseline approach for regularization-based approaches), SKLD-EWC improves relative average WER by 8.3%–12.2%, and the relative gap to the domain-specific performance is reduced from 28.6% → 17.9% for IND, and 37.42% → 20.6% for AUS datasets.

Model Averaging (MA) Results – We first examine performance of the average model along the linear path connecting the initial model $\mathcal{M}^o$ (corresponding to the Initial model in Table I) and the adapted model $\mathcal{M}^{adp}$ (corresponding to the Adaptation approach in Table I). By taking a linear average using (9), we create new models that have not been reached during regular adaptation steps. As demonstrated in Fig. 3, starting from the initial model and shifting towards the adapted model along the linear path, performance of the resulting model drops smoothly for the initial data. However, we do not observe any performance spikes along this path. Furthermore, we observe the same trend for new data, where performance of the average model improves smoothly by increasing the contribution of the adapted model. These experiments support the hypothesis that for our settings, the objective loss along the linear path between initial and adapted models is smooth and convex-like for both datasets.

To investigate model averaging (MA) for domain expansion, we first examine performance of Adaptation Model Averaging. The results demonstrate the efficacy of this approach in proving a trade-off between retaining the initial model and learning new domains. However, as shown in Fig. 3, since the adapted

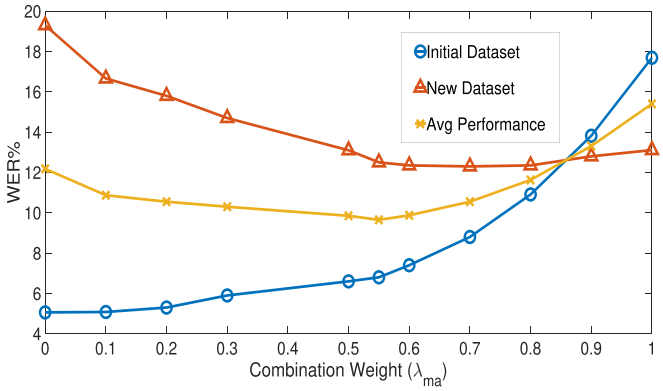

Fig. 3. Examining word error rate (WER) along a linear path connecting the initial model $\mathcal{M}^{o}$ and adapted model $\mathcal{M}^{adp}$. WERs are computed for held-out development sets of the Libri dataset (Initial Dataset) and Indian accent dataset (New Dataset). $\lambda_{ma} = 0$ results in the initial model, which performs poorly for the new dataset. By increasing λ_{ma}, the WER improves for the new dataset, and it drops for the initial data. For $0.5 < \lambda_{ma} < 0.6$, we reach a performance balance for both domains.

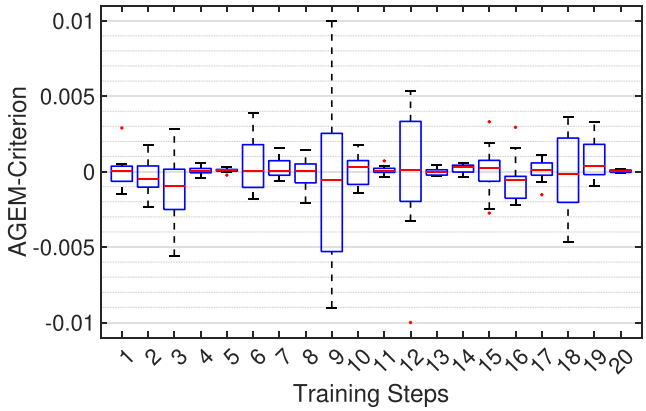

Fig. 4. Examining the range of AGEM-Criterion values in each training step. To analyze the AGEM approach, in each training step, for a mini-batch of new data, AGEM-Criterion is computed for ten different mini-batches of initial data. This figure presents the resulting AGEM-Criterion distribution for 20 training steps using a Box and Whisker plot. **Key observation**: AGEM-Criterion fluctuates between positive and negative values; therefore, it is not a reliable criterion.

model catastrophically forgets the initial data, performance of the average model for the initial data drops quickly as the model shifts toward the adapted model. Therefore, the final average model reaches a suboptimal performance. The results demonstrate that the hybrid SKLD-MA addresses this problem and outperforms other regularization-based approaches. Comparing performance of Adaptation-MA with SKLD-MA shows that SKLD regularization is a critical component because it alleviates cartographic forgetting without impacting performance on new data. Note that we leverage the SKLD constraint with varying contributions along different settings. For example, for SKLD-MA, we set $\lambda_s = 0.9$ to regularize the adaptation. However, for SKLD domain expansion, the optimum value of λ_s is in the range of [0.6,0.7]. Compared to ES-Adaptation, SKLD-MA improves the relative average WER by 11.8%-16.25%. The relative performance gap between domain-specific models and the SKLD-MA is narrowed to 15.16% for AUS and 13.44% for IND datasets.

B. Rehearsal-Based Domain Expansion Approaches

In this section, we evaluate and compare performance of rehearsal-based approaches. For results provided in Table II, all domain expansion approaches leverage the entire initial and new training datasets.

AGEM Results – Results show that AGEM significantly underperforms other domain expansion approaches. By analyzing results in Table II, the poor performance is mainly due to forgetting the initial domain data. Therefore, as we hypothesized, the assumption of making the angle between the gradient vectors less than 90° might not be sufficient to alleviate forgetting of the initial domain. To further analyze the problem, we examine the AGEM-Criterion. To this end, in each training step, for a mini-batch of the new data, we compute AGEM-Criterion for ten different mini-batches of the initial data. As shown in Fig. 4, in every training step, depending on each mini-batch for which the initial gradients are computed, the AGEM-Criterion fluctuates greatly between positive and negative values. The gradients of the initial data have varying directions because the model is already fitted to the initial dataset. Consequently, some gradients are too small, with potential unreliable directions. In addition, it is known that gradients computed for a mini-batch do not accurately represent overall true gradients for the entire dataset, and their directions change for different mini-batches [33]. Therefore, the AGEM-Criterion disregards the initial gradient directions in many training steps, which results in forgetting the previously learned information from the initial dataset.

By comparing performance of AGEM-GA and AGEM, it can be inferred that we can significantly alleviate forgetting of the initial data by increasing the contribution of the initial gradients. In addition to a significant gain in initial data performance, AGEM-GA performs better than AGEM for the new domains as well. Here, WER improves from 7.82% to 7.47% for AUS accent, and from 13.15% to 12.14% for IND accent. This observation suggests that AGEM-GA results in a generalized model by leveraging initial data in every iteration, which also regularizes adapting the model to the new small-sized datasets.

Gradient Averaging (GA) Results – Compared to AGEM and AGEM-GA, the results demonstrate the superiority of GA. Based on the results, it can be inferred that the angle-based criterion does not offer additional information in finding an effective solution for domain expansion. In each training iteration, GA computes an averaged vector of the two gradient vectors regardless of their relative directions. Therefore, it is only necessary to tune the relative weight of the gradient vectors, and then over training iterations, these averaged gradients would lead to an optimum domain expansion solution.

The GA method does outperform other domain expansion approaches explored in this study, which narrows the relative gap to the performance of domain-specific models to 1.6% for AUS and 2.12% for IND datasets. The gap towards domain-specific models is mainly due to the difficulty in maintaining initial data performance. One can retain initial data performance by

increasing the contribution of the initial gradients. However, such an approach will constrain the model to learn the new domain data, consequently leading to an overall reduced optimum domain expansion solution. For example, we have examined GA in two settings, with one being a tuning of the hyperparameters to reach optimum average performance, corresponding to GA ($\lambda_{base} = 0.5$) in Table II. For the other setting, the two gradients of the initial and new data are added with the same weight, corresponding to GA ($\lambda_{base} = 1.0$). As results show, GA with $\lambda_{base} = 1.0$ performs slightly better for the initial data, but degrades for the new domain as compared with the optimal setting of GA. Additionally, performing GA with pseudo labels (described in Section III-C, corresponding to GA-Pseudo Labels in Table II) results in performance degradation. The poor performance can also be due to the fixed pseudo-labels, which constrain the adapted model from training efficiently on new datasets.

The proposed GA approach outperforms multi-condition training, which can be attributed to leveraging the initial model as well as benefiting from a flexible weight to combine gradients for the initial and new datasets. In terms of computational complexity, GA also performs better than multi-condition training. The computational burden of processing the entire initial dataset plus the new dataset (to perform multi-condition training) is significantly higher than processing two times the size of new datasets (to perform GA). This observation is valid for ASR domain expansion scenarios, where new datasets are usually significantly smaller than initial datasets.

Impact of Buffer Size – The rehearsal-based approaches considered leveraging the entire initial dataset to maintain model performance. However, storing and reusing the entire initial dataset can be an obstacle to applying these approaches for real large-scale ASR scenarios. Therefore, it is generally preferred in such cases to store and reuse a small amount of the initial dataset. This section investigates the performance impact of using a limited amount of initial data for domain expansion. The number of stored utterances of the initial dataset is referred to as the buffer size. Here, we study performance GA as a function of buffer size. The buffer size changes from zero (i.e., Early Stopped Adaptation) to the size of the entire initial domain dataset. The results are presented in Fig. 5. Note that for each buffer size setting, the final model is selected based on its combined average performance for initial and new datasets. Consequently, we observe that limiting the buffer size affects model performance for both domains.

The results demonstrate that leveraging initial data as small as 1 k utterances significantly impacts performance. Here a model with a "buffer size = 1 k" improves relative performance by 8.3% compared to a "buffer size = 0". Furthermore, increasing the buffer size to 5 k and 30 k utterances also results in a significant relative improvement. However, an interesting observation is that further increasing the buffer size to 60 k or 100 k utterances only slightly improves performance. Therefore, for our experimental settings, it can be inferred that a reasonable compromise between performance and buffer size is a "buffer size = 30 k". After this point, the rate of improvement declines, where increasing the buffer size from 30 k to 281 k utterances (entire utterances of Libri960) improves the relative performance by only 3.33%, which is minimal.

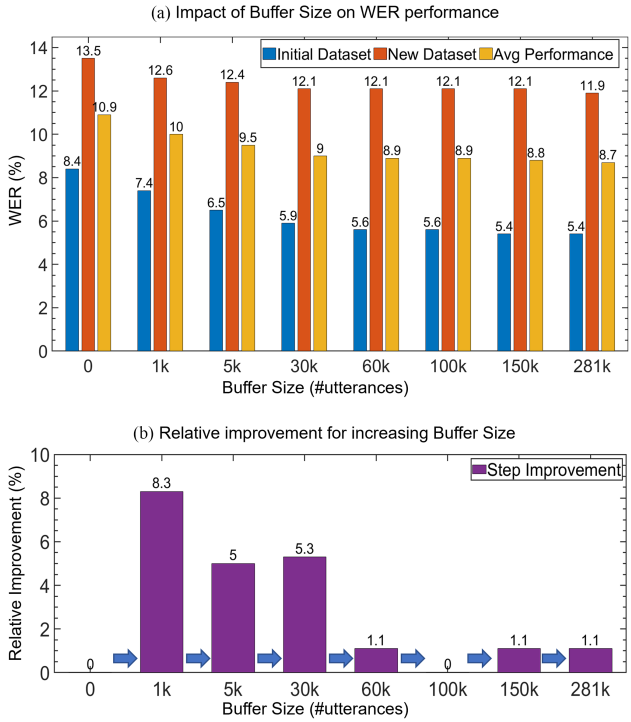

Fig. 5. WER performance of the Gradient Averaging approach as a function of initial data buffer size. (a) WER performance of Gradient Averaging on initial data, new data (for this figure, we use Indian accent data), and average WER for each buffer size. (b) Relative improvement of average WER compared to a preceding buffer size.

Comparing performance of GA for different buffer sizes with SKLD-MA (reported in Table I) demonstrates the significant advantage of leveraging initial data to preserve initial domain model performance. Results show that leveraging 1k-5 k utterances of the initial dataset leads to a comparable performance to the best regularization-based approach.

VI. Conclusion

This study has proposed a number of novel continual learning-based techniques for an effective domain expansion solution in ASR. We examined the efficacy of approaches through experiments on adapting a model trained with native English to two unseen English accents: Australian and Indian. We demonstrate that adapting the initial model to new domains results in improved performance for new domains but at the expense of a significant performance degradation for the initial domian dataset. We addressed performance degradation assuming two scenarios; (1) where only data for the new domain is available during adaptation, and (2) where data for the initial domain and new domains are both available during adaptation.

For the first scenario, we investigated regularization-based approaches to mitigate performance loss for the initial domain data. Weight constrained-based approaches (i.e., WCA, EWC, SI) improve domain expansion performance compared to ES-Adaptation (i.e., the baseline domain expansion approach).

However, results demonstrated that improvement is mainly attributed to the weight constrained regularization terms, and that parameter importance weights (computed by EWC and SI) do not provide significant performance gains. Although these approaches have performed well for small models and simple image recognition tasks, we hypothesized that these approaches would not offer the same performance for ASR, where tasks and models are complicated. The proposed SKLD approach offers the same or better performance compared to weight-constrained approaches. The superior performance of the hybrid SKLD-EWC approach supports our hypothesis that curvature information (computed by EWC) can be complementary to the SKLD approach.

Next, results demonstrated the efficacy of our proposed MA approach in providing a trade-off between retaining the initial domain model and learning new domains. The experiments also support our hypothesis behind MA, where the objective loss for the datasets is smooth and convex-like along a linear path. Performing MA after adaptation regularized with SKLD (i.e., SKLD-MA) outperforms other regularization-based approaches and improves the relative average WER by 11.8%–16.25%, compared to the baseline ES-Adaptation.

For the second domain expansion scenario, where initial domain data is also available, we explored rehearsal-based approaches to leverage initial data to prevent initial domain performance loss. The AGEM approach, which operates based on the angle between gradient vectors, underperforms other rehearsal-based approaches, mainly due to forgetting initial domain data. AGEM-GA however increases the contribution of the initial gradients, which improves overall performance. In general, the superior performance of GA demonstrates that an angle-based criterion does not offer additional information in finding an effective domain expansion solution. GA narrows the relative performance gap to the best domain-specific models of 1.6% for AUS and 2.12% for IND datasets. Furthermore, GA offers better WER performance and improved computational cost compared to multi-condition training. This can be attributed to leveraging the initial model and benefiting from a flexible weight to combine gradients for the initial and new datasets. Examining the impact of the available initial domain data buffer size shows that increasing the buffer size to 30 k utterances improves performance significantly. However, further increasing the buffer size to 281 k utterances (i.e., entire initial utterances) only slightly improves the performance. Therefore, it can be inferred that a reasonable compromise between performance and buffer size is using a "buffer size $=$ 30 k".

This study has, therefore, demonstrated the merits of proposed regularization-based and rehearsal-based approaches as practical solutions for domain expansion in ASR systems. Future work could explore using our proposed approaches for multi-step domain expansion scenarios as well as improving ASR performance under other domain mismatch scenarios.

REFERENCES

[1] G. Saon, Z. Tüske, D. Bolanos, and B. Kingsbury, "Advancing RNN transducer technology for speech recognition," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2021, pp. 5654–5658.

[2] Q. Xu et al., "Self-training and pre-training are complementary for speech recognition," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2021, pp. 3030–3034.

[3] Y. Zhang et al., "Pushing the limits of semi-supervised learning for automatic speech recognition," in *Proc. NeurIPS SAS Workshop*, Vancouver, Canada, 2020.

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

[5] A. Hinsvark et al., "Accented speech recognition: A survey," 2021, *arXiv:2104.10747*. [Online]. Available: https://arxiv.org/abs/2104.10747

[6] S. Ghorbani and J. H. L. Hansen, "Leveraging native language information for improved accented speech recognition," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2018, pp. 2449–2453.

[7] S. Sun, B. Zhang, L. Xie, and Y. Zhang, "An unsupervised deep domain adaptation approach for robust speech recognition," *Neurocomputing*, vol. 257, pp. 79–87, 2017.

[8] Z. Meng, J. Li, Y. Gaur, and Y. Gong, "Domain adaptation via teacher-student learning for end-to-end speech recognition," in *Proc. IEEE Autom. Speech Recognit. Understanding Workshop*, 2019, pp. 268–275.

[9] A. Narayanan et al., "Toward domain-invariant speech recognition via large scale training," in *Proc. IEEE Spoken Lang. Technol. Workshop*, 2018, pp. 441–447.

[10] H. Hu et al., "REDAT: Accent-invariant representation for end-to-end ASR by domain adversarial training with relabeling," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2021, pp. 6408–6412.

[11] S. Ghorbani, A. E. Bulut, and J. H. L. Hansen, "Advancing multi-accented LSTM-CTC speech recognition using a domain specific student-teacher learning paradigm," in *Proc. IEEE Spoken Lang. Technol. Workshop*, 2018, pp. 29–35.

[12] A. A. Rusu et al., "Progressive neural networks," 2016, *arXiv:1606.04671*. [Online]. Avilable: https://arxiv.org/abs/1606.04671

[13] S. Sadhu and H. Hermansky, "Continual learning in automatic speech recognition," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2020, pp. 1246–1250.

[14] X. Li and J. Bilmes, "Regularized adaptation of discriminative classifiers," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2006, pp. 237–240.

[15] J. Kirkpatrick et al., "Overcoming catastrophic forgetting in neural networks," *Proc. Nat. Acad. Sci. USA*, vol. 114, no. 13, pp. 3521–3526, 2017.

[16] F. Zenke, B. Poole, and S. Ganguli, "Continual learning through synaptic intelligence," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 3987–3995.

[17] S. Ghorbani, S. Khorram, and J. H. L. Hansen, "Domain expansion in DNN-based acoustic models for robust speech recognition," in *Proc. IEEE Autom. Speech Recognit. Understanding Workshop*, 2019, pp. 107–113.

[18] Z. Li and D. Hoiem, "Learning without forgetting," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 12, pp. 2935–2947, Dec. 2018.

[19] H. Jung, J. Ju, M. Jung, and J. Kim, "Less-forgetful learning for domain expansion in deep neural networks," in *Proc. AAAI Conf. Artif. Intell.*, 2018, pp. 3358–3365.

[20] B. Houston and K. Kirchhoff, "Continual learning for multi-dialect acoustic models," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2020, pp. 576–580.

[21] T. L. Hayes, N. D. Cahill, and C. Kanan, "Memory efficient experience replay for streaming learning," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2019, pp. 9769–9776.

[22] T. J. Draelos et al., "Neurogenesis deep learning: Extending deep networks to accommodate new classes," in *Proc. IEEE Int. Joint Conf. Neural Netw.*, 2017, pp. 526–533.

[23] D. Lopez-Paz and M. Ranzato, "Gradient episodic memory for continual learning," in *Proc. 31st Int. Conf. Neural Inf. Process. Syst.*, 2017, pp. 6470–6479.

[24] A. Chaudhry, R. Marc'Aurelio, M. Rohrbach, and M. Elhoseiny, "Efficient lifelong learning with A-GEM," in *Proc. Int. Conf. Learn. Representations*, 2019, pp. 392–411.

[25] M. Farajtabar, N. Azizan, A. Mott, and A. Li, "Orthogonal gradient descent for continual learning," in *Proc. Int. Conf. Artif. Intell. Statist.*, 2020, pp. 3762–3773.

[26] T. Hospedales, A. Antoniou, P. Micaelli, and A. Storkey, "Meta-learning in neural networks: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 9, pp. 5149–5169, Sep. 2022.

[27] G. Gupta, K. Yadav, and L. Paull, "Look-ahead meta learning for continual learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, pp. 11588–11598.

[28] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 1126–1135.
[29] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks," in *Proc. 23rd Int. Conf. Mach. Learn.*, 2006, pp. 369–376.
[30] D. Maltoni and V. Lomonaco, "Continuous learning in single-incremental-task scenarios," *Neural Netw.*, vol. 116, pp. 56–73, 2019.
[31] D. Yu, K. Yao, H. Su, G. Li, and F. Seide, "KL-divergence regularized deep neural network adaptation for improved large vocabulary speech recognition," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2013, pp. 7893–7897.
[32] S.-W. Lee, J.-H. Kim, J. Jun, J.-W. Ha, and B.-T. Zhang, "Overcoming catastrophic forgetting by incremental moment matching," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2017, pp. 4652–4662.
[33] I. Goodfellow, Y. Bengio, A. Courville, and Y. Bengio, *Deep Learning*, vol. 1. Cambridge, MA, USA: MIT Press, 2016.
[34] A. Graves and J. Schmidhuber, "Framewise phoneme classification with bidirectional LSTM and other neural network architectures," *Neural Netw.*, vol. 18, no. 5/6, pp. 602–610, 2005.
[35] A. L. Maas, A. Y. Hannun, and A. Y. Ng, "Rectifier nonlinearities improve neural network acoustic models," in *Proc. Int. Conf. Mach. Learn.*, 2013, Art. no. 3.
[36] Y. Miao, M. Gowayyed, and F. Metze, "EESEN: End-to-end speech recognition using deep RNN models and WFST-based decoding," in *Proc. IEEE Autom. Speech Recognit. Understanding Workshop*, 2015, pp. 167–174.
[37] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2015, pp. 5206–5210.
[38] A. Paszke et al., "Pytorch: An imperative style, high-performance deep learning library," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 8026–8037.
[39] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Representations*, 2015, pp. 1–13.
[40] M. L. Seltzer, D. Yu, and Y. Wang, "An investigation of deep neural networks for noise robust speech recognition," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2013, pp. 7398–7402.

Shahram Ghorbani (Student Member, IEEE) received the B.S. degree in computer engineering from Shahid Chamran University of Ahvaz, Ahvaz, Iran, in 2011, and the M.S. degree in computer science from the Sharif University of Technology, Tehran, Iran, in 2013. He is currently working toward the Ph.D. degree with the Center for Robust Speech Systems, University of Texas at Dallas, Richardson, TX, USA. In the summers of 2019 and 2020, he was a Research Intern with Tencent AI Lab, Bellevue, WA, USA, and Microsoft Corp, Redmond, WA, respectively. His research interests include speech signal processing, speech enhancement, multimodal speech recognition, accented speech recognition, domain expansion, continual learning, machine learning, and deep learning.

John H.L. Hansen (Fellow, IEEE) received the B.S.E.E. degree from the College of Engineering, Rutgers University, New Brunswick, NJ, USA, the M.S. and Ph.D. degrees in electrical engineering from the Georgia Institute of Technology, Atlanta, GA, USA, and the honorary Doctor Technices Honoris Causa degree from Aalborg University, Aalborg, Denmark, in recognition of his contributions to the field of speech signal processing and speech/language/hearing sciences, in April 2016. In 2005, he joined the Erik Jonsson School of Engineering and Computer Science, University of Texas at Dallas (UTDallas), Richardson, TX, USA, where he is currently the Associate Dean of research, and a Professor of electrical and computer engineering, the Distinguished University Chair of telecommunications engineering, and holds a joint appointment as a Professor with the School of Behavioral and Brain Sciences (Speech & Hearing). From August 2005 to December 2012, he was the Department Head of electrical engineering with UTDallas. At UTDallas, he established the Center for Robust Speech Systems (CRSS). He was the Department Chairman and a Professor of speech, language and hearing sciences, and a Professor of electrical and computer engineering with the University of Colorado - Boulder, Boulder, CO, USA during 1998–2005, where he Co-founded and was the Associate Director of the Center for Spoken Language Research. In 1988, he established the Robust Speech Processing Laboratory and continues to direct research activities in CRSS, UTDallas. He has supervised 99 Ph.D./M.S. thesis candidates (58 Ph.D., 41 M.S./M.A.). He is the author or coauthor of 863 journal and conference papers, including 13 textbooks in the field of speech processing and language technology, signal processing for vehicle systems, coauthor of textbook *Discrete-Time Processing of Speech Signals* (IEEE Press, 2000), *Advances for In-Vehicle and Mobile Systems: Challenges for International Standards* (Springer, 2006), *In-Vehicle Corpus and Signal Processing for Driver Behavior* (Springer, 2008), and the lead author of the report The Impact of Speech Under Stress on Military Speech Technology, (NATO RTO-TR-10, 2000). His research interests include digital speech processing, analysis and modeling of speech and speaker traits, speech enhancement, feature estimation in noise, signal processing for hearing impaired/cochlear implants, robust speech recognition with emphasis on machine learning and knowledge extraction, and in-vehicle interactive systems for hands-free human-computer interaction. He was recognized as the IEEE Fellow in 2017 for his contributions in Robust Speech Recognition in Stress and Noise, International Speech Communication Association Fellow in 2010 for his contributions on research for speech processing of signals under adverse conditions. He previously served two terms as ISCA President (2017–19; 2020–22), and currently serves as ISCA Treasurer and Member of the ISCA Board. He currently serves as a member of OSAC-SPEAKER (U.S. Office of Scientific Advisory Committees) in the voice forensics domain, having served as Vice-Chair from 2015–21. Previously, he was the IEEE Technical Committee Chair and a Member of the IEEE Signal Processing Society: Speech-Language Processing Technical Committee (SLTC) (during 2005–2008 and 2010–2014; elected IEEE SLTC Chairman during 2011–2013, Past-Chair for 2014), and an elected ISCA Distinguished Lecturer during 2011–2012. He was also a Member of the IEEE Signal Processing Society Educational Technical Committee during 2005–2008 and 2008–2010; a Technical Advisor to the U.S. Delegate for NATO (IST/TG-01), IEEE Signal Processing Society Distinguished Lecturer during 2005–2006, an Associate Editor for the IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING during 1992–1999 and IEEE SIGNAL PROCESSING LETTERS during 1998–2000, Editorial Board Member of IEEE SIGNAL PROCESSING MAGAZINE during 2001–2003, and the Guest Editor in October 1994 for the Special Issue on Robust Speech Recognition of the IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING. He is an Associate Editor for the *Journal of the Acoustical Society of America*, and was on the Speech Communications Technical Committee for the *Acoustical Society of America* during 2000–2003. He was the recipient of the Acoustical Society of America's 25 Year Award in 2010 in recognition of his service, contributions, and membership to the Acoustical Society of America, 2020 Provost's Award for Excellence in Graduate Student Supervision - University of Texas-Dallas, and 2005 University of Colorado Teacher Recognition Award as voted on by the student body. He organized and was the General Chair of ISCA Interspeech-2002, September 16–20, 2002, a Co-organizer and the Technical Program Chair of IEEE ICASSP-2010, Dallas, TX, March 15–19, 2010, and Co-Chair and Organizer for IEEE SLT-2014, December 7–10, 2014, Lake Tahoe, NV, USA. He served as Co-Chair and Organizer for ISCA INTERSPEECH-2022, and will serve as Technical Program Chair for IEEE ICASSP-2024.