G-IFT: A Gated Linear Unit adapter with Iterative Fine-Tuning for Low-Resource Children's Speech Verification

Vishwas Shetty, Jiusi Zheng*, Abeer Alwan*

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

shettyvishwas@ucla.edu, zheng94@ucla.edu, alwan@ee.ucla.edu

Abstract

Speaker Verification (SV) systems trained on adults speech often underperform on children's SV due to the acoustic mismatch, and limited children speech data makes fine-tuning not very effective. In this paper, we propose an innovative framework, a Gated Linear Unit adapter with Iterative Fine-Tuning (G-IFT), to enhance knowledge transfer efficiency between the high-resource adults speech domain and the low-resource children's speech domain. In this framework, a Gated Linear Unit adapter is first inserted between the pre-trained speaker embedding model and the classifier. Then the classifier, adapter, and pre-trained speaker embedding model are optimized sequentially in an iterative way. This framework is agnostic to the type of the underlying architecture of the SV system. Our experiments on ECAPA-TDNN, ResNet, and X-vector architectures using the OGI and MyST datasets demonstrate that the G-IFT framework yields consistent reductions in Equal Error Rates compared to baseline methods.

Index Terms: Speaker Verification, Children's Speech, Domain Adaption, Low-Resource Scenario

1. Introduction

With the rise of technological innovations, children increasingly engage with social media and e-learning platforms. While these tools foster development, they also raise security concerns, necessitating protective measures. Advancements in voice-based Children's Speaker Verification (C-SV) systems are a crucial step in ensuring children's online safety [1]. These systems aim to accurately verify a child's identity based on their voice, addressing security concerns and ensuring personalized, safe user experiences.

Compared to other domain adaptation problems, a major challenge in C-SV arises from the constraints imposed by the scarcity of children's speech data. Although Speaker Verification (SV) systems [2, 3, 4] trained on adult speech have achieved notable success across various datasets [5, 6], they often face challenges when applied to children speech due to the significant acoustic differences between adult and children speech [7]. Fine-tuning pre-trained models initially trained on adult speech with children speech datasets is one of the widely adopted strategies. However, fine-tuning pre-trained verification models with limited children's speech data may not adequately address the acoustic mismatch due to the scarcity of children's speech resources, resulting in inefficient knowledge transfer and hence subpar adaptation to the target domain [8, 9].

To address the data scarcity challenge in C-SV, researchers have primarily adopted two strategies. The first is out-of-domain data augmentation [10, 11, 12, 13], which supplement existing child speech data by adding perturbed adult speech data so that its acoustic properties resemble children's speech. This includes methods such as perturbing adult speech by modifying parameters like speaking rate, pitch, duration, and vocal tract length, as well as leveraging cycle-consistent GAN-based voice conversion to transform adult speech into child-like speech [11, 12, 13].

Another strategy involves model-level adaptations, including optimizing model architectures [14], designing more effective loss functions [15], and fine-tuning of the pre-trained adult SV model using children speech [16]. Adapter-based fine-tuning, achieved by integrating lightweight adapter modules into pre-trained speaker embedding models, effectively enhances the efficiency of knowledge transfer compared to the baseline methods. Although adapter-based approaches have shown promise in domain adaptation tasks [17, 18, 19, 20, 21, 22], to the best of our knowledge, no prior work has explored their use for C-SV, where data scarcity presents greater challenges and distinguishes it from other adaptation tasks such as cross language adaptation [23].

In this paper, we propose a novel framework, Gated Linear Unit adapter (GLU) with Iterative Fine-Tuning, which we refer to as G-IFT, to enhance knowledge transfer efficiency between the adults speech domain and the children speech domain. We first propose a novel adapter module using GLU to fine-tune the speaker embedding models trained on adult speech, alleviating the domain shifting problem in C-SV. Additionally, we introduce a novel iterative fine-tuning strategy, which builds upon the GLU adapter fine tuning, wherein different components of the model are alternatively fine-tuned to enhance the SV accuracy. Experimental results indicate that our method amplifies the positive impact of adapter fine-tuning particularly in the low resource scenario, achieving a further reduction in the EER of the SV systems. The proposed G-IFT framework also exhibits potential for broader applications in other low-resource SV tasks, including verification tasks for individuals with speech disorders and similar under-resourced scenarios.

The following sections are structured as follows: Section 2 introduces the GLU adapter and iterative fine-tuning in the G-IFT framework. Section 3 covers the experimental setup. Section 4 presents the results of the proposed approach under varying conditions. Section 5 concludes and suggests future work.

2. Methods

2.1. Gated Linear Unit adapters

We first propose a novel adapter structure inspired from the Gated Linear Unit (GLU) mechanism [24]. The motivation behind using a GLU layer was to allow the network to decide how

*These authors contributed equally to this work.

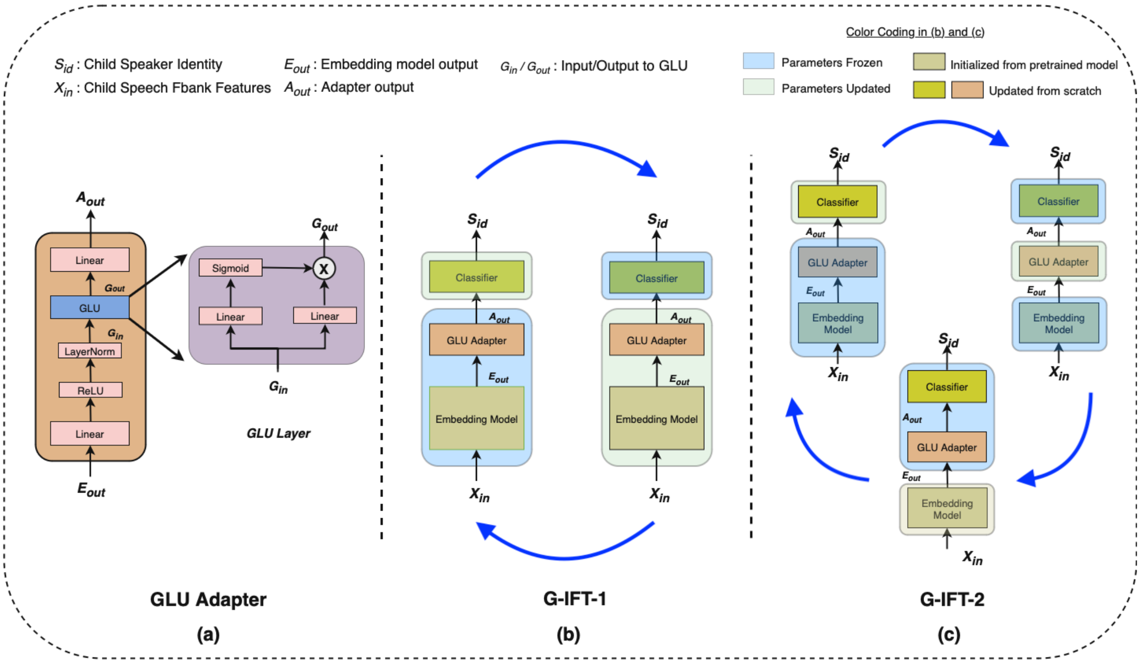

Figure 1: *An overview of the proposed G-IFT framework. The GLU adapter architecture is shown in (a), and illustration of the steps involved in the G-IFT-1 and G-IFT-2 methods (discussed in Section 2.2) are shown in (b) and (c) respectively.*

much information should flow through the adapter, i.e., gating weights based on the adaptation data. The GLU layer learns the gating weights with the help of two linear layers and a sigmoid operation. The GLU adapter structure used in this paper is shown in Figure 1 (a). Firstly, the output of the speaker embedding model, E_{out} is fed into the adapter module, passing through a linear layer followed by a ReLU activation function. Then after performing layer normalization, G_in is passed through the GLU layer. Finally, the output of GLU layer, G_out is passed through a linear layer before being passed on to perform speaker identification by the classifier. The operation performed by GLU layer is given in Equation (1). W, V, b, and c are learnable parameters. In GLU layer, the sigmoid operation on the output of one of the linear layers acts as the weights to the output of the second linear layer, deciding how much information from the pre-trained embedding model should be passed on ahead.

$$G_{out} = (G_{in} * W + b) \otimes \sigma(G_{in} * V + c) \qquad (1)$$

The proposed adapter is inserted between the output of the speaker embedding model and the input of the classifier. When evaluating the proposed GLU adapter based approach, the outputs of the adapter module, denoted as A_{out} in Figure 1, are utilized as speaker embeddings.

2.2. G-IFT: GLU adapter with Iterative Fine-Tuning

Fine-tuning pretrained verification models in SV tasks differs from fine-tuning in other tasks such as Automatic Speech Recognition (ASR) due to the need for adjusting the classifiers based on the number of speakers in the target domain dataset. This requires the classifier's parameters to be trained from scratch, rather than relying solely on transfer learning. Traditional adapter fine-tuning methods commonly update the adapter and classifier simultaneously; in some cases, the adapter, classifier, and embedding model are all updated simultaneously. However, since the classifier and the inserted adapters typically require more substantial adjustments than the pretrained embedding model, this approach can be inefficient especially in the low-resource scenario, potentially slowing down convergence and reducing overall performance.

To enhance the efficiency of knowledge transfer, we propose a novel framework, Gated Linear Unit Adapter with Iterative Fine-Tuning (G-IFT), which applies an iterative fine-tuning strategy to the GLU adapter. The motivation behind this approach is to address potential inefficiencies associated with simultaneous updates by allowing more focused updates to the adapter and classifier initially, which helps them adapt more effectively to in-domain children speech data. This framework aims to make more efficient use of target data during fine-tuning, especially in the low-resource scenario, which leads to better performance in C-SV. Additionally, this method is highly adaptable and is not constrained by the underlying model architecture.

The proposed G-IFT framework includes two variants, G-IFT-1 and G-IFT-2, which share the same foundational structure but differ in their iterative fine-tuning strategies as shown in Figure 1 (b) and (c). In G-IFT-1, the process involves three main steps. First, the model is initialized and trained on the source domain dataset to obtain a pre-trained speaker embedding model. Second, the GLU adapter is inserted between the pretrained speaker embedding model and a classifier designed for fine-tuning on the target domain dataset. Third, the training alternates iteratively between fine-tuning the GLU adapters and the classifier jointly, followed by fine-tuning the pretrained speaker embedding model. G-IFT-2 follows the same first two steps as G-IFT-1 but introduces a slightly different iterative strategy. In this variant, the process alternates iteratively by first fine-tuning the classifier, then the GLU adapters, and then the pretrained speaker embedding model. This adjustment allows

Table 1: *The train-eval splits of the OGI and MyST databases used for training/fine-tuning the models are presented. #Spks refers to the number of speakers, #Hrs to the duration in hours, and #Trials to the number of SV trials in the evaluation set.*

	Train		Eval	
	# Spks	# Hrs	# Spks	# Trials
OGI	120	2.65	993	190972
MyST-1	1210	2.00	91	100000
MyST-2	1210	8.00	91	100000
MyST-3	1210	85.00	91	100000
MyST-4	1210	268.00	91	100000

G-IFT-2 to prioritize classifier adaptation in isolation before integrating adjustments from the adapters and embedding model.

3. Experimental Settings

3.1. Databases

We conduct experiments on two children's speech databases, OGI [25] and MyST [26]. For OGI, we use the scripted portion, which contains children's speech from approximately 100 speakers per grade, spanning kindergarten to grade 10. The train/eval split follows the protocol established in [27]. The MyST dataset consists of 499 hours of speech data collected from 1,372 students in grades 3 to 5. For our experiments we utilize only the annotated portion totaling approximately 268 hours. For MyST, we also create multiple smaller training subsets from the MyST training set, maintaining a consistent evaluation split to analyze the impact of in-domain training data size on the proposed method. While creating the MyST train splits MyST-1 to MyST-4, we ensure that data from all the speakers in the complete training set is included. Hence, as we go from MyST-1 to MyST-4, the amount of speech data per speaker increases. Across all train and evaluation splits in both datasets, there is no overlap in speakers. It should be noted that the MyST evaluation set is the same, irrespective of the MyST train split. The train eval splits of both databases are shown in Table 1.

3.2. Baseline Systems

In this study, we utilized ECAPA-TDNN [2], ResNet [28], and X-vector [29] as the SV systems. All experiments were conducted using the SpeechBrain toolkit [30]. Models trained from scratch on the OGI and MyST datasets served as the baseline systems and are referred to as *Baseline*. Pretrained speaker embedding models are trained on the VoxCeleb dataset [5]. The experiment where these pretrained models are directly tested on children speech data is referred to as *Pretrained*. We refer to fine-tuning the *Pretrained* model with child speech data for a fixed number of epochs as vanilla fine-tuning and this model is referred to as *Finetune* in this paper. 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. Equal Error Rate was used to evaluate the SV systems' performance.

Optimization was performed using the Adam optimizer with a learning rate of 0.001 and a weight decay of 0.000002, along with a cyclic learning rate scheduler to balance exploration and exploitation, with a base learning rate of 1×10^{-8}, a maximum learning rate of 0.001, and a step size of 65,000. Notably, the number of training epochs varied across different methods due to differences in fine-tuning strategies. Specifically, the vanilla fine-tuning method updates all model parameters, including both the speaker embedding model and classifier, simultaneously within each training epoch. In contrast, the G-IFT framework adopts an iterative strategy, alternately fine-tuning the classifier, GLU adapters, and the pretrained speaker embedding model in a iterative manner. As a result, the total number of epochs for the GIFT-1 method was twice that of the vanilla fine-tuning method, and for the GIFT-2 method, it was three times that of the vanilla fine-tuning method. This adjustment ensures that the number of parameter updates across all methods is equivalent.

Table 2: *Performance in terms of Equal Error Rate (EER %) comparison across OGI and MyST-1 datasets.* ***Pretrained*** *is the open source adult SV systems tested directly on children speech data.* ***Baseline*** *is the SV systems trained using children speech data from scratch, and* ***Finetuned*** *is the adult SV systems finetuned using children speech data.* ***GLU*** *is adapter fine tuning,* ***G-IFT-1****, and* ***G-IFT-2*** *are the proposed adapter fine tuning framework.* ***RA*** *is the Residual Adapter [9]. Significant improvements (paired t-test; p-value=0.05) over Finetune are represented with "*".*

	ECAPA-TDNN		**ResNet**		**X-vector**	
	OGI	MyST-1	OGI	MyST-1	OGI	MyST-1
Pretrained	17.39	17.48	12.62	14.47	23.59	24.95
Baseline	12.15	24.80	14.23	36.23	14.71	25.70
Finetune	11.10	20.04	7.45	10.64	15.89	21.26
RA [9]	11.34	22.89	7.47	12.76	15.61	21.54
GLU	9.70	22.30	6.92	12.48	12.31	17.96
G-IFT-1	**8.88***	16.38	**6.77***	9.29	12.71	18.20
G-IFT-2	9.06	**14.22***	6.88	**8.37***	**12.09***	**16.79***

4. Results and Discussion

4.1. Performance of the G-IFT framework

Table 2 presents results on the OGI and MyST-1 datasets using various training methods. The proposed G-IFT framework, including G-IFT-1 and G-IFT-2, consistently improves performance across all architectures. Rows 2 and 3 show baseline and vanilla fine-tuning results. Fine-tuning generally outperforms training from scratch, except for X-vector on OGI. Row 4 reports results of Residual Adapter (RA) fine-tuning, configured to match our GLU adapter in parameter size and training steps (15 epochs). Although RA performed well in [9], it is less effective here. In contrast, direct fine-tuning with the GLU adapter (without iteration) outperforms vanilla fine-tuning in four of six cases. The last two rows show that G-IFT-1 and G-IFT-2 consistently outperform vanilla fine-tuning in all test cases. Compared to GLU fine-tuning, one of the G-IFT variants always achieves the best result, validating the benefit of iterative strategies in enhancing GLU adapter adaptability under limited data. We also examined Iterative Fine Tuning (IFT) alone on MyST-1, which yields EERs of 22.5%, 10.2%, and 20.9% for ECAPA-TDNN, ResNet, and x-vector, without clear improvement. In contrast, applying IFT to GLU reduces EERs to 14.2%, 8.4%, and 16.8%, The adapter-classifier setup provides greater modeling capacity.

4.2. Impact of Training Data Size on Model Adaptability

To further investigate the impact of training dataset size on the proposed G-IFT framework, we use the four MyST subsets

Table 3: *Equal Error Rate (EER %) across different models—ECAPA-TDNN, X-vector, and ResNet—on various MyST training splits. The duration of the training splits increases progressively from MyST-1 to MyST-4. The evaluation results are obtained using a consistent testing split, with no overlap in speakers between the training and test splits in all experiments. Significant improvements (paired t-test; p-value=0.05) over Finetune are represented with "*".*

	ECAPA-TDNN				ResNet				X-vector			
	MyST-1	MyST-2	MyST-3	MyST-4	MyST-1	MyST-2	MyST-3	MyST-4	MyST-1	MyST-2	MyST-3	MyST-4
Baseline	24.80	14.42	10.76	10.11	36.23	19.26	18.05	11.48	25.70	19.76	17.08	16.24
Finetune	20.04	7.97	5.81	6.05	10.64	9.33	6.60	4.87	21.26	19.16	16.70	16.11
RA [9]	22.89	8.19	6.13	6.07	12.76	8.62	5.70	4.87	21.54	19.50	16.36	17.83
GLU	22.30	7.51	5.92	5.97	12.48	7.48	5.50	5.07	17.96	15.05	**11.49***	**11.75***
G-IFT-1	16.38	**7.03**	**5.49***	**5.42***	9.29	6.42	**4.89***	**4.82**	18.20	15.37	12.29	12.58
G-IFT-2	**14.22***	7.81	5.93	5.99	**8.37***	**5.87***	5.52	5.27	**16.79***	**14.58***	14.70	15.30

(Table 1) to analyze the performance as training resources increased. As shown in Table 3, the EERs of all methods decrease overall with increasing training data. Comparing *Finetune* with *Baseline* in Table 3, ECAPA-TDNN and ResNet exhibit significantly better performance with fine-tuning using larger MyST datasets, while X-vector shows comparable results between the two approaches. This discrepancy can likely be attributed to differences in model complexity. The larger and more intricate architectures of ECAPA-TDNN and ResNet may better fit the children speech data when trained using larger amounts of in-domain data, whereas the simpler architecture of X-vector might not. Residual Adapter (RA) fine-tuning results are given in row 3 of Table 3. Additionally, GLU adapter i.e., row 4 of Table 3 performs better than RA in 11 out of 12 test cases.

In terms of the G-IFT framework, all three model architectures achieve lower EERs across different MyST training subsets compared to vanilla fine-tuning. Specifically, as shown in Figure 2, the absolute reduction in EER (%) obtained by G-IFT-1 and G-IFT-2 methods on the MyST-1 and MyST-2 datasets are more pronounced and consistent than those on MyST-3 and MyST-4 datasets across three architectures, highlighting the effectiveness of our proposed approach particularly in the low-resource scenarios. This phenomenon is likely due to our method's focus on iteratively fine-tuning the adapter and classifier first, which helps mitigate embedding overfitting. However, when sufficient training data is available, the large dataset naturally reduces embedding overfitting, making the efficiency of our method less critical and diminishing its advantage. We also observed that both G-IFT-1 and G-IFT-2 outperform vanilla fine-tuning on MyST-1 and MyST-2 across all three model architectures, though the best-performing method varied between the G-IFT-1 and G-IFT-2. However, with larger training datasets like MyST-3 and MyST-4, G-IFT-1 consistently achieved better results than G-IFT-2. A possible explanation is that in low-resource scenarios, sequentially fine-tuning individual components - the classifier , adapter and base model, as implemented in G-IFT-2 may be more effective. Conversely, under high-resource conditions, it appears sufficient to update relatively larger sub-modules, as in G-IFT-1, where the adapter and classifier are updated jointly followed by base model update. This hypothesis is further supported by the observation that vanilla fine-tuning also performs well in high-resource settings, as reflected in our results in Table 3. These observations prompt us to further investigate the respective strengths of G-IFT-1 and G-IFT-2 in future work, in order to refine the framework and propose a more universally effective approach applicable across diverse scenarios.

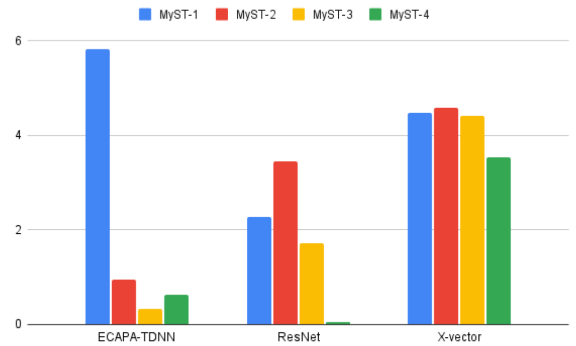

Figure 2: *Absolute reduction in EER (%) (y-axis) by the proposed G-IFT framework (Rows 4–5, Table 3) over Finetune (Row 2) on different MyST training splits. Larger reductions are observed in low-resource settings (MyST-1/2) than in high-resource ones (MyST-3/4), underscoring the framework's effectiveness in low-resource scenarios.*

5. Conclusion

In this paper, we proposed the G-IFT framework to alleviate the domain mismatch challenge when adapting adults (i.e., high-resource) SV models to perform children (i.e., low-resource) SV. Through the integration of the GLU adapter and an iterative fine-tuning strategy, the proposed framework achieved consistent performance improvements across three neural network architectures on the OGI and MyST datasets. The G-IFT framework outperforms baseline methods, achieving competitive performance using significantly lesser in-domain data. By varying amounts of in-domain training data using MyST dataset we evaluate the framework's performance under different resource constraints and our results highlight the potential of the proposed G-IFT framework in improving speaker verification systems in low-resource scenarios. Future work will focus on refining the G-IFT framework and extending its application to other low-resource SV tasks, including those involving disordered speech.

6. Acknowledgment

The 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 IES, DoE, or the NSF.

7. References

[1] J. H. Graafland, "New technologies and 21st century children: Recent trends and outcomes," *Organization for Economic Cooperation and Development Working Papers (OECD)*, 2018.

[2] B. Desplanques, J. Thienpondt, and K. Demuynck, "ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification," in *Interspeech*, 2020, pp. 3830–3834.

[3] T. Zhou, Y. Zhao, and J. Wu, "Resnext and res2net structures for speaker verification," in *IEEE Spoken Language Technology Workshop (SLT)*, 2021, pp. 301–307.

[4] C. yeong Lim, H. seo Shin, J. ho Kim, J. Heo, K.-W. Koo, S. bin Kim, and H.-J. Yu, "Improving noise robustness in self-supervised pre-trained model for speaker verification," in *Interspeech*, 2024, pp. 2665–2669.

[5] A. Nagrani, J. S. Chung, W. Xie, and A. Zisserman, "Voxceleb: Large-scale speaker verification in the wild," *Computer Speech & Language*, vol. 60, p. 101027, 2020.

[6] Y. Fan, J. Kang, L. Li, K. Li, H. Chen, S. Cheng, P. Zhang, Z. Zhou, Y. Cai, and D. Wang, "CN-Celeb: A challenging chinese speaker recognition dataset," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 7604–7608.

[7] E. T. Stathopoulos, J. E. Huber, and J. E. Sussman, "Changes in acoustic characteristics of the voice across the life span: Measures from individuals 4–93 years of age," *Journal of Speech, Language, and Hearing Research*, vol. 54, no. 4, pp. 1011–1021, 2011.

[8] T. Graave, Z. Li, T. Lohrenz, and T. Fingscheidt, "Mixed children/adult/childrenized fine-tuning for children's asr: How to reduce age mismatch and speaking style mismatch," in *Interspeech*, 2024, pp. 5188–5192.

[9] R. Fan and A. Alwan, "Draft: A novel framework to reduce domain shifting in self-supervised learning and its application to children's asr," in *Interspeech*, 2022, pp. 4900–4904.

[10] H. K. Kathania, V. Kadyan, S. R. Kadiri, and M. Kurimo, "Spectral warping based data augmentation for low resource children's speaker verification," *Multimedia Tools and Applications*, vol. 83, no. 16, pp. 48 895–48 906, 2024.

[11] S. Aziz and S. Shahnawazuddin, "Role of data augmentation and effective conservation of high-frequency contents in the context children's speaker verification system," *Circuits, Systems, and Signal Processing*, vol. 43, no. 5, pp. 3139–3159, 2024.

[12] S. Aziz, Ankita, and S. Shahnawazuddin, "Short-utterance-based children's speaker verification in low-resource conditions," *Circuits, Systems, and Signal Processing*, vol. 43, no. 3, pp. 1715–1740, 2024.

[13] S. Aziz and S. Shahnawazuddin, "Enhancing children's short utterance-based asv using inverse gamma-tone filtered cepstral coefficients," *Circuits, Systems, and Signal Processing*, vol. 43, no. 5, pp. 3020–3041, 2024.

[14] K. Radha, M. Bansal, and R. B. Pachori, "Automatic speaker and age identification of children from raw speech using sincnet over erb scale," *Speech Communication*, vol. 159, p. 103069, 2024.

[15] S. Kataria, P. S. Nidadavolu, J. Villalba, and N. Dehak, "Analysis of deep feature loss based enhancement for speaker verification," in *The Speaker and Language Recognition Workshop (Odyssey)*, 2020, pp. 459–466.

[16] V. M. Shetty, J. Zheng, S. M. Lulich, and A. Alwan, "Enhancing age-related robustness in children speaker verification," in *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2025, pp. 1–5.

[17] M. Sang and J. H. Hansen, "Efficient adapter tuning of pre-trained speech models for automatic speaker verification," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 12 131–12 135.

[18] W. Huang, B. Han, S. Wang, Z. Chen, and Y. Qian, "Robust cross-domain speaker verification with multi-level domain adapters," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 11 781–11 785.

[19] T. Wang, L. Li, and D. Wang, "SE/BN Adapter: Parametric efficient domain adaptation for speaker recognition," in *Interspeech*, 2024, pp. 2145–2149.

[20] S. Han, Y. Ahn, K. Kang, and J. W. Shin, "Short-segment speaker verification using ecapa-tdnn with multi-resolution encoder," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[21] Z.-C. Chen, Y.-S. Sung, and H.-y. Lee, "Chapter: Exploiting convolutional neural network adapters for self-supervised speech models," in *IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*, 2023, pp. 1–5.

[22] J. Peng, T. Stafylakis, R. Gu, O. Plchot, L. Mošner, L. Burget, and J. Černockỳ, "Parameter-efficient transfer learning of pre-trained transformer models for speaker verification using adapters," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[23] W. Xia, J. Huang, and J. H. Hansen, "Cross-lingual text-independent speaker verification using unsupervised adversarial discriminative domain adaptation," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 5816–5820.

[24] Y. N. Dauphin, A. Fan, M. Auli, and D. Grangier, "Language modeling with gated convolutional networks," in *International conference on machine learning*. PMLR, 2017, pp. 933–941.

[25] K. Shobaki, J.-P. Hosom, and R. Cole, "The OGI kids' speech corpus and recognizers," in *Proc. of ICSLP*. Citeseer, 2000, pp. 564–567.

[26] W. Ward, R. Cole, D. Bolanos, C. Buchenroth-Martin, E. Svirsky, S. V. Vuuren, T. Weston, J. Zheng, and L. Becker, "My science tutor: A conversational multimedia virtual tutor for elementary school science," *ACM Transactions on Speech and Language Processing (TSLP)*, vol. 7, no. 4, pp. 1–29, 2011.

[27] V. P. Singh, M. Sahidullah, and T. Kinnunen, "Childaugment: Data augmentation methods for zero-resource children's speaker verification," *The Journal of the Acoustical Society of America*, vol. 155, no. 3, pp. 2221–2232, 2024.

[28] J. Villalba, N. Chen, D. Snyder, D. Garcia-Romero, A. McCree, G. Sell, J. Borgstrom, L. P. García-Perera, F. Richardson, R. Dehak, P. A. Torres-Carrasquillo, and N. Dehak, "State-of-the-art speaker recognition with neural network embeddings in NIST SRE18 and Speakers in the Wild evaluations," *Computer Speech Language*, vol. 60, p. 101026, 2020.

[29] D. Snyder, D. Garcia-Romero, A. McCree, G. Sell, D. Povey, and S. Khudanpur, "Spoken language recognition using X-vectors," in *The Speaker and Language Recognition Workshop (Odyssey)*, 2018, pp. 105–111.

[30] M. Ravanelli, T. Parcollet, P. Plantinga, A. Rouhe, S. Cornell, L. Lugosch, C. Subakan, N. Dawalatabad, A. Heba, J. Zhong *et al.*, "Speechbrain: A general-purpose speech toolkit," *arXiv preprint arXiv:2106.04624*, 2021.