CROSS-ATTENTIVE ADAPTER WITH REGULARIZED DOMAIN ADAPTATION FOR SPEAKER VERIFICATION

Vishwas M. Shetty, Anthony Wong, Abeer Alwan

University of California, Los Angeles
Department of Electrical and Computer Engineering
Los Angeles, USA

ABSTRACT

We address catastrophic forgetting (CF) in automatic speaker verification (SV) during domain adaptation from Adult English (VoxCeleb) to Child English (MyST) and Adult Chinese (CNCeleb), under the assumption of no access to VoxCeleb data. We leverage embeddings of target-data inputs extracted from the pretrained VoxCeleb model, hypothesizing that these embeddings retain source-domain knowledge without requiring source data. These target-data embeddings extracted from the pretrained VoxCeleb model are referred to as pretrained-model embeddings (PMEs). We propose a cross-attentive (CA) adapter that reduces CF by dynamically balancing information between learnable target-domain embeddings and PMEs during adaptation. We design two regularization strategies: (i) K-means–based elastic weight consolidation (K-EWC), where clustered PMEs provide pseudo labels for parameter-importance estimation, and (ii) a moment-matching (MM) loss that constrains learnable target-domain embeddings to remain close to the PME distribution. We evaluate across target data training splits of increasing duration. Results show that our approaches consistently reduce CF by not degrading performance (%EER) on the source domain regardless of target data size.

***Index Terms*—** Catastrophic Forgetting, Zero-source domain data, Adapter, Regularization

1. INTRODUCTION

A neural network should be stable enough to preserve the information it has learned over time while remaining plastic enough to adapt to new tasks. An imbalance between stability and plasticity leads to the problem of Catastrophic forgettting (CF) [1, 2]. Approaches to strike this balance can be grouped into three categories: **(a) Architectural based approaches** that modify or expand the network architecture to mitigate interference between tasks include adding a new classification layer (output head) for each new task [3], using adapters [4], isolating task-specific parameters, or dynamically expanding network capacity [5, 6, 7]. These approaches allow shared representation layers to remain stable while dedicating new or isolated parameters to new tasks, effectively preserving prior knowledge. **(b) Rehearsal Based Approaches** involve retraining the network using a subset of samples from previous tasks alongside data from the new task [8, 9]. However, access to prior (or source) task data can be limited due to storage, privacy, or legal constraints. To address lack of source task data, "rehearsal-free" approaches [10, 11] and "pseudo-rehearsal" mechanisms are used where representative samples that approximate past data are generated synthetically [8, 12]. A related approach, "generative replay", leverages generative models such as Variational Autoencoders (VAEs) or Generative Adversarial Networks (GANs) to produce artificial samples that mimic the distribution of previous tasks [13, 14, 15]. These generated samples are then replayed during training, allowing the network to retain prior knowledge without storing real data. **(c) Regularization Based Approaches** introduce constraints on network parameters or outputs to prevent significant interference with previously learned tasks. Regularization methods include *Elastic Weight Consolidation (EWC)* [16], *Synaptic Intelligence (SI)* [17], and L2 regularization [18]. Knowledge distillation techniques also fit in this context. Here the outputs of a pre-trained network (teacher) are used to regularize the training of the current network (student), ensuring that prior task knowledge is preserved while learning new tasks [12, 19, 20].

In this work, we tackle catastrophic forgetting in speaker verification under the challenging zero-source-data constraint. Our primary focus lies in two application settings: *cross-age* adaptation (Adult → Child) and *cross-lingual* adaptation (English → Chinese). We propose approaches to operate effectively under zero-source-data constraint. We leverage embeddings of target-domain inputs extracted from the pretrained model, which we refer to as pretrained-model embeddings (PMEs), to approximate source-domain knowledge, without requiring actual source data. We propose three strategies: (i) **Cross-Attentive (CA) Adapter**, a hybrid of knowledge distillation and adapter methods; (ii) **K-means-based modified Elastic Weight Consolidation (K-EWC)**, for parameter regularization, and (iii) a **Moment-Matching Regularization Loss**, which constrains fine-tuning by aligning learnable target-domain embeddings distribution with PME distributions.

The most closely related approaches to our proposed techniques include [12, 21, 22, 23, 24, 25]. These works similarly exploit representations from a pre-trained model to guide adaptation without full source-data access. However, [21] requires access to source data to initially estimate the Fisher matrix for EWC. While [23, 24] rely on target test data for estimating Fisher matrix, but this might not work well in highly mismatched domains such as Adult-to-Child. In Adult-to-Child SV adaptation, [25] proposed an age-agnostic SV system, but their approach depended on both source and target data. Our proposed approach removes the dependency on source data by using PMEs. The K-EWC method circumvents the need for labeled source data by leveraging PMEs and their corresponding K-Means labels. CA adapter extends knowledge distillation strategies [12, 22] by dynamically balancing information between PMEs and the learnable target-domain representations during adaptation.

2. METHOD

2.1. Cross-Attentive (CA) Adapter

The proposed Cross-Attentive (CA) adapter (**CA Block** in Fig 1) is integrated into a Siamese-style network architecture as shown in Fig. 1 (a). It modifies the Attentive Statistics Pooling (ASP) layer of the ECAPA-TDNN [26] model, enabling more effective retention

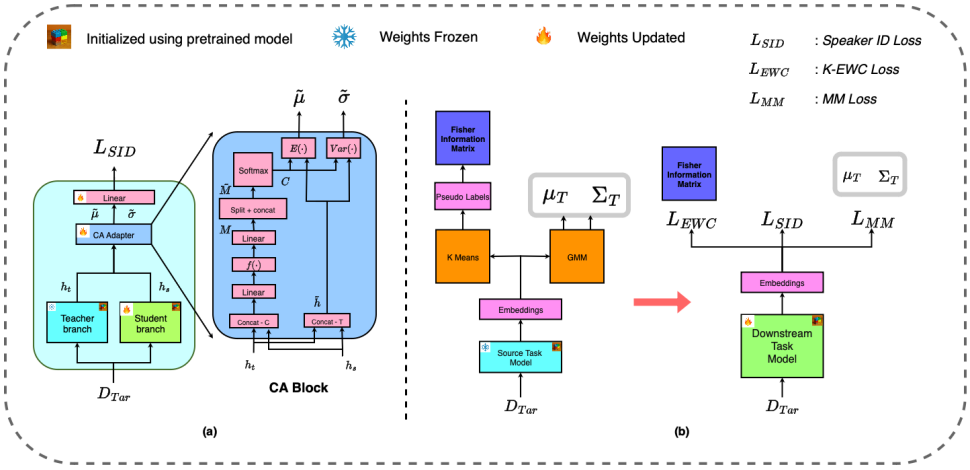

Fig. 1. *An overview of the proposed approaches. The CA adapter framework is shown in (a). Hidden representations h_t and h_s have dimension $C \times T$. In the CA Block, "Concat-C" and "Concat-T" refer to concatenation operation on the hidden representations along the C (channel) and T (time) axes, respectively. Estimation of Fisher Information Matrix (FIM) for K-EWC using pseudo-labels / Estimation of pseudo μ_T and Σ_T for MM regularization is shown in (b). D_{Tar} refers to the target domain data.*

of source-domain information while adapting to the target domain. The teacher and student embedding networks in Fig 1 (a), consist of the layers up to the Multi-layer Feature Aggregation (MFA) module of the ECAPA-TDNN architecture, and are initialized with the corresponding weights from the pre-trained model. The ASP layer is then modified to incorporate the proposed CA adapter, enabling cross-attentive integration of teacher outputs (i.e., fixed PMEs) and student outputs (i.e., learnable target-domain representations).

The teacher branch, is kept frozen throughout. Given the data from the target domain D_{Tar}, the output of the MFA layer, i.e., the hidden representations from the teacher and student branches - h_t, h_s, each $\in \mathbb{R}^{1536\times T}$, are stacked channel-wise ("Concat-C" within **CA Block** in Fig. 1), followed by a linear transformation with parameters $W \in \mathbb{R}^{l\times 3072}$ and $b \in \mathbb{R}^{l\times 1}$ to down-project into a smaller latent space with dimension l, using a non-linearity $f(.)$. After the non-linearity $f(.)$, the self-attention scores M across both hidden representations are calculated through a second linear transformation with parameters $V \in \mathbb{R}^{3072\times l}$ and $k \in \mathbb{R}^{3072}$.

$$M = V\left[f\left(W\begin{bmatrix}h_t\\ h_s\end{bmatrix} + b\right)\right] + k; \qquad \in \mathbb{R}^{3072\times T} \tag{1}$$

The "*Split+concat*" (shown within **CA Block** in Fig. 1), reshapes the self-attention scores M into $\tilde{M}$ to ensure that channel information across the two hidden representations is aligned along the channel dimension. A softmax is then applied channel-wise across time to obtain normalized scores C.

$$\tilde{M} = \begin{bmatrix} M_{1:1536,1:T} & M_{1537:3072,1:T}\end{bmatrix}; \qquad \in \mathbb{R}^{1536\times 2T} \tag{2}$$

$$C = softmax(\tilde{M}); \qquad \in \mathbb{R}^{1536\times 2T} \tag{3}$$

The original hidden representations h_t and h_s, are stacked time-wise ("Concat-T" within **CA Block** in Fig. 1) and the normalized attention scores C, which represent the importance of each frame given the channel, are used to compute mean ($\tilde{\mu}$) and standard deviation ($\tilde{\sigma}$).

$$\tilde{h} = \begin{bmatrix} h_t & h_s\end{bmatrix}; \qquad \in \mathbb{R}^{1536\times 2T} \tag{4}$$

$$\tilde{\mu} = \sum_{t=1}^{2T} C \odot \tilde{h}; \qquad \in \mathbb{R}^{1536} \tag{5}$$

$$\tilde{\sigma} = \sqrt{\sum_{t=1}^{2T} C \odot \tilde{h}^2 - \tilde{\mu}^2}; \qquad \in \mathbb{R}^{1536} \tag{6}$$

The estimated $\tilde{\mu}$ and $\tilde{\sigma}$ are concatenated and passed on to the linear layer to produce 192-dimensional speaker embeddings, which are then used to perform speaker identification. The **CA block** increases ECAPA-TDNN parameters by $\sim$1.5M (i.e., 20.8M $\rightarrow$ 22.3M).

2.2. K-means-based Modified Elastic Weight Consolidation (K-EWC)

In EWC [16], the importance of each model parameter for the source task is estimated using the Fisher Information Matrix (FIM). During adaptation, a regularization term that penalizes changes to parameters that are deemed important is added to the loss function, thereby preserving performance on the source task. Formally, if θ represents the model parameters and θ^* the optimal parameters for the source task, the EWC loss for a target task is given by:

$$\mathcal{L}_{\text{EWC}} = \sum_i F_i(\theta_i - \theta_i^*)^2 \tag{7}$$

where, F_i is the diagonal element of the FIM corresponding to the parameter θ_i.

In zero-source data scenario, direct computation of FIM is infeasible. Intuitively, the parameters of the pretrained-model define a probability distribution over the source-domain embedding space, and any representation extracted from this set of unadapted parameters will lie within this distribution. Consequently, embeddings for target-data produced by the pretrained-model, i.e., PMEs, provide an approximate view of this source-domain distribution, which we leverage to estimate parameter importance during adaptation. Motivated by this hypothesis, we propose K-means-based modified EWC (K-EWC). Let D_{Tar} denote the target-data. As shown in Fig. 1 (b) we extract PMEs (output of the "Source Task Model") for D_{Tar}. Since EWC requires class labels to estimate FIM, we perform K-means clustering on PMEs to generate pseudo-class labels. These pseudo-labels allow us to approximate the FIM element F_i for each parameter θ_i. Once the FIM is estimated, we incorporate it into the EWC loss as in Eq. 7. Thus, the proposed approach circumvents the need for source-data in the estimation of FIM.

2.3. Moment-Matching (MM) Regularization

To further mitigate catastrophic forgetting during adaptation, we add a moment-matching (MM) regularization term to the training loss. The central intuition behind MM regularization is that, during adaptation, the evolving target-domain embeddings should remain anchored within source-domain distribution, while still adapting to capture target-domain specific speaker-discriminative characteristics. If the evolving target-domain embeddings shift to an entirely different space, the model risks forgetting source-domain knowledge. Following our hypothesis from K-EWC, we use PMEs (output of the "Source Task Model" in Fig. 1 (b)) to obtain an approximate (pseudo) representation of the source-domain distribution. These PMEs are used to fit a Gaussian Mixture Model (GMM), where the number of components is selected based on the Akaike Information Criterion (AIC). The GMM provides an estimate of the mean vector μ_T and covariance matrix Σ_T for the pseudo source-domain (teacher) distribution. During fine-tuning, the learned target-data embedding distribution produced by the model are regularized to match the pseudo source-domain distribution statistics. The moment-matching regularization term is defined as:

$$\mathcal{L}_{\text{MM}} = \|\mu_T - \mu_S\|_2^2 + \|\Sigma_T - \Sigma_S\|_F^2 \tag{8}$$

where μ_S and Σ_S are the mean vector and covariance matrix of the learned target-data (student) embeddings in a given batch, $\|\cdot\|_2$, $\|\cdot\|_F$ denote the L2 and Frobenius norms, respectively.

2.4. Overall training objective

The overall training objective is given in Eq. 9, where $\mathcal{L}_{\text{SID}}$ is the speaker identification loss, λ_{EWC}, and λ_{MM} control the weights of the $\mathcal{L}_{\text{EWC}}$ and $\mathcal{L}_{\text{MM}}$ regularization losses, respectively.

$$\mathcal{L} = \mathcal{L}_{\text{SID}} + \lambda_{EWC} \cdot \mathcal{L}_{\text{EWC}} + \lambda_{\text{MM}} \cdot \mathcal{L}_{\text{MM}} \tag{9}$$

3. EXPERIMENTAL DETAILS

3.1. Databases

We perform adaptation experiments on two target domain databases: MyST [27], and CNCeleb [28]. In MyST, we use only the annotated portion, totaling approximately 268 hours. We adopt the training splits MyST-1 through MyST-4, and employ the same evaluation set used in [29] for consistency. The CNCeleb dataset contains 273 hours of speech from 997 speakers across diverse Chinese open-media sources. Similar to the MyST setup, we partition the CNCeleb training data into four subsets (CNCeleb-1 through CNCeleb-4), where CNCeleb-1 is the smallest and CNCeleb-4 corresponds to the biggest training set. In both databases, we use only the 3 sec long audio segments in the train set, while discarding audio files less than 3 seconds long as done in [30]. Hence, CNCeleb-4 has only 162 hours of data from 781 speakers. For evaluation, we use the standard CNCeleb evaluation set. The train set statistics for both MyST and CNCeleb are given in Table 1. For reporting VoxCeleb results we use the standard VoxCeleb-O evaluation set.

Table 1. *The train set splits of MyST and CNCeleb databases used for training/fine-tuning the models are presented. #Spks refers to the number of speakers, #Hrs is the duration in hours*

Split	MyST		CNCeleb	
	# Spks	# Hrs	# Spks	# Hrs
1	1210	2.00	781	1.94
2	1210	8.00	781	10.00
3	1210	85.00	781	98.55
4	1210	268.00	781	162.20

3.2. System Configurations

We use ECAPA-TDNN [26] as the underlying neural network architecture and the Speechbrain [30] toolkit. We refer to the models trained from scratch on the target domain datasets: MyST and CNCeleb datasets as *Baseline* in this paper. ECAPA-TDNN model pretrained on VoxCeleb 1 and 2 [32, 33] is used as the pretrained model in all adaptation experiments. Fine-tuning the pretrained model with the target data for a fixed number of epochs is referred to as *Finetune*. The input features for all the models were 80-dimensional filter bank features extracted with a frame length of 25 ms and a hop size of 10 ms. All experiments use five-fold data augmentation from [30]. Scoring uses cosine similarity; performance is measured by Equal Error Rate (EER).

Optimization was performed using the Adam optimizer with a learning rate of 10^{-3} and a weight decay of 2×10^{-6}, along with a cyclic learning rate scheduler to balance exploration and exploitation, with a base learning rate of 10^{-8}, a maximum learning rate of 10^{-3}, and a batch size of 8. Experiments were conducted on a single NVIDIA GeForce RTX 2080 Ti GPU. Training epochs for the four MyST splits (MyST-1 to MyST-4) were 15,15,2, and 1, respectively, while those for the CNCeleb splits (CNCeleb-1 to CNCeleb-4) were 10,10,2, and 2. For larger splits, the number of epochs was reduced due to computational resource constraints.

The K-EWC regularization loss weight λ_{EWC} was empirically fixed at 100 in all experiments. For K-means clustering, we used embeddings of the target-domain data extracted from the pretrained model, i.e., PMEs, with only 320 randomly selected target samples and 50 clusters. This choice provided a good trade-off between efficiency and minimal computational overhead across experiments. For the MM regularization loss, the pseudo source-domain distribution was approximated by fitting a Gaussian Mixture Model (GMM) to the PMEs extracted using the MyST-2 and CNCeleb-2 subsets for all MyST and CNCeleb experiments, respectively. The optimal number of components of the GMM was determined by the Akaike Information Criterion (AIC). These subsets (about 10h of data each) were chosen because larger subsets would incur significant computational cost, while smaller ones yielded poor approximations. The MM regularization loss weight λ_{MM} was scheduled to increase linearly from 0.2 to 2.0 over training epochs, ensuring stronger regularization as the model became more adapted to the downstream task. This schedule helps prevent the adapted embeddings from drifting too far from the pseudo source-domain representations. MM losses were accumulated over four batches prior to backpropagation. In experiments combining CA with K-EWC and/or MM, the teacher branch of the Siamese network was always frozen, while the MM and K-EWC regularizations were applied only to the student branch.

4. RESULTS AND DISCUSSION

For reference, no-adaptation (VoxCeleb pretrained) model EERs are: Vox-O 0.89%, MyST 17.48%, CNCeleb 15.26%. Experimental results from our experiments are given in Table 2. *Baseline* corresponds to training exclusively on target-domain data, while *Finetune* initializes the model with pretrained VoxCeleb weights and optimizes using only the speaker identification loss on target-domain data. In Table 2, rows 5 (+*MM*) and 6 (+*K-EWC*) denote adding the respective regularization term to *Finetune*. *CA* refers to the proposed Cross-Attentive adapter, and the addition of regularization terms to *CA* is similarly indicated with a "+" in rows 8, 9, and 10. For comparison with prior work, we also include G-IFT [29], an adapter-based approach, and Sparse Filterbank [31]. For MyST, the G-IFT approach is directly comparable to our methods, as it also employs an adapter-based architecture and uses a MyST training setup comparable to ours In contrast, for CNCeleb, the Sparse Filterbank sys-

Table 2. *Model performance across the proposed approaches in terms of Equal Error Rate (%EER). Results are reported as Target/Source %EER. The target training set duration increases progressively from Split-1 to Split-4. For each domain, the evaluation set is fixed across all splits (i.e., the same MyST evaluation set and the same CNCeleb evaluation set). The best %EERs from our models (rows 4–10) are boldfaced, with "*" denoting statistically significant improvements over Finetune (row 4) based on a paired t-test at $p = 0.05$.*

		MyST/VoxCeleb-O				CNCeleb/VoxCeleb-O			
	Method	**MyST-1**	**MyST-2**	**MyST-3**	**MyST-4**	**CNCeleb-1**	**CNCeleb-2**	**CNCeleb-3**	**CNCeleb-4**
1	**Baseline**	21.84/28.97	12.72/18.08	10.07/21.10	8.64/21.18	32.06/25.16	19.51/15.77	17.13/13.23	16.17/11.76
2	**G-IFT [29]**	14.22/8.00	7.03/8.94	5.49/12.51	5.42/11.23	-	-	-	-
3	**Sparse FilterBank [31]**	-	-	-	-	-	-	-	12.25/10.81
4	**Finetune**	20.3/7.73	7.22/7.60	5.67/8.39	**5.57**/14.91	17.96/6.05	10.00/3.52	10.57/4.63	10.98/6.52
5	+ ***MM***	22.58/6.92	7.64/7.06	**5.54***/11.14	5.59/15.57	18.6/5.33	10.19/3.56	10.89/5.23	10.94/6.16
6	+ ***K-EWC***	21.00/5.49	8.86/6.00	7.41/7.18	7.26/7.76	**14.61*/2.04***	10.40/1.90	10.20/2.63	10.26/2.85
7	**CA**	**16.53***/4.16	**6.93***/5.24	5.56/6.55	5.78/11.15	16.99/4.01	10.14/3.22	10.36/3.96	10.56/3.97
8	+ ***MM***	16.79/4.42	7.04/5.51	5.75/6.56	5.70/10.05	16.96/4.17	10.16/3.24	10.50/4.14	10.44/3.94
9	+ ***K-EWC***	17.29/3.89	7.64/3.11	6.54/3.83	6.28/**4.39***	15.47/3.02	9.96/**1.80***	**10.14*/2.45***	10.27/2.73
10	+ ***K-EWC*** + ***MM***	17.29/**3.88***	7.58/**3.01***	6.57/**3.74***	6.37/4.40	15.46/3.03	**9.95***/1.81	10.19/2.52	**10.17*/2.72***

tem is based on ECAPA-TDNN and trained on the same CNCeleb corpus (approximately 800 speakers), but its primary objective was not to address catastrophic forgetting. We report its results not as a direct benchmark, but to provide context on performance range of CNCeleb-based systems. Results on MyST versus VoxCeleb are reported in [34], but direct comparison is avoided due to different splits and speaker overlap.

4.1. Baseline Experiments

As expected, our results demonstrate that *Finetune* improves performance over *Baseline* approach. We also observe a direct correlation between target domain training data size and source domain performance degradation. This behavior is consistent with CF theory, where exposure to larger amounts of target data during adaptation leads to greater interference with previously learned representations. A comparative analysis between corresponding data splits (i.e., MyST-1 vs CNCeleb-1, MyST-2 vs CNCeleb-2, etc) reveals domain-dependent degradation patterns. Specifically, MyST experiments exhibit higher VoxCeleb EER degradation compared to CNCeleb counterparts. This disparity can be attributed to the substantial domain mismatch between MyST (child speech) and the source VoxCeleb model (adult speech), whereas CNCeleb maintains adult speech characteristics, resulting in reduced domain shift and consequently lower source domain performance degradation. We establish *Finetune* as our primary baseline for our experiments.

4.2. MM and K-EWC Regularization

Adding the *MM* regularization to *Finetune* yields mixed outcomes. For MyST, *MM* helps preserve VoxCeleb performance in the low-resource case (MyST-1 and MyST-2) but significantly degrades VoxCeleb %EER in MyST-3 and MyST-4, likely because these larger subsets were trained for fewer epochs, preventing the *MM* regularization from adequately pulling embeddings toward the pseudo-source distribution. In CNCeleb, *MM* improves in some cases (splits 1 and 4), is neutral in others, and slightly harmful in split 3, suggesting its effect depends on both domain mismatch and training dynamics. By contrast, *K-EWC* consistently improves VoxCeleb %EER across all splits in both datasets. This uniform effectiveness highlights the *K-EWC* method's robustness to varying training data sizes and domain characteristics. However, both regularization methods generally come at the cost of reduced target-domain accuracy.

4.3. CA Adapter Analysis

The *CA* approach demonstrates superior target domain performance compared to *Finetune* and its regularized variants (+*MM* and +*K-EWC*) for MyST-1 through MyST-3 configurations. Additionally, *CA* mitigates source domain performance degradation in these scenarios. However, this advantage diminishes in high-resource conditions (MyST-4), where source domain EER deteriorates to 11.15%, contrasting with *Finetune*+*K-EWC*'s ability to maintain 7.76% EER, albeit at the cost of target domain performance (7.26% vs 5.78%) - refer rows 6 and 7 of MyST-4. For CNCeleb experiments, *CA* outperforms standard *Finetune* in target domain performance across three of four training splits while consistently reducing source domain degradation. Yet, because CNCeleb and VoxCeleb are both adult speech datasets, *Fintune*+*K-EWC* (row 6) proves particularly effective, i.e., finetuning with K-EWC often surpasses CA in preserving both source and target performance.

Adding *K-EWC* to *CA* further mitigates source-domain forgetting, particularly in MyST, though with minor reductions in target-domain performance compared to CA alone. In CNCeleb, where the domain mismatch is minimal, *CA*+*K-EWC* improves both target- and source-domain performance across all splits, clearly outperforming *CA* alone. Incorporating *MM* alongside *K-EWC* (row 10) yields modest additional gains in a few cases (e.g., MyST-2, MyST-3, CNCeleb-4) but otherwise maintains similar performance. This indicates that while MM holds promise, it may require further tuning and more careful integration to deliver consistent benefits.

In general, for a given CNCeleb train split, the same approach often achieves the best performance on both source and target test sets, showing that target gains need not come at the expense of the source domain. In contrast, for a given MyST train split, the method that best improves target performance is not always the one that minimizes source degradation, highlighting a stronger trade-off in the challenging *cross-age* adaptation scenario. In comparison to G-IFT [29], which is another adapter-based approach, our experiments show that while G-IFT achieves stronger performance on the target-domain task, it suffers from a degradation in source-domain accuracy. We included Sparse Filter bank [31] to provide a sense of the performance range for CNCeleb based system.

5. CONCLUSION

We propose a CA adapter along with K-EWC and MM regularization strategies to address catastrophic forgetting in speaker verification domain adaptation under zero source-data access. Our approaches consistently succeed in reducing source-domain performance degradation across different target-domain train data sizes (in hours). This demonstrates that our methods effectively mitigate catastrophic forgetting, by providing a more stable balance between learning from target data and retaining pretrained model knowledge. As a direction for future work, we aim to extend these approaches beyond ECAPA-TDNN to assess their generalizability across a wider range of speaker verification frameworks.

6. ACKNOWLEDGMENTS

This research is supported in part by the NSF and the IES, U.S. Department of Education (DoE), through Grant R305C240046 to the U. at Buffalo. The opinions expressed are those of the authors and do not represent views of the NSF, IES, or the DoE.

7. REFERENCES

[1] M. McCloskey and N. J. Cohen, "Catastrophic interference in connectionist networks: The sequential learning problem," *Psychology of learning and motivation*, 1989.

[2] R. M. French, "Catastrophic forgetting in connectionist networks," *Trends in Cognitive Sciences*, 1999.

[3] S. Ruder, "An overview of multi-task learning in deep neural networks," *arXiv preprint arXiv:1706.05098*, 2017.

[4] S. Vander Eeckt and H. Van Hamme, "Using adapters to overcome catastrophic forgetting in end-to-end automatic speech recognition," in *ICASSP*, 2023.

[5] A. A. Rusu, N. C. Rabinowitz, G. Desjardins, et al., "Progressive neural networks," *arXiv preprint arXiv:1606.04671*, 2016.

[6] J. Yoon, E. Yang, J. Lee, and S. J. Hwang, "Lifelong learning with dynamically expandable networks," in *ICLR*, 2018.

[7] A. Mallya and S. Lazebnik, "Packnet: Adding multiple tasks to a single network by iterative pruning," in *CVPR*, 2018.

[8] A. Robins, "Catastrophic forgetting, rehearsal and pseudorehearsal," *Connection Science*, 1995.

[9] R. Ratcliff, "Connectionist models of recognition memory: constraints imposed by learning and forgetting functions.," *Psychological review*, 1990.

[10] T. Xu, K. Huang, P. Guo, et al., "Towards Rehearsal-Free Multilingual ASR: A LoRA-based Case Study on Whisper ," in *Interspeech*, 2024.

[11] S. Vander Eeckt et al., "Rehearsal-free online continual learning for automatic speech recognition," *Interspeech*, 2023.

[12] Z. Li and D. Hoiem, "Learning without forgetting," *TPAMI*, 2017.

[13] H. Shin, J. K. Lee, J. Kim, and J. Kim, "Continual learning with deep generative replay," *NeurIPS*, 2017.

[14] C. V. Nguyen, Y. Li, T. D. Bui, and R. E. Turner, "Variational continual learning," in *ICLR*, 2018.

[15] C. Wu, L. Herranz, X. Liu, et al., "Memory replay gans: Learning to generate new categories without forgetting," *NeurIPS*, 2018.

[16] J. Kirkpatrick, R. Pascanu, N. Rabinowitz, et al., "Overcoming catastrophic forgetting in neural networks," *PNAS*, 2017.

[17] F. Zenke, B. Poole, and S. Ganguli, "Continual learning through synaptic intelligence," in *ICML*, 2017.

[18] A. Krogh and J. Hertz, "A simple weight decay can improve generalization," *NeurIPS*, 1991.

[19] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *NIPS*, 2014.

[20] N. Monaikul, G. Castellucci, S. Filice, and O. Rokhlenko, "Continual learning for named entity recognition," in *AAAI*, 2021.

[21] Y. Xu, X. Zhong, A. J. Jimeno Yepes, and J. H. Lau, "Forget me not: Reducing catastrophic forgetting for domain adaptation in reading comprehension," in *IJCNN*, 2020.

[22] S. Kar, G. Castellucci, S. Filice, S. Malmasi, et al., "Preventing catastrophic forgetting in continual learning of new natural language tasks," in *KDD*, 2022.

[23] S. Niu, J. Wu, Y. Zhang, et al., "Efficient test-time model adaptation without forgetting," in *ICML*, 2022.

[24] M. Honarmand, O. C. Mutlu, P. Azizian, et al., "Selective test-time domain adaptation using fisher information for robust facial expression recognition in-the-wild," in *CVPR*, 2025.

[25] J. Zheng, V. M. Shetty, N. B. Shankar, and A. Alwan, "An age-agnostic system for robust speaker verification," *WOCCI, Interspeech*, 2025.

[26] B. Desplanques, J. Thienpondt, and K. Demuynck, "ECAPA-TDNN: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification," in *Interspeech*, 2020.

[27] W. Ward, R. Cole, D. Bolaños, et al., "My science tutor: A conversational multimedia virtual tutor.," *Journal of Educational Psychology*, 2013.

[28] Y. Fan, J. W. Kang, L. T. Li, et al., "CN-CELEB: A challenging chinese speaker recognition dataset," in *ICASSP*, 2020.

[29] V. M. Shetty, J. Zheng, and A. Alwan, "G-IFT: A gated linear unit adapter with iterative fine-tuning for low-resource children's speaker verification," *WOCCI, Interspeech*, 2025.

[30] M. Ravanelli, T. Parcollet, P. Plantinga, et al., "Speechbrain: A general-purpose speech toolkit," *arXiv preprint arXiv:2106.04624*, 2021.

[31] J. Peng and R. Gu and L. Mošner and others, "Learnable Sparse Filterbank for Speaker Verification," in *Interspeech*, 2022.

[32] A. Nagrani, J. S. Chung, and A. Zisserman, "Voxceleb: A large-scale speaker identification dataset," in *Interspeech*, 2017.

[33] J. S. Chung, A. Nagrani, and A.Zisserman, "Voxceleb2: Deep speaker recognition," in *Interspeech*, 2018.

[34] S. Tabatabaee, J. Liu, and C. Espy-Wilson, "FT-Boosted SV: Towards Noise Robust Speaker Verification for English Speaking Classroom Environments ," in *Interspeech*, 2025.