

Personalized Speech Recognition for Children with Test-Time Adaptation

Zhonghao Shi^{*†}, Harshvardhan Srivastava^{†‡}, Xuan Shi^{*}, Shrikanth Narayanan^{*} and Maja Mataric^{*}

^{*}University of Southern California, Los Angeles, USA

[†]Columbia University, New York, USA

[‡]Sara Technology Inc., USA

Abstract—Accurate automatic speech recognition (ASR) for children is crucial for effective real-time child-AI interaction, especially in educational applications. However, off-the-shelf ASR models primarily pre-trained on adult data tend to generalize poorly to children’s speech due to the data domain shift from adults to children. Recent studies have found that supervised fine-tuning on children’s speech data can help bridge this domain shift, but human annotations may be impractical to obtain for real-world applications and adaptation at training time can overlook additional domain shifts occurring at test time. We devised a novel ASR pipeline to apply unsupervised test-time adaptation (TTA) methods for child speech recognition, so that ASR models pre-trained on adult speech can be continuously adapted to each child speaker at test time without further human annotations. Our results show that ASR models adapted with TTA methods significantly outperform the unadapted off-the-shelf ASR baselines both on average and statistically across individual child speakers. Our analysis also discovered significant data domain shifts both between child speakers and within each child speaker, which further motivates the need for test-time adaptation.

Index Terms—Child Speech Recognition, Test-Time Adaptation

I. INTRODUCTION

Child-AI interaction enabled by AI software agents [1] or socially assistive robots [2] has shown great potential for many application domains, for example education [3]. Conversational capabilities for these AI agents can support natural interaction with the child in achieving task goals [4]. To enable such human-like interaction, these AI agents and robots require accurate recognition of child speech [5]. Despite tremendous progress in machine learning methods for automatic speech recognition (ASR), a large body of recent work has shown that off-the-shelf pre-trained ASR models do not generalize well to children’s speech data, due to the high amount of acoustic and linguistic variability [6], resulting in data domain shifts between the adult data used for pre-training and child data used for testing [7], [8].

As detailed in Table I, recent work on child ASR [9]–[16] has experimented with various supervised methods within the setting of fine-tuning to adapt pre-trained ASR models at training time before model deployment. Prior studies have proposed applying methods such as transfer learning [9]–[12], continued pre-training [13], adapters [14], [15], and low-rank adaptation [16] to fine-tune and adapt the pre-trained ASR

TABLE I
PROPOSED TEST-TIME ADAPTATION (TTA) VS FINE-TUNING.

Adaptation Method	Supervision Setting	Adaptation Loss
fine-tuning [9]–[16]	supervised	at training time: $L(x^c, y^c)$
test-time adaptation (Ours)	unsupervised	at test time: $L(x^c)$

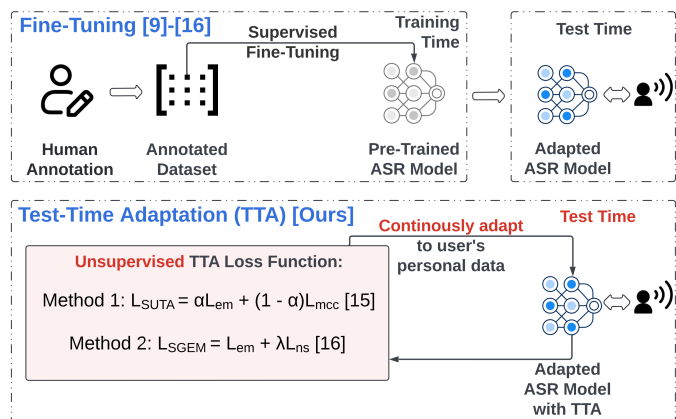


Fig. 1. Proposed Test-Time Adaptation (TTA) vs Fine-tuning for Child Speech Recognition. Pre-trained ASR models with TTA can be continuously adapted to personal data from each child test user locally on their own device, so that their performance can be more robust to domain shift and more privacy-aware without the need for data transfer.

models with annotated children’s speech data at training time. The results have shown that supervised fine-tuning with annotated children’s speech data can significantly adapt models to children’s speech [17].

Despite recent progress, as shown in Figure 1, the setting of supervised fine-tuning may not be feasible in real-world model deployments for the following reasons: 1) each new child speaker may introduce further domain shifts at test time; 2) supervised fine-tuning requires additional labeled annotations, but annotation requires considerable financial and human labor investment; and 3) users may prefer to keep their data private and stay on their local devices with limited computing capabilities, which makes fine-tuning at training time infeasible.

For these reasons, our work addresses the following two re-

search questions: RQ1) can unsupervised test-time adaptation (TTA) methods help adapt ASR models pre-trained on adult speech to child speech recognition at test time? and RQ2) if yes, why is it helpful to adapt ASR models at test time for child speech recognition?

For the first question, we devised a child speech ASR system by combining pre-trained wav2vec 2.0 ASR models [18] with two state-of-the-art TTA methods: SUTA [19] and SGEM [20]. Our results show that both TTA methods significantly outperform the unadapted baseline ASR models both on average and statistically across individual child speakers. The results support that TTA methods can continuously adapt pre-trained ASR models to a new user's personal data without the need for additional human annotations.

To address the second question, we individually analyzed the performance of off-the-shelf ASR models on each child speaker's data in the test set of MyST dataset. We also obtained the t-distributed stochastic neighbor embedding (t-SNE) for each child speaker's data to study the individual data domain shift for child speakers. Our findings suggest that there exist significant domain shifts and variance both across and within child speakers' speech data, confirming the necessity of TTA for child ASR after model deployment. This work aims to pave the way for more personalized and accurate child speech recognition models that can be robustly deployed in real-world applications.

II. METHODS

We propose a novel child speech recognition pipeline to adapt a pre-trained speech model on out-of-domain child speech by using test-time adaptation (TTA) techniques. As shown in the Figure II, TTA can continuously adapt the pre-trained model on a new user's speech data without further annotation and fine-tuning.

A. Problem Formulation

A canonical ASR model can be denoted as $z = f(x, \theta)$, where x is the input speech waveform, θ refers to the ASR model parameters, and $z \in \mathbb{R}^{L \times C}$ is the predicted context logits. L is the total number of timestamp, and C is the number of word class. Then a CTC-loss is applied on z to get the linguistic context prediction \hat{y} of x . ASR models are typically trained on a training dataset $D_{train} = \{(x_i^{tr}, y_i^{tr})\}_I$ in a supervised or unsupervised manner to estimate θ . The model is then used to transcribe \hat{y}_j^{te} from $x_j^{te} \in D_{test}$, which assumes identical distribution with D_{train} . However, in recognition on child speech D_{child} , an ASR model trained primarily on adult speech D_{adult} suffers performance degradation due to the wide acoustic and linguistic variability in child speech and associated data scarcity [7], [8]. To address this challenge, we propose a novel ASR system with a TTA function to modify the parameters of the pre-trained ASR model $\theta \rightarrow \tilde{\theta}$ in the test stage without supervision.

B. Pre-Trained Speech Model

We used a widely-used off-the-shelf ASR model, wav2vec 2.0 [18], as the baseline to recognize child speech. To achieve

better performance, we selected the most well-trained version: wav2vec2-base-960h, which is pre-trained on 960 hours of adult speech from Librispeech [21].

C. Test-Time Adaptation Methods

The goal of TTA is to design optimization objectives based on the output logits $z \in \mathbb{R}^{L \times C}$ to adapt the whole ASR model to the current test child's speech by continuously updating a small portion of model's parameters.

This work experimented with two state-of-the-art TTA methods: 1) single-utterance test-time adaptation (SUTA) [19]; and 2) sequential-level generalized entropy minimization (SGEM) [20]. The unsupervised optimization objective of SUTA [19] consists of two parts: 1) Shannon entropy minimization loss (\mathcal{L}_{em}); and 2) negative sampling loss (\mathcal{L}_{mcc}). With the weighting hyper-parameter α , the overall loss function is denoted as follows:

$$\mathcal{L}_{SUTA} = \alpha \mathcal{L}_{em} + (1 - \alpha) \mathcal{L}_{mcc} \quad (1)$$

$$\mathcal{L}_{em} = \frac{1}{L} \sum_{i=1}^L \mathcal{H}_i = -\frac{1}{L} \sum_{i=1}^L \sum_{j=1}^C \mathbf{P}_{ij} \log \mathbf{P}_{ij}, \quad (2)$$

$$\mathcal{L}_{mcc} = \sum_{j=1}^C \sum_{j' \neq j}^C \mathbf{P}_{\cdot j}^\top \mathbf{P}_{\cdot j'}, \quad (3)$$

The unsupervised optimization objective of SGEM [20] consists of two parts: 1) generalized Rényi entropy minimization (\mathcal{L}_{GEM}); and 2) negative sampling loss (\mathcal{L}_{NS}). With a weighting hyper-parameter λ , the overall loss function is denoted as below:

$$\mathcal{L} = \mathcal{L}_{GEM} + \lambda_{NS} \mathcal{L}_{NS}, \quad (4)$$

$$\mathcal{L}_{GEM} = \frac{1}{L} \sum_{i=1}^L \frac{1}{1 - \alpha} \log \left(\sum_{j=1}^C p_{ij}^\alpha \right), \quad (5)$$

$$\mathcal{L}_{NS} = -\frac{1}{L} \sum_{i=1}^L \log \left(1 - \sum_{j=1}^C \mathbb{I}_{[p'_{ij} < \tau]} p_{ij} \right) \quad (6)$$

III. EXPERIMENTAL SETUP

A. Datasets

This work uses the My Science Tutor (MyST) dataset [22], currently one of the largest publicly available datasets for child speech recognition¹. After removing low-quality recordings and utterances with missing annotation, we included 86 children's data in our experiments. There are on average 134 (SD=99) utterances for each child speaker in the dataset. The duration of the utterances varies from less than 1 second to 111 seconds across all speakers. Overall, 25.8 hours of data were included in our experiments.

¹<https://catalog.ldc.upenn.edu/LDC2021S05>

TABLE II
WORD ERROR RATE (%) FOR UNSUPERVISED TTA METHODS (SUTA AND SGEM) VS UNADAPTED BASELINE
FOR PARTICIPANT P1-P11 AND P77-P86.

Setting	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11
Unadapted	72.6	57.5	56.3	55.5	54.1	52.4	50.1	47.7	47.4	47.2	46.8
SUTA	69.0	53.7	59.4	50.9	48.8	42.1	46.7	45.9	38.1	45.5	42.1
SGEM	68.8	53.1	56.3	48.8	48.3	42.9	46.6	45.8	39.0	44.8	42.1

Setting	P77	P78	P79	P80	P81	P82	P83	P84	P85	P86	Average
Unadapted	20.0	18.9	18.6	18.5	17.2	16.8	16.1	13.0	12.0	10.7	31.1
SUTA	18.2	17.5	17.3	15.5	16.4	17.1	14.6	12.4	11.6	6.7	28.1
SGEM	18.3	17.6	17.1	15.3	16.6	17.0	14.3	12.4	11.3	7.3	27.8

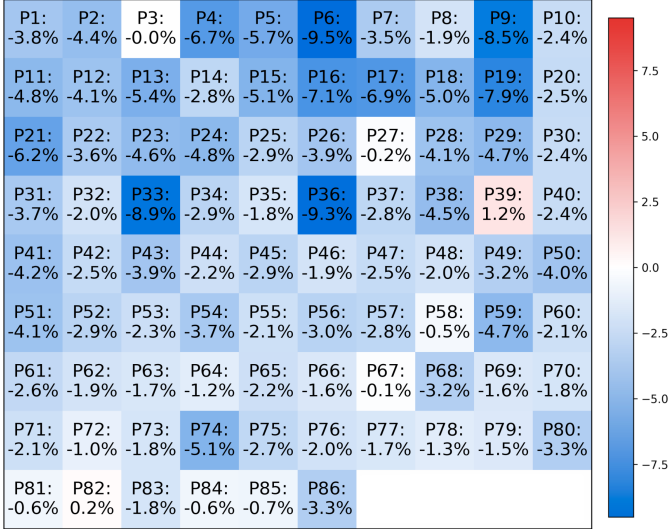


Fig. 2. Heatmap of performance gains in word error rate (%) for all child speakers with our proposed TTA pipeline using SGEM. We found that the bottom 50% of speakers who had worse unadapted WER also benefited significantly more from our TTA pipeline than the top 50% of speakers with better unadapted WER.

B. Experimental Setup

In the unadapted condition, we used the off-the-shelf pre-trained wav2vec2 model [18]². In the test-time adaptation conditions, we used the official implementation in SGEM [20]³ to develop our codebase. All experimental settings such as learning rate and optimizer were kept consistent between settings. For both TTA methods, the adaptation step was set as $N = 10$. The weighting hyper-parameter was set to the following default values: $\alpha = 0.3$ for SUTA and $\lambda = 0.3$ for SGEM. We evaluated ASR models using word error rate (WER). Due to the data distribution imbalance—a small group of speakers have many utterances—we report the unweighted average WER based on the subject rather than the utterance, so the results accurately represent the ASR system’s performance for each child.

²<https://huggingface.co/facebook/wav2vec2-base-960h>

³<https://github.com/drumpt/SGEM>

IV. RESULTS AND DISCUSSION

RQ1: Can test-time adaptation (TTA) methods help adapt pre-trained ASR models to test child speakers in an unsupervised fashion?

Our experiments compared our unsupervised test-time adaptation approach and the unadapted baseline. Our results indicate that pre-trained ASR models adapted with unsupervised test-time adaptation can significantly outperform the unadapted baselines both on average and statistically across child speakers on the MyST dataset.

As shown in Table II, on average, both SGEM (Mean=27.8%, SD=10.9%) and SUTA (Mean=28.1%, SD=11.1%) outperformed the unadapted WER baseline (Mean=31.1%, SD=11.8%) by 3.3% and 3%, respectively. We further conducted individual two-sided Wilcoxon signed-rank tests to validate whether the improvements between the TTA condition and the unadapted condition extend across all child speakers, to ensure the overall improvements were not driven by gains from a small groups of speakers. The Wilcoxon signed-rank tests found significantly better WER for the SGEM condition over the unadapted condition ($Z = 7.960, p < .001, r = 0.607$), and better WER for SUTA condition over the unadapted baseline ($Z = 7.805, p < .001, r = 0.595$). Between the SUTA and SGEM conditions, despite the similar performance on average, Wilcoxon signed-rank tests found that the SGEM condition significantly outperformed the SUTA condition, with lower WER ($Z = 3.404, p < .001, r = 0.260$).

In Table II, we present the performance gain enabled by unsupervised TTA methods over the unadapted baseline per speaker, to further analyze the impact of TTA methods on individual child speakers. Due to the page limit, of the 86 speakers in the dataset, in Table II, we present the results for the 6 speakers with the worst WER in the unadapted condition, and the 5 speakers with the best WER in the unadapted condition. We found that P6, who benefited from the TTA method the most, had a 10% gain with SUTA and 9.5% with SGEM. However, we also observed that P3 had a significant performance drop with 3.1% for SUTA, while SGEM was able to maintain the same WER. For P82, both SUTA (0.3%) and SGEM (0.2%) performed slightly worse than the unadapted baseline. This results indicate that SGEM may perform more robustly than SUTA on more challenging

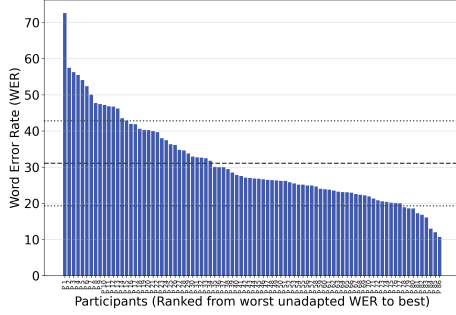


Fig. 3. Unadapted baseline performance in word error rate (WER). Unadapted pre-trained ASR model was evaluated on 86 individual child speakers’ data from the MyST dataset; results indicate significant differences in model performance across different child speakers, motivating the need for domain adaptation for each child speaker individually.

participants’ data. This helps explain why SGEM statistically out-performed SUTA across participants. Since TTA methods (SUTA and SGEM) are based on the assumption that incoming test streams only contain benign test data worth adapting to, the observed performance degradation on P3 and P82 may be caused by noisy data commonly found in real-world applications, which could misguide the adaptation and lead to worsened model performance [23]. Our findings validate the overall performance gains enabled by TTA, and also support the need for the development of more robust TTA methods for child speech recognition against noisy test data.

We visualized the performance gain enabled by SGEM over the unadapted baseline for every individual child speaker, as shown in Figure 2. Darker blue color indicates better performance gain from TTA, and lower participant numbers mean worse unadapted WER with off-the-shelf ASR models. We found that the top 50% of child speakers (P44-P86) with better unadapted WER had a 2.2% ($SD = 1.2\%$) WER improvement on average, while the bottom 50% of child speakers (P1-P43) with worse unadapted WER had larger 4.3% ($SD = 2.4$) WER improvements on average. We further conducted two-sided Mann-Whitney tests and found that the bottom 50% of child speakers (P1-P43) benefited significantly more from the SGEM TTA method than the top 50% ($Z = 4.772, p < .001, r = 0.515$). This finding shows that TTA methods potentially benefit more child speakers who initially cannot be well generalized by a pre-trained speech model.

RQ2: Why is it helpful to adapt ASR models at test time for child speech recognition?

The majority of prior work has reported average WER as a measure of model performance for child speech recognition, but has not yet analyzed individual WER differences between child speakers. As shown in Figure 3, we analyzed model performance of unadapted pre-trained ASR models for each child speaker individually, and observed significant differences across speakers. The child speaker with the worst WER result (P1) had a WER of 72.6%, while the child speaker with the

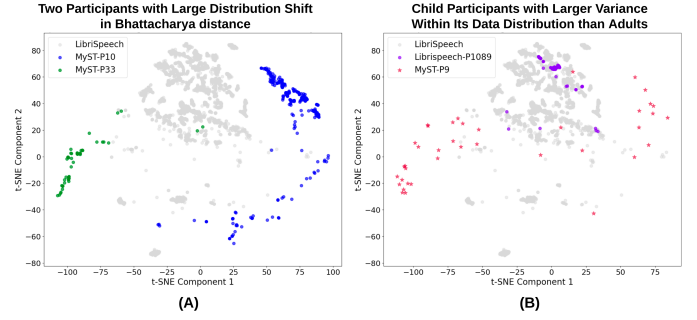


Fig. 4. Domain shift across and within child speakers. (A) Significant Domain shift between child speakers; (B) Significant variance and domain shift within child speakers

best result (P86) had a WER of 10.7%. Our findings suggest that off-the-shelf pre-trained models also may not generalize robustly across different child speakers, highlighting the need for more individual-level adaptation to ensure model robustness for all speakers.

We also analyzed and visualized the feature space using the T-distributed stochastic neighbor embedding (t-SNE). We calculated the pair-wise Bhattacharya distance across all pairs of child speakers in MyST data, and visualized the pair with one of the largest distances in in Figure 4A. The results validate that there may be significant distribution shift between child speakers. As shown in Figure 4B, we also calculated the variance within each child speaker’s embedding (Mean=1545.53 , $SD=915.32$) and found that it is significantly larger than the variance within each adult speaker from Librispeech (Mean=259.14, $SD=275.02$). Our findings suggest that child speakers may also have larger distribution shifts within their data distribution due to more expressive speech, which can only be addressed at test time. These findings further motivate the need for applying test-time adaptation for child speech recognition to ensure robust model generalization in real-world applications.

V. CONCLUSION AND FUTURE WORK

We described a novel pipeline to adapt off-the-shelf pre-trained ASR models to out-of-domain children’s speech using unsupervised test-time adaptation. Our results show that the proposed ASR pipeline significantly outperformed the baselines both on average and statistically across speakers, without the need for additional human annotations. Our analysis also revealed that there may exist significant domain shifts both between and within child speakers, further motivating the need for test-time adaptation. In future work, we aim to develop more robust child speech recognition systems. First, we will improve the ASR system to perform more robustly in sophisticated scenarios, for example in noisy and far-field speech. Second, we will keep exploring the unique characteristics of child speech to continue gaining insights that can assist improvements of child speech recognition toward more robust systems.

REFERENCES

- [1] W. Huang, K. F. Hew, and L. K. Fryer, "Chatbots for language learning—are they really useful? a systematic review of chatbot-supported language learning," *Journal of Computer Assisted Learning*, vol. 38, no. 1, pp. 237–257, 2022.
- [2] T. Belpaeme, P. Baxter, J. De Greeff, J. Kennedy, R. Read, R. Looije, M. Neerinx, I. Baroni, and M. C. Zelati, "Child-robot interaction: Perspectives and challenges," in *Social Robotics: 5th International Conference, ICSR 2013, Bristol, UK, October 27-29, 2013, Proceedings 5*. Springer, 2013, pp. 452–459.
- [3] W. Holmes and I. Tuomi, "State of the art and practice in ai in education," *European Journal of Education*, vol. 57, no. 4, pp. 542–570, 2022.
- [4] S. S. Narayanan and A. Potamianos, "Creating conversational interfaces for children," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 2, pp. 65–78 [IEEE Signal Processing Society Best Paper Award Wnner, 2005], feb 2002.
- [5] A. Potamianos and S. S. Narayanan, "Robust recognition of children's speech," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 6, pp. 603–616, nov 2003.
- [6] S. Lee, A. Potamianos, and S. S. Narayanan, "Acoustics of children's speech: Developmental changes of temporal and spectral parameters," *Journal of the Acoustical Society of America*, vol. 105, no. 3, pp. 1455–1468, mar 1999, selected Research Article.
- [7] V. Bhardwaj, M. T. Ben Othman, V. Kukreja, Y. Belkhier, M. Bajaj, B. S. Goud, A. U. Rehman, M. Shafiq, and H. Hamam, "Automatic speech recognition (asr) systems for children: A systematic literature review," *Applied Sciences*, vol. 12, no. 9, p. 4419, 2022.
- [8] P. Gurunath Shivakumar and S. Narayanan, "End-to-end neural systems for automatic children speech recognition: An empirical study," *Computer Speech & Language*, vol. 72, p. 101289, 2022. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0885230821000905>
- [9] S. Shraddha, S. Kumar *et al.*, "Child speech recognition on end-to-end neural asr models," in *2022 2nd International Conference on Intelligent Technologies (CONIT)*. IEEE, 2022, pp. 1–6.
- [10] R. Jain, A. Barcovich, M. Yiwere, P. Corcoran, and H. Cucu, "Adaptation of whisper models to child speech recognition," *arXiv preprint arXiv:2307.13008*, 2023.
- [11] J. Thienpondt and K. Demuynck, "Transfer learning for robust low-resource children's speech asr with transformers and source-filter warping," *arXiv preprint arXiv:2206.09396*, 2022.
- [12] T. Rolland, A. Abad, C. Cucchiari, and H. Strik, "Multilingual transfer learning for children automatic speech recognition," 2022.
- [13] A. A. Attia, D. Demszky, T. Ogunremi, J. Liu, and C. Espy-Wilson, "Continued pretraining for domain adaptation of wav2vec2.0 in automatic speech recognition for elementary math classroom settings," *arXiv preprint arXiv:2405.13018*, 2024.
- [14] T. Rolland and A. Abad, "Exploring adapters with conformers for children's automatic speech recognition," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 12 747–12 751.
- [15] R. Fan and A. Alwan, "Draft: A novel framework to reduce domain shifting in self-supervised learning and its application to children's asr," *arXiv preprint arXiv:2206.07931*, 2022.
- [16] W. Liu, Y. Qin, Z. Peng, and T. Lee, "Sparsely shared lora on whisper for child speech recognition," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 11 751–11 755.
- [17] P. G. Shivakumar and P. Georgiou, "Transfer learning from adult to children for speech recognition: Evaluation, analysis and recommendations," *Computer speech & language*, vol. 63, p. 101077, 2020.
- [18] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in neural information processing systems*, vol. 33, pp. 12 449–12 460, 2020.
- [19] G.-T. Lin, S.-W. Li, and H.-y. Lee, "Listen, adapt, better wer: Source-free single-utterance test-time adaptation for automatic speech recognition," *arXiv preprint arXiv:2203.14222*, 2022.
- [20] C. Kim, J. Park, H. Shim, and E. Yang, "Sgem: Test-time adaptation for automatic speech recognition via sequential-level generalized entropy minimization," *arXiv preprint arXiv:2306.01981*, 2023.
- [21] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.
- [22] W. Ward, R. Cole, and S. Pradhan, "My science tutor and the myst corpus," *Boulder Learning Inc*, 2019.
- [23] T. Gong, Y. Kim, T. Lee, S. Chottananurak, and S.-J. Lee, "Sotta: Robust test-time adaptation on noisy data streams," *Advances in Neural Information Processing Systems*, vol. 36, 2024.