

CPT-Boosted Wav2vec2.0: Towards Noise Robust Speech Recognition for Classroom Environments

Ahmed Adel Attia¹, Dorottya Demszky², Tolulope Ògúnrémi², Jing Liu¹, Carol Espy-Wilson¹

¹*University of Maryland College Park, MD, USA*, ²*Stanford University, CA, USA*

aadel@umd.edu, ddemszky@stanford.edu, tolulope@cs.stanford.edu, jliu28@umd.edu, espy@umd.edu

Abstract—Creating Automatic Speech Recognition (ASR) systems that are robust and resilient to classroom conditions is paramount to the development of AI tools to aid teachers and students. In this work, we study the efficacy of continued pretraining (CPT) in adapting Wav2vec2.0 to the classroom domain. We show that CPT is a powerful tool in that regard and reduces the Word Error Rate (WER) of Wav2vec2.0-based models by upwards of 10%. More specifically, CPT improves the model’s robustness to different noises, microphones and classroom conditions.

Index Terms—self-supervised learning, wav2vec2, asr, automatic speech recognition, classrooms

I. INTRODUCTION

Developing interactive AI tools for classrooms can help create a more fair and equitable learning environment while providing tools for teachers to aid in the teaching process. Automatic Speech Recognition (ASR) is a critical part of such pipelines. Transcripts generated by ASR systems can be analyzed on many levels to understand the dynamics in the classrooms — if they are sufficiently accurate [1]–[3]. However, classroom ASR remains a largely unsolved challenge. ASR systems face significant challenges when dealing with children’s speech, even under ideal conditions. These systems are predominantly trained on adult speech, which leaves them ill-equipped to handle the distinct characteristics of children’s speech. Children generally articulate less clearly than adults [18], and their speech possesses unique acoustic and linguistic properties that differ significantly from those of adults [17].

In our previous work [4], we analyzed Whisper’s [5] performance on children’s speech and our findings suggest that while it can achieve adult-level performance with simple and scripted prompts—demonstrating its ability to adapt to the acoustic characteristics of children’s speech—it still struggles with the linguistic characteristics of children’s speech. Several other studies have shown that fine-tuning improves the performance of popular ASR systems, yet a significant gap remains between the effectiveness of these systems on adult versus children’s speech [4], [14]–[16].

In classrooms, the problem gets more complicated. Babble noise is considered one of the most challenging noises even in adult speech with adult babble [30]. However, children’s babble noise like in classrooms is more challenging as it is less likely to occur in public datasets that most current ASR models are trained on. Other conditions that affect the accuracy

This work is supported by the Grand Challenge Award at the University of Maryland

of ASR systems, like far-field speech [13], and multi-speaker conditions [19], [20] are also abundant in classrooms.

The previously mentioned challenges are further exacerbated by the scarcity of transcribed classroom datasets, many of which are not publicly available due to the sensitive nature of data involving minors. While non-transcribed classroom recordings exist, transcription costs can be prohibitively high. As such, this low-resource environment is particularly well-suited for self-supervised speech representation models like Wav2vec2.0 [6] and HuBERT [25], which can leverage untranscribed data for pretraining, while the available transcribed data can be used for fine-tuning the model for ASR tasks.

In this paper, we propose continued pretraining (CPT) as an effective way to adapt Wav2vec2.0 to domain-specific noisy speech recognition, namely classroom speech recognition. Starting from different Wav2vec2.0 versions as initialization, we perform self-supervised pretraining on unlabeled noisy classroom data, then finetune on small labeled subsets.

Our contributions in this research are as follows:

- We show that CPT is the most effective tool to adapt Wav2vec2.0 to noisy conditions like classrooms compared to existing methods.
- We perform an ablation study to determine the optimal pretrained model to use as initialization for CPT.
- We show how CPT can improve the robustness of Wav2vec2.0, not only to noise but to different microphone configurations and demographics.
- We show that our proposed method is more robust to noise than State Of The Art (SOTA) ASR models.
- We demonstrate the use of existing classroom text corpora for Language Model (LM) training.

To facilitate further research and reproducibility, our training code as well as our model checkpoints will be available at the time of the camera-ready submission.

II. BACKGROUND AND RELATED PREVIOUS WORKS

A. Wav2vec2.0

Wav2vec2.0 is a Self Supervised Learning (SSL) speech representation model developed by [6] which utilizes the contextualization capabilities of transformers to learn contextual self-supervised representations from unlabeled audio.

While supervised speech models, like Whisper, learn directly on human-annotated, task-specific labeled data to achieve SOTA performance, Wav2vec2.0 models are first *pretrained* on unlabeled data using contrastive learning to

extract contextual representation. Then, a single classification layer is added on top of the model and the entire model is *fine-tuned* on a smaller labeled dataset using CTC loss [31]. *Continued pretraining (CPT)* refers to performing additional self-supervised pre-training on a model that was already pre-trained before fine-tuning.

B. Adaptation of Wav2vec2.0 to low-resource languages through CPT

Several research papers [10]–[12] investigated the effectiveness of CPT in adapting multi-lingual self-supervised ASR systems like XLSR53 [22] and XLS-R [24] to low-resource languages. The research by [10] developed an ASR system for Ainu, a critically endangered and low-resource language. Starting from XLSR53 which was already pretrained on 56K hours from 53 languages, they performed CPT on 234 hours of Ainu recordings. They describe CPT as “clearly the most effective way to adapt a speech representation model for a new language”. CPT decreased their WER by up to 40% relative to the unadapted model.

III. DATASETS

1) *NCTE*: The NCTE dataset consists of video and audio recordings of 2128 4th and 5th-grade elementary math classrooms collected as part of the National Center for Teacher Effectiveness (NCTE) Main Study [26]. The observations took place between 2010 and 2013 across four districts serving historically marginalized students.

For each classroom, 2 to 3 video and audio recordings exist, from different angles and microphones, each lasting 45 minutes to an hour. The total duration of the recordings from all microphones and classrooms is 5235 hours. We resampled the audio from 44.1KHz to 16KHz and cut it into 20-second chunks. About 10% of the data was reserved for validation.

Out of these recordings, 6 were randomly chosen to be transcribed to create a low-resource unbalanced problem. Classrooms within this subset have varying demographic makeups. One of these recordings, which is denoted in Table II as 2619, comes from a far field microphone to test The duration of this subset is about 5.15 hours, with the duration of each file between 45-60 minutes, and it was used for supervised *fine-tuned* generalization to unseen microphone configuration.

2) *MPT dataset*: We recorded six classrooms as part of the larger M-Powering Teachers (MPT) dataset we are compiling. Two 8th-grade classrooms from a Washington, DC charter school, labeled **DC-1** and **DC-2**, served predominantly African-American and Hispanic low-income students, including special education and English language learners. Two 5th-grade classrooms in Eastlake, Ohio, referred to as **OH-1** and **OH-2**, had mostly White students, with some in special education. Lastly, two 6th-grade classrooms from a private school in San Jose, California, labeled **CA-1** and **CA-2**, had predominantly White and Asian high-income students, with no special education or English language learners. In each classroom, five microphones were placed at different places and the audio streams were added together. This resulted in

a good capture of all the audio in the classroom but also extremely noisy audio in one particularly noisy classroom, CA-1. We keep this configuration to test the model’s ability to handle extremely noisy conditions. We use this dataset to test the effect of CPT on improving performance in different but related domains than the one seen during CPT.

3) *NCTE-Text*: NCTE-Text is a dataset of anonymized transcriptions of 1660 classrooms from the NCTE dataset [9]. To protect subjects’ privacy, all names in the text corpus were de-identified and replaced by their initials. These transcripts were not intended for ASR training and are not suitable for it as they are not verbatim. However, this text corpus can be used to train task-specific lightweight n-gram LM for beam-search decoding of Wav2vec2.0 outputs. To make this data suitable for LM training, we replaced the initials with randomly chosen names. To ensure that the names are unbiased towards race or gender, we referenced a list of the most popular baby names by race between 2011 and 2019 in New York City¹. For each classroom transcription, de-identified initials were replaced by names sampled from this list, to ensure equitable representation across gender and racial identity.

IV. EXPERIMENTS

A. CPT Experiments

We performed three CPT experiments to contrast the effect of the starting checkpoint for CPT on domain adaptation. We considered three 300M parameter checkpoints of Wav2vec2.0.

- **W2V-LV60**, pretrained on 60K hours of LibriVox [27].
- **W2V-Robust**, pretrained on 60K hours of LibriVox, noisy telephone speech and crowd sourced data. [23].
- **XLS-R**, pretrained on 436K hours of speech from 128 languages, including English [24].

Fine-tuning the resulting models serves as an ablation study to determine the best criteria for the starting checkpoint of CPT. The contrast between the performance of W2V-LV60 and W2V-Robust will showcase the effect of additional out-of-domain (OOD) noisy data during initial pretraining on the efficacy of CPT domain adaptation. The performance of XLS-R will showcase the impact of having more pretraining data from completely different domains and languages.

We also pretrain Wav2vec2.0 from scratch, which we denote as **W2V-SCR** to contrast the effect of CPT versus pre-training from scratch. For both pre-training from scratch as well as CPT, we use 5235 hours of untranscribed NCTE audio.

We perform two separate 6-fold cross-validation finetuning on each of the two datasets used for training, leaving one classroom recording for validation in each fold. We finetune each off-the-shelf Wav2vec2.0 model (W2V-LV60, W2V-Robust, and XLS-R), and their CPT counterparts as well as W2V-SCR.

We preprocess our data by cutting the audio into chunks of 30 seconds or less depending on the timestamps of the transcription. We normalize the text, removing casing and punctuation as described in the seminal Whisper paper.

¹<https://data.cityofnewyork.us/Health/Popular-Baby-Names/25th-nujf/data>

TABLE I
AVERAGE CROSS-VALIDATION RESULTS FROM FINETUNING VARIOUS
WAV2VEC2.0 CHECKPOINTS, WITH AND WITHOUT CONTINUED
PRETRAINING, COMPARED TO THE WHISPER SMALL ENGLISH-ONLY
CHECKPOINT. “WHISPER-FT” INDICATES FINETUNING ON THE TARGET
DATASET. WER STANDARD DEVIATIONS ARE IN BRACKETS.

Model	LM	WER(STD)	
		NCTE	MPT
<i>Pretraining from Scratch</i>			
W2V-SCR	None	47.34(5.73)	51.39 (6.83)
	5-gram LM	30.25(15.44)	38.59(12.93)
<i>No Continued Pretraining</i>			
W2V-LV60K	None	39.11(13.01)	37.82(12.30)
	5-gram LM	30.39(14.48)	33.56(10.86)
XLS-R	None	38.19(10.39)	39.12(13.60)
	5-gram LM	29.02(10.96)	32.49(10.76)
W2V-Robust	None	35.07(11.85)	36.36(11.54)
	5-gram LM	27.99(13.28)	31.49(9.97)
<i>Continued Pretraining</i>			
W2V-LV60K (CPT)	None	22.52(4.89)	32.26(8.92)
	5-gram LM	18.13(5.50)	26.72(7.72)
XLS-R (CPT)	None	26.53(5.13)	32.16(10.78)
	5-gram LM	19.37(4.91)	26.80(8.00)
W2V-Robust (CPT)	None	25.04(5.28)	30.97(9.99)
	5-gram LM	17.71(5.06)	26.50(8.09)
Whisper	-	24.46(12.37)	30.15(12.99)
Whisper-FT	-	19.14(6.77)	28.53(14.07)

B. N-Gram LM Training

We train a 5-gram LM on the deanonymized NCTE-Text dataset. We normalize the training data using the Whisper Normalizer library to match the transcription text.

V. RESULTS AND DISCUSSION

A. Pretraining from scratch on target domain data

Table I shows the average cross-validation WER across all test folds for every model. Pre-training Wav2vec2.0 from scratch, we get high WER of 30.25% and 38.59% in NCTE and MPT, respectively. The performance gap between the two datasets indicates that the NCTE task is easier, which aligns with the noise levels observed in the datasets.

The standard deviation in the results is quite high, indicating that each recording has unique characteristics. As a result, the folds perform differently on each one. Although we consider classroom recordings to be a single domain, the type of classroom environment—whether collaborative or instructional—affects the noise level and the ratio of teacher to student speech. All of these factors influence performance.

B. Off-the-shelf pre-trained models

Previous work [23] shows that pre-training on target domain data improves performance compared to pre-training on OOD data. However, our results indicate that pre-training from scratch on target data underperforms off-the-shelf models, including W2V-LV60K which is pre-trained entirely on clean speech. Unlike [23], where in-domain and out-of-domain pre-training datasets were the same size, the OOD pre-training

datasets are at least 12 times larger than the target domain pre-training dataset, highlighting the crucial role of pre-training dataset size in learning high-quality speech embeddings.

Without CPT, fine-tuning W2V-Robust outperforms both XLS-R and W2V-LV60K. W2V-Robust was pre-trained on the same data as W2V-LV60K, plus additional noisy English speech, showing that adding OOD noisy data improves performance. XLS-R, despite being trained on a much larger, cross-lingual dataset, performs worse than W2V-Robust, but better than W2V-LV60K. This supports previous findings [23], [24] that larger cross-domain corpora can help. However, smaller, domain-relevant data (noisy adult English) performs better. All models still outperform training from scratch on smaller in-domain data.

To summarize, pre-training on larger OOD datasets generally yields better performance than small in-domain pre-training. The closer the pre-training data aligns with the target domain, the better the results—though more data still proves beneficial regardless.

C. CPT on target domain data

Looking at CPT results in the third section of Table I, we can see that for NCTE with LM decoding, CPT improves WER by up to 12.26%. This shows that CPT is a powerful tool for domain adaptation when labeled data is scarce and more unlabeled data is available. For the MPT dataset which is different from the pre-training data, CPT is still effective, yielding an improvement of up to 6.84%, proving that CPT is effective in domain adaptation in a generalizable fashion. It is also noticeable how the standard deviation of the WER decreases in both datasets. This shows that performing CPT on diverse classroom environments and noise conditions improves the ability of the model to generalize to these conditions.

In terms of the choice of starting point for CPT, the order of performance with CPT does not follow the order observed with off-the-shelf models. W2V-LV60K outperforms XLS-R with CPT, meaning that the performance edge XLS-R had from initially from large cross-lingual pre-training does not carry forward with CPT, and initial pretraining on English-only datasets is always superior. However, W2V-Robust still provides the best performance, showing that initial pretraining on OOD noisy English data yields better speech representations when adapted to the target domain.

The gap between LM-decoded and non-LM decoded results is much larger when pretraining from scratch, suggesting the model learns acoustics but lacks sufficient linguistic representation, which LM decoding compensates for. This indicates that initial pretraining on clean adult speech, even in other languages, captures useful linguistic features not learned from smaller, noisy in-domain data. Notably, there's a 25% gap in non-LM decoded WER between W2V-LV60K (CPT) and W2V-SCR, highlighting that CPT benefits from clean OOD speech and further refinement on in-domain data with CPT.

D. Comparison with Whisper

Finally, we can see that with CPT, W2V-Robust outperforms the Whisper checkpoint of comparable size, even with

TABLE II

DETAILED RESULTS ON EACH FOLD OF THE CROSS-VALIDATION ON BOTH THE NCTE AND MPT DATASETS, FOR THE OFF-THE-SHELF AND THE CPT VERSION OF W2V-ROBUST AND THE FINETUNED SMALL ENGLISH-ONLY WHISPER CHECKPOINT. "FOLD" REFERS TO THE FILE USED FOR TESTING.

NCTE			
Fold	Whisper-FT	W2V-Robust	W2V-Robust (CPT)
144	14.78	19.86	15.20
622	16.97	26.31	18.77
2619	28.4	54.03	27.15
2709	12.74	19.09	13.12
2944	26.98	28.07	17.61
4724	14.95	20.55	14.43
Average	19.14	27.99	17.71
MPT			
Fold	Whisper-FT	W2V-Robust	W2V-Robust (CPT)
OH-1	19.76	18.55	15.42
OH-2	16.42	19.89	16.88
DC-1	28.79	29.42	24.90
DC-2	26.42	37.32	30.34
CA-1	55.77	49.03	41.54
CA-2	24.04	34.47	29.91
Average	28.53	31.45	26.50

finetuning in both NCTE and MPT datasets. The nature of self-supervised speech models breaks down the problem into three parts: pretraining/CPT, finetuning, and LM decoding. This gives us more flexibility to utilize unmappable representations, like unlabeled audio or text that doesn't correspond well to said audio. Without CPT or LM decoding, Wav2vec2.0 performs much worse than Whisper, with WER being higher by 15-19% in the NCTE dataset. The flexibility that Wav2vec2.0 allows beyond supervised finetuning improved the performance by up to 19% through a combination of CPT and LM decoding.

E. Detailed analysis of cross-validation results

In this section, we take a closer look at the results. We start by discussing the results in Table II which shows the WER of each fold of the cross-validation in Whisper-FT and W2V-Robust with and without CPT.

1) *NCTE*: Looking at the results, it is apparent that CPT significantly improves the performance in each fold of the cross-validation with one notable example, recording 2619. This recording is the only one that comes from a far-field microphone, with the entirety of the training data in this cross-validation fold coming from near-field microphones. It is thus no surprise that without CPT, the model performs much worse in this fold than the others. However, with CPT, the error is cut in half, with an absolute improvement of almost 27%, proving that CPT helps the model generalize to microphone configurations unseen in the labeled data but present in the unlabeled pretraining data.

In terms of comparison with Whisper, looking at the NCTE results, CPT allows the model to achieve close performance or improve upon Whisper in every fold, with one major improvement noticed with recording 2944. Upon manual inspection, it was noted that this class started as an instructional class for 15 minutes and then the teacher assigned a set of

questions to the students and started making rounds in the class. This resulted in a much noisier environment than other classrooms with a higher degree of children's babble noise. Whisper was unable to deal with children's babble noise, often interleaving the target speaker with whatever it could discern from the background noise and sometimes exhibiting characteristic Whisper hallucinations by repeating a single word or phrase, for example, "*how do you know that what what is the denominator what is the denominator the the the the the...*" with the word "*the*" repeating 206 times. Wav2vec does not suffer from the same hallucination problem, and with CPT, it's much more capable of dealing with children's babble noise and focusing on the target speaker correctly.

2) *MPT dataset*: The differences in WER between Whisper and W2V-Robust with CPT on each fold are higher in the MPT dataset. W2V-Robust with CPT outperforms Whisper in 3 folds, with the main improvement coming from the recording of class CA-1. This recording is perhaps the noisiest of all the datasets used in this study, as discussed in the **Dataset** section. Both W2V-Robust and Whisper have high WER for this class recording, but even the non-CPT W2V-Robust outperformed Whisper. Upon manual inspection of the transcriptions, we again see that Whisper suffers from extreme hallucinations with high children babble noise. On the other hand, W2V-Robust with CPT is more capable of handling extremely noisy conditions, outperforming Whisper in this fold by 14%.

VI. CONCLUSIONS AND FUTURE WORK

We have demonstrated how CPT is the most effective method to adapt Wav2vec2.0 models to different domains, when compared to existing methods, improving the model's ability to generalize to different noise conditions. We show that when fine-tuning exclusively on near-field recordings, CPT cuts the WER of far-field recordings in half, showing that the model learns to generalize to acoustic conditions not found in the labeled set. We've shown that CPT can improve the WER on noisy classroom data by up to 12.26% on average and up to 27% in specific conditions. Our results suggest that CPT should be the baseline for low-resource domain adaptation experiments, especially in noisy applications as it is superior to other configurations and SOTA models like Whisper. We provide guidelines for selecting a CPT starting point, emphasizing that domain similarity is more important than the amount of pre-training data. We also propose a race-aware deanonymized classroom text dataset for LM training.

There is an urgent need for more balanced labeled classroom datasets. To that end, we are developing tools to sample recordings from the unlabeled NCTE dataset for transcription in a way that ensures balanced demographics and fair representation. We are also working towards a larger CPT trial, with an additional 15K hours of unlabeled classroom recordings. Lastly, we plan to expand on the work of [8], using speech enhancement-based Wav2vec2.0 pretraining. We are working on simulating classroom noises to augment clean speech to develop educational tools for physical and virtual classrooms.

REFERENCES

- [1] Jacobs, Jennifer, Scornavacco, Karla, Clevenger, Charis, Suresh, Abhijit, Sumner, Tamara. Automated feedback on discourse moves: teachers' perceived utility of a professional learning tool. *Educational technology research and development* vol. , no. , pp. 1–23, 2024.
- [2] Jacobs, Jennifer, Scornavacco, Karla, Harty, Charis, Suresh, Abhijit, Lai, Vivian, Sumner, Tamara. Promoting rich discussions in mathematics classrooms: Using personalized, automated feedback to support reflection and instructional change. *Teaching and Teacher Education* vol. 112, no. , pp. 103631, 2022.
- [3] Demszky, Dorotya, Liu, Jing, Hill, Heather C., Sanghi, Shyamoli, Chung, Ariel. Improving Teachers' Questioning Quality through Automated Feedback: A Mixed-Methods Randomized Controlled Trial in Brick-and-Mortar Classrooms. *EdWorkingPapers* vol. , no. , pp. , 2023.
- [4] Attia, Ahmed Adel, Liu, Jing, Ai, Wei, Demszky, Dorotya, Espy-Wilson, Carol. Kid-Whisper: Towards Bridging the Performance Gap in Automatic Speech Recognition for Children VS. Adults. *arXiv preprint arXiv:2309.07927* vol. , no. , pp. , 2023.
- [5] Radford, Alec, Kim, Jong Wook, Xu, Tao, Brockman, Greg, McLeavy, Christine, Sutskever, Ilya. Robust speech recognition via large-scale weak supervision, vol. , no. , pp. 28492–28518, 2023.
- [6] Baevski, Alexei, Zhou, Yuhao, Mohamed, Abdelrahman, Auli, Michael. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems* vol. 33, no. , pp. 12449–12460, 2020.
- [7] Hsu, W., Bolte, B., Tsai, Y., Lakhotia, K., Salakhutdinov, R. & Mohamed, A. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions On Audio, Speech, And Language Processing*. **29** pp. 3451–3460 (2021)
- [8] Zhu, Qiu-Shi, Zhang, Jie, Zhang, Zi-Qiang, Wu, Ming-Hui, Fang, Xin, Dai, Li-Rong. A noise-robust self-supervised pre-training model based speech representation learning for automatic speech recognition, vol. , no. , pp. 3174–3178, 2022.
- [9] Demszky, Dorotya, Hill, Heather. The NCTE transcripts: A dataset of elementary math classroom transcripts. *arXiv preprint arXiv:2211.11772* vol. , no. , pp. , 2022.
- [10] Nowakowski, Karol, Ptaszynski, Michal, Murasaki, Kyoko, Nieuwāzy, Jagna. Adapting multilingual speech representation model for a new, underresourced language through multilingual fine-tuning and continued pretraining. *Information Processing & Management* vol. 60, no. 2, pp. 103148, 2023.
- [11] San, Nay, Paraskevopoulos, Georgios, Arora, Aryaman, He, Xiluo, Kaur, Prabhjot, Adams, Oliver, Jurafsky, Dan. Predicting positive transfer for improved low-resource speech recognition using acoustic pseudo-tokens. *arXiv preprint arXiv:2402.02302* vol. , no. , pp. , 2024.
- [12] Paraskevopoulos, G., Kouzelis, T., Rouvalis, G., Katsamanis, A., Katsouras, V. & Potamianos, A. Sample-Efficient Unsupervised Domain Adaptation of Speech Recognition Systems: A Case Study for Modern Greek. *IEEE/ACM Transactions On Audio, Speech, And Language Processing*. (2023)
- [13] Zhu, Qiushi, Zhang, Jie, Gu, Yu, Hu, Yuchen, Dai, Lirong. Multichannel AV-wav2vec2: A Framework for Learning Multichannel Multi-Modal Speech Representation. *arXiv preprint arXiv:2401.03468* vol. , no. , pp. , 2024.
- [14] Shahnawazuddin, S., others. Developing children's ASR system under low-resource conditions using end-to-end architecture. *Digital Signal Processing* vol. 146, no. , pp. 104385, 2024.
- [15] Jain, Rishabh, Barcovschi, Andrei, Yiwere, Mariam, Corcoran, Peter, Cucu, Horia. Adaptation of Whisper models to child speech recognition. *arXiv preprint arXiv:2307.13008* vol. , no. , pp. , 2023.
- [16] Southwell, Rosy, Ward, Wayne, Trinh, Viet Anh, Clevenger, Charis, Clevenger, Clay, Watts, Emily, Reitman, Jason, D'Mello, Sidney, Whitehill, Jacob. Automatic Speech Recognition Tuned for Child Speech in the Classroom, vol. , no. , pp. 12291–12295, 2024.
- [17] Gerosa, Matteo, Giuliani, Diego, Narayanan, Shrikanth, Potamianos, Alexandros. A review of ASR technologies for children's speech, vol. , no. , pp. 1–8, 2009.
- [18] Lee, Sungbok, Potamianos, Alexandros, Narayanan, Shrikanth. Acoustics of children's speech: Developmental changes of temporal and spectral parameters, *The Journal of the Acoustical Society of America* vol. 105, no. 3, pp. 1455–1468, 1999.
- [19] Chang, Xuankai, Zhang, Wangyou, Qian, Yanmin, Le Roux, Jonathan, Watanabe, Shinji. End-to-end multi-speaker speech recognition with transformer, vol. , no. , pp. 6134–6138, 2020.
- [20] Chang, Xuankai, Qian, Yanmin, Yu, Kai, Watanabe, Shinji. End-to-end monaural multi-speaker ASR system without pretraining, vol. , no. , pp. 6256–6260, 2019.
- [21] Jain, Rishabh, Barcovschi, Andrei, Yiwere, Mariam Yahayah, Corcoran, Peter, Cucu, Horia. Exploring Native and Non-Native English Child Speech Recognition With Whisper, *IEEE Access* vol. 12, no. , pp. 41601–41610, 2024.
- [22] Conneau, Alexis, Baevski, Alexei, Collobert, Ronan, Mohamed, Abdelrahman, Auli, Michael. Ünsupervised cross-lingual representation learning for speech recognition. *arXiv preprint arXiv:2006.13979* vol. , no. , pp. , 2020.
- [23] Hsu, Wei-Ning, Sriram, Anuroop, Baevski, Alexei, Likhomanenko, Tatiana, Xu, Qiantong, Pratap, Vineel, Kahn, Jacob, Lee, Ann, Collobert, Ronan, Synnaeve, Gabriel, others. Robust wav2vec 2.0: Analyzing domain shift in self-supervised pre-training. *arXiv preprint arXiv:2104.01027* vol. , no. , pp. , 2021.
- [24] Babu, Arun, Wang, Changhan, Tjandra, Andros, Lakhotia, Kushal, Xu, Qiantong, Goyal, Naman, Singh, Kritika, von Platen, Patrick, Saraf, Yatharth, Pino, Juan, others. XLS-R: Self-supervised cross-lingual speech representation learning at scale. *arXiv preprint arXiv:2111.09296* vol. , no. , pp. , 2021.
- [25] Hsu, Wei-Ning, Bolte, Benjamin, Tsai, Yao-Hung Hubert, Lakhotia, Kushal, Salakhutdinov, Ruslan, Mohamed, Abdelrahman, Hubert: Self-supervised speech representation learning by masked prediction of hidden units, *IEEE/ACM Transactions on Audio, Speech, and Language Processing* vol. 29, no. , pp. 3451–3460, 2021.
- [26] Kane, Thomas, Hill, Heather, Staiger, Douglas. National Center for Teacher Effectiveness Main Study, vol. , no. , pp. , 2022.
- [27] Kearns, Jodi. Librivox: Free public domain audiobooks, *Reference Reviews* vol. 28, no. 1, pp. 7–8, 2014.
- [28] Martin, Joshua L, Wright, Kelly Elizabeth. Bias in automatic speech recognition: The case of African American Language, *Applied Linguistics* vol. 44, no. 4, pp. 613–630, 2023.
- [29] Garofolo, John S. Timit acoustic phonetic continuous speech corpus, *Linguistic Data Consortium*, 1993 vol. , no. , pp. , 1993.
- [30] Simic, Christopher, Bocklet, Tobias. Šelf-Supervised Adaptive AV Fusion Module for Pre-Trained ASR Models, vol. , no. , pp. 12787–12791, 2024.
- [31] Graves, A., Fernández, S., Gomez, F. & Schmidhuber, J. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. *Proceedings Of The 23rd International Conference On Machine Learning*. pp. 369–376 (2006)

This figure "fig1.png" is available in "png" format from:

<http://arxiv.org/ps/2409.14494v1>