



From Weak Labels to Strong Results: Utilizing 5,000 Hours of Noisy Classroom Transcripts with Minimal Accurate Data

Ahmed Adel Attia¹, Dorottya Demszky², Jing Liu³, Carol Espy-Wilson¹

¹Electrical and Computer Engineering, University of Maryland,

²Graduate School of Education, Stanford University,

³College of Education, University of Maryland,

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

Abstract

Recent progress in speech recognition has relied on models trained on vast amounts of labeled data. However, classroom Automatic Speech Recognition (ASR) faces the real-world challenge of abundant weak transcripts paired with only a small amount of accurate, gold-standard data. In such low-resource settings, high transcription costs make re-transcription impractical. To address this, we ask: what is the best approach when abundant inexpensive weak transcripts coexist with limited gold-standard data, as is the case for classroom speech data? We propose Weakly Supervised Pretraining (WSP), a two-step process where models are first pretrained on weak transcripts in a supervised manner, and then fine-tuned on accurate data. Our results, based on both synthetic and real weak transcripts, show that WSP outperforms alternative methods, establishing it as an effective training methodology for low-resource ASR in real-world scenarios.

Index Terms: ASR, weakly supervised learning

1. Introduction

Automatic speech recognition (ASR) systems have seen remarkable improvements in recent years, largely driven by the availability of large-scale labeled speech datasets and advances in deep learning. State-of-the-art models like Whisper[1] are trained on more than half a million hours of webscale transcripts. However, in many domains, acquiring high-quality transcriptions remains a costly and time-consuming process, limiting the amount of labeled data available for training. In fact, high-quality transcriptions can cost upwards of \$150 per hour of speech [2, 3]. This high cost results in data scarcity in many domains which hinders the development of robust ASR models for low-resource and domain-specific tasks.

A potential solution is to leverage weak transcriptions, which may be imperfect or noisy with many errors but still retain useful information and can be acquired cheaply. The central question we explore in this work is how to effectively integrate weak transcriptions with limited gold-standard transcriptions to improve ASR performance in low-resource domains. This approach falls under the domain of Weakly Supervised Learning (WSL) [4]. Unlike fully supervised learning, where the model is trained on a precisely labeled dataset, and self-supervised or unsupervised learning where the model is trained on unlabeled data, WSL is a paradigm of machine learning where models are trained on partially labeled or imprecisely labeled data.

In the task of ASR, WSL, particularly inaccurate supervision [5, 6], has enabled the use of imprecise or inaccurate transcription. WSL has proved to be crucial in many ASR models, due to the large costs of accurate transcription. Perhaps the most well-known example of a weakly-supervised ASR model is Whisper, where the large training corpus was not precisely

vetted, and portions of it were known to be inaccurate. However, the model’s contextual capacity and the large size of the dataset proved to be adequate, so that the model can learn well from the noisy labels.

Whisper’s success with inaccurate supervision motivated us to explore the applicability of weak transcripts in low-resource ASR, primarily in classroom speech. Classrooms present a unique challenge in ASR since they are multi-speaker environments with multiple target speakers (teachers and students) whose speech sometimes overlap [7, 8]. Additionally, classrooms are noisy environments characterized by children’s babble noise, which is the noise of multiple background speakers and is considered one of the most difficult noises in ASR [9]. However, there is a huge potential for sufficiently accurate ASR to help in the objective analysis and understanding of classroom dynamics[10, 11, 12].

All of these challenges are complicated by the fact that classroom speech is a low-resource task [13, 14] due to the protections around children [15, 16]. This means great care must be taken to best utilize existing data. One of the largest resources of classroom audio data is the National Center for Teacher Effectiveness (NCTE) classroom speech corpus [17]. The NCTE dataset contains 5000 hours of transcribed classroom speech, but these transcripts are very inaccurate as they were not intended for ASR training. They are not verbatim and all names of the students and teachers were de-identified to protect their privacy. These transcripts have not been used, and a small number of recordings were accurately re-transcribed using human gold-standard transcriptions [18]. While Whisper has shown that inaccurate transcripts are useful if paired with sufficient accurate transcripts, it is not clear if imprecise transcripts in the NCTE dataset would be helpful, given that they far outnumber precise transcripts.

In this study, we explore how to best utilize these weak transcriptions to improve the performance of classroom ASR systems. We preface our experiments using the NCTE corpus with controlled experiments on synthetically corrupted transcripts from the TEDLIUM-3 [19] corpus.

We propose Weakly Supervised Pretraining (WSP), a weakly supervised learning paradigm where we initially use imprecise transcription as an intermediate supervised step, followed by fine-tuning on a limited amount of gold-standard data. Our results indicate that this WSP paradigm is very effective in increasing the utility of limited gold-standard data, outperforming previous methods with limited classroom data.

2. Datasets

2.1. TEDLIUM

We use the TEDLIUM 3 [19] dataset for our synthetic experiments. TEDLIUM is a collection of transcribed recordings of

TED talks amounting to 540 hours from 2028 speakers. We chose TEDLIUM as it is a high-quality and public dataset that provides a good benchmark for our experiments. Unlike Librispeech [20], TED talks are a better representation of everyday spontaneous speech, making them more suitable for experiments with transcription errors.

2.2. NCTE

The NCTE dataset consists of video and audio recordings of 2128 4th and 5th-grade elementary math classrooms [17]. Each classroom is recorded with 2 to 3 microphones that range from lanyards worn by teachers capturing near-field speech to far-field stationary microphones placed in the corner of the classroom. In total, the dataset contains 5235 hours of recordings.

2.2.1. NCTE-Weak

The majority of these recordings were transcribed, but they were not intended for ASR tasks. These transcriptions contain a large number of substitutions and deletions, and all the names of students and teachers were omitted to protect their privacy. Most importantly, the transcripts were not properly timestamped. Loose inaccurate minute marks are provided that are often off by tens of seconds. In addition, in many instances, utterances have been labeled as “unintelligible” or “side conversation” that are still perfectly intelligible to both ASR models and humans. Thus far, these transcriptions have not been used for ASR training, and previous works only utilized them for analysis [17] or for training n-gram language models (LMs) for beam-search decoding [18]. However, in this research, we attempt to utilize these transcriptions as weak transcriptions in intermediate training. We call this dataset **NCTE-Weak**. We pass the dataset through a forced aligner [21, 22] to obtain more accurate timestamps for preprocessing the dataset before ASR training. While the forced aligner gave better timestamps than the ones provided, deletion errors and the mismatch between the forced alignment model and the classroom speech domains resulted in errors in timestamps.

Ground Truth: *So what would your final answer be?*

Provided Transcription: *line, he is using a number line. Good. So then what would your final answer be?*

2.2.2. NCTE-Gold

In previous works[18], a small subset of the recordings were re-transcribed using gold-standard human transcriptions. In this work, we transcribe 11 more classrooms raising the number of gold-standard transcribed classroom recordings to 17 classes, amounting to 13 hours of recordings. We call this subset **NCTE-Gold** and we split it into a training set (10 hours from 13 classes) and a validation set (3 hours from 4 classes).

3. Experiments

In this section, we outline our experiments to better understand the effect of weak transcriptions on ASR performance and showcase the efficacy of WSP. First, we perform an ablation study using the TEDLIUM dataset where we synthetically corrupt increasing portions of the training data with common transcription mistakes. We then apply our findings to real-world scenarios using the NCTE dataset. For both experiments, we train Wav2vec2.0-based models, using the fairseq [23] implementation. We chose to use Wav2vec in our experiments as it does not have an internal LM like Whisper, which allows for better measurement of the effect of weak transcripts on ASR.

3.1. Synthetic Corruption

For the synthetic corruption experiments, we finetune Robustwav2vec [24], as it is the Wav2vec model pre-trained with the largest amount of English speech audio.

3.1.1. Weakly Supervised Pre-training Step

For the intermediate WSP step, we corrupt the transcriptions to model and approximate common human transcription mistakes. We consider 3 types:

- **Deletion:** We randomly drop words from the sentence.
- **Misspellings:** We use the Datamuse API [25] to replace a word with another word that either sounds like it to model hearing mistakes, or one that is spelled like it, to model common spelling mistakes.
- **Timestamp Inaccuracies:** Long recordings are often partitioned into smaller utterances; however, the timestamps for the beginning or end can be inaccurate. We model this inaccuracy by dropping a few words from the beginning and/or end of a sentence or adding a few random words.

We use two methods to corrupt each sentence:

- **Random Corruption:** Where at least one or more of the above mistakes are randomly applied, but the rest of the sentence is mainly correct. Each word had a 5% probability of being deleted, a 20% probability of being replaced by a soundalike, or a common misspelling, and a 5% probability of being repeated. Each sentence had a 50% probability of timestamp inaccuracies.
- **Full Corruption:** To model extreme corruption, every word in the sentence is misspelled or deleted. The rest followed the same paradigm as random corruption.

Below is an example of random and full corruption as compared to the ground truth transcription.

Ground Truth: *Was offered a position as associate professor of medicine.*

Random Corruption: *Okay was offeree a position as associate of medicine.*

Full Corruption: *Coffered a exposition exposition ass assonate professore off medicines.*

For each corruption method, we interpolate the corrupted transcriptions with gold-standard accurate transcriptions to create different training sets with different degrees of corruption. We create 4 training configurations for each corruption method at 25%, 50%, 75% and 100% corruption, where the percentages refer to the number of corrupted utterances in the training set. These configurations model the case where we mix weakly transcribed data with high-quality transcripts, similar to Whisper’s training data. We also train the same model using full, uncorrupted transcriptions as a baseline. We trained the models for 150,000 steps, and all models converged around 70,000 steps.

3.1.2. Precise Fine-tuning

All the models in the previous sections are then fine-tuned using 10 minutes of precisely transcribed and uncorrupted data. This approximates an extremely low-resource scenario.

3.2. Real World Case Study - Classroom Data

For a real-world case study, we consider the NCTE dataset. We use CPT-Boosted Wav2vec2.0[18] which was specifically adapted to classroom speech using the NCTE dataset.

3.2.1. Weakly Supervised Pre-training Step

We first train the intermediate model using the **NCTE-Weak** dataset. We partition the dataset into training and validation splits with no test data using a 90/10 split after removing all the recordings that were re-transcribed in the NCTE-Gold dataset to prevent data leakage. We test that model on the test set from NCTE-Gold. We trained the model using the configuration file used for training Wav2vec on all 960 hours of Librispeech which trains the model for 320,000 steps.

3.2.2. Precise Fine-tuning

We set the model trained on NCTE-Weak as initialization for our precise fine-tuning. We fine-tune that model using the NCTE-Gold training data for 500 steps using a few-shot learning scenario. We compare that model against two base-lines:

- **Direct Fine-tuning:** we directly fine-tune the CPT-Wav2vec2.0 model for 20,000 steps on NCTE-Gold. [18]
- **Self-training:** [26, 27] We use an already fine-tuned model to transcribe a portion of the NCTE-Weak dataset and add it to the NCTE-Gold dataset. This adds 40 hours to the NCTE-Gold dataset. We fine-tune the CPT-Wav2vec2.0 model using this dataset for our second baseline.

4. Results and Discussion

4.1. Synthetic Corruption

We test each model regardless of corruption type or percentage on the uncorrupted TEDLIUM test set. We consider the results with and without LM beam-search decoding. Table 1 gives the Word Error Rate (WER) results for each model.

The first row shows the performance of the model trained on the original dataset without any corruption. The results from the randomly corrupted training sets in the second column, show that while random inconsistent transcription mistakes negatively affect the model’s performance, they are not detrimental overall. We can see that even when half of the training data had transcription mistakes, the performance was barely affected and the degradation in WER was less than 1%. Even when all the transcriptions in the training data had random mistakes, the performance was degraded by about 2% with greedy decoding, and less than 1% with LM decoding.

In contrast, in fully corrupted training sets, where every word in the corrupted sentence was misspelled, the degradation in performance is noticeably higher. In the 25% and 50% corrupted training sets, the model could still perform reasonably well **with LM decoding**, however, the performance with non-LM greedy decoding is noticeably worse. This highlights one important distinction between the two corruption methods, where the gap between greedy decoding and LM beam search decoding is significantly larger with full corruption. This points to the fact that most of the misspelling mistakes are usually caught by the LM and corrected accordingly. While this doesn’t work as well with higher degrees of corruption (75% and 100%), the improvement in performance with LM decoding is still significant, although the utility of the models at this level of performance is limited. Notably, the model still learns even under severe corruption, likely because our method replaces words with phonetically similar alternatives or common misspellings. While this maintains some phonetic alignment with the gold transcription, the resulting WER remains too high for practical use.

The results of fine-tuning the models trained on corrupted data on 10 minutes of accurate gold-standard data are in Table 2. We also fine-tuned the model that was trained on the full

Table 1: WER results from training Wav2vec2.0 with corrupted TEDLIUM transcriptions. “% Corr.” indicates the proportion of utterances corrupted in the training set. Results before the slash (/) are from greedy decoding (no LM), while those after the slash use LM decoding.

% Corr.	Random Corr.	Full Corr.
0	7.40/6.77	
25	7.50/6.82	12.72/8.16
50	7.89/7.06	46.73/10.57
75	8.76/7.03	72.45/ 37.76
100	9.73/7.72	80.36/52.40

Table 2: WER results from finetuning models from Table 1 on 10 minutes of accurate uncorrupted TEDLIUM data. “% Corr.” indicates the proportion of utterances corrupted in the training set of the pretrained model.

% Corr.	Random Corr.	Full Corr.
0	7.27/6.48	
25	7.11/6.50	10.31/7.53
50	7.49/6.65	13.30/8.33
75	8.10/6.55	33.91/19.0
100	8.31/6.86	43.90/ 24.49

original uncorrupted training data as a baseline, which is shown in row 1. We can see that fine-tuning with a small amount of labeled data seems to cancel out the effect of random inconsistent corruption in the original training data with less than 50% random corruption. Further fine-tuning these models matches the performance of the model trained without any corruption in the first row of Table 1. With higher degrees of corruption, we still see some improvement with greedy decoding, but we also see, interestingly, that with LM decoding, the model originally trained with some random corruption in the transcription in 100% of the training labels only slightly underperforms the model trained without any corruption.

With models that were originally trained with fully corrupted labels, fine-tuning with 10 minutes of accurate data helps them become more usable. With LM decoding, even with up to 50% full corruption of the training labels, fine-tuning on a small amount of labeled data reduces the performance gap caused by corruption to below 2%. Even with 75% and 100% corruption in the training data, fine-tuning on 10 minutes of accurate data reduces the WER significantly, by around 40% with greedy decoding, and 25% with LM decoding.

To understand the impact of such results, we note that correcting 10 minutes of weak transcription by a human can take less than an hour, as each minute of transcription can take between 5-8 minutes [28, 29] and can cost around \$25 [2]. However, a model trained with 100% of its training data fully corrupted, a very extreme case, can get a WER as low as 24.49% with fine-tuning on just 10 minutes of accurate data. While 24.49% is a high WER, it can be usable for many low-resource tasks and matches some baselines in classroom ASR as seen in Table 3.

Finally, we attempted to train the model directly using 10 minutes and 1 hour of labeled data without prior initial training, but the model did not converge. We did not include that model in the table, but the reader can assume its error to be 100% since it did not learn to output any characters. While Wav2vec2.0 has been successfully trained with just 10 minutes of Librispeech before, Librispeech is a much cleaner and more straightforward task than TEDLIUM. This highlights an important finding: not only does further fine-tuning on accurate data cancel out the effect of moderate corruption in the transcription, but even severe corruption significantly improves the utility of small precise data. While it remains potentially possible

Table 3: WER results on the NCTE and MPT test sets with different training configurations. NCTE-Weak \rightarrow TED 10-Hr and NCTE-Weak \rightarrow NCTE-Gold refer to models initially trained on the NCTE-Weak dataset and then finetuned on 10 hours of TEDLIUM and the NCTE-Gold dataset respectively.

Training Data	NCTE	MPT
TEDLIUM	55.82/50.56	55.11/50.50
NCTE-Weak	36.23/32.30	50.84/46.09
NCTE-Gold	21.12/16.47	31.52/27.93
NCTE-Self Training	17.45/15.09	27.42/26.24
NCTE-Weak \rightarrow TED 10-Hr	25.59/21.14	42.62/37.22
NCTE-Weak \rightarrow NCTE-Gold	16.54/13.51	25.07/23.70

to train a Wav2vec model using limited amounts of data from TEDLIUM, the fact remains that prior training on severely corrupted data makes this training much more straightforward.

4.2. Real World Case Study - Classroom Data

Looking at Table 3, we see that the model trained on TEDLIUM performs poorly with both test sets which corroborates previous research [18, 30] that shows that ASR models trained on off-the-shelf data struggle with classroom speech. While training on the NCTE-Weak improves the results, it is outperformed by training on NCTE-Gold, a precisely transcribed dataset that is 500 times smaller. Using the resultant model to generate self-training data further improves the results on both datasets. Unlike our synthetic experiments, we do not mix NCTE-Weak with NCTE-Gold for two reasons because NCTE-Gold is too small to make a measurable impact, being only 0.2% the size of NCTE-Weak.

We ran two WSP experiments, both relying on first training on NCTE-Weak followed by precise fine-tuning. First, we consider the case where no in-domain accurate data exists. We use 10 hours of TEDLIUM, which we have already seen to be a poor match for classroom speech as evident from the first row of Table 3. However, few-shot fine-tuning on this dataset improves the results significantly from NCTE-Weak training, by about 10% in both configurations, while still underperforming direct precise fine-tuning on in-domain data. If in-domain accurate data exists, such as NCTE-Gold, we get further improvements, resulting in our best configuration for both NCTE and MPT test sets. In any case, this configuration outperforms both direct fine-tuning and self-training, further showing that initial imprecise training increases the utility of limited precise data.

5. Error Analysis

In this section, we discuss sample transcription from models trained with weak supervision, and those further fine-tuned on small accurate data. With this analysis, we attempt to understand the effect of transcription errors on performance, and why initial large-scale weak pre-training improves performance.

5.1. Synthetic Corruption

The example below shows a sample transcription from the extreme case of 100% of the training data being fully corrupted as the WSP model, and model that was subsequently fine-tuned on 10 minutes of precise data. WER is shown in parentheses.

Ground Truth: *Everybody talks about happiness these days.*
100% Full Corr. (WSP): *e bod tal abou hapne thel da.* (116.67%)
WSP \rightarrow 10 min FT: *Ever body talks about hapines thees das.* (83.34%)

At first glance, the transcription from the WSP model might seem complete gibberish. However, we can see that the model correctly detects some phonemic features from the sentence. For example “bo” matches “everybody”, “ta” matches “talks”,

etc. This can be attributed to the fact that even though the majority of words in the training set were replaced by their soundalikes or common misspellings, these alternative words still match a lot of the sounds in the audio, which the model partially succeeds in detecting. Further precise fine-tuning builds on this weak foundation to output a legible transcription with a few spelling and structure mistakes. These mistakes are usually fixed by LM beam search decoding. In fact, LM decoding reduces the WER for this example to 16.67% with a single substitution error (day \rightarrow das).

5.2. Real World Case Study - Classroom Data

The example below shows a sample transcription from three models, the model directly trained on NCTE-Gold, the model trained on NCTE-Weak (WSP), and the model initially trained on NCTE-Weak and then fine-tuned on NCTE-Gold. The models do not predict casing or punctuation, however they were added to improve legibility.

Reference: *33. You are right. Yours is correct okay. I won. Clean up. Go back to your seat. Do not put anything away. You are going to need it. 54. Faithful, you have to stay there, you can not move because of the camera.*
NCTE-Weak (WSP): *33. You re right. Yours is correct. 54. You have to stay there. n me csthe me. (75.00%)*
NCTE-Gold: *33. You are right. Yours is correct. kay. on one. Cleen up. Go back to your sseat. Do not put anything away, you are going to need it. 54. Faithfueel, you have to stay there, you can not move because of the camera. (15.91%)*
WSP \rightarrow NCTE-Gold: *33. You are right. Yours is correctokay. On one. Cleen up. Go back to your seat. Do not put anything away, you are going to need it. 54. Faithful, you have to stay there, you can not move because the camera. (11.36%)*

We can that the worst performance is when we train with NCTE-Weak. Similarly to how the training data suffered from deletion errors, the transcription skips over large portions of the audio and picks up later. For instance, the phrases “Clean up ... You are going to need it.” are completely missing, followed by correctly transcribing the number 54, and so on. By inspecting the audio, there does not seem to be any pattern to the deleted portions in tone or noise level. However, the model picks up the initial phrases of the audio correctly. The model often gets 0% error for shorter instances. This might highlight some ways the utility of this dataset can be improved further.

However, even with the deletion mistakes, the model often captures portions of the audio correctly, which explains why fine-tuning this model with NCTE-Gold outperforms direct fine-tuning on the same dataset. By comparing both transcriptions, we see that initial WSP improves several aspects of the transcription, such as the misspelling on “clean” as “cleen”, and the substitution of the name Faithful. Fine-tuning on NCTE-Gold does not only fix the deletion mistakes caused by timestamp inaccuracies in NCTE-Weak, but also utilizes the exposure to large amounts of weak data, resulting in improved robustness.

6. Conclusion

In this paper, we introduced Weakly Supervised Pretraining (WSP) as an effective approach for leveraging highly inaccurate transcriptions in low-resource ASR settings. Our synthetic corruption study on TEDLIUM demonstrated that models trained on even severely corrupted transcriptions, when followed by fine-tuning on just 10 minutes of precise data, can achieve usable ASR performance. We extended this insight to real-world classroom speech using the NCTE dataset, showing that WSP on weak transcripts outperforms direct fine-tuning on limited accurate data and even outperforms self-training approaches.

7. Acknowledgment

This work is supported by the Grand Challenge Award at the University of Maryland and through the generous support of the Bill and Melinda Gates Foundation

8. References

- [1] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *International Conference on Machine Learning*, PMLR, 2023, pp. 28 492–28 518.
- [2] J. D. Williams, I. D. Melamed, T. Alonso, B. Hollister, and J. Wilpon, "Crowd-sourcing for difficult transcription of speech," in *2011 IEEE Workshop on Automatic Speech Recognition & Understanding*. IEEE, 2011, pp. 535–540.
- [3] S. Novotney and C. Callison-Burch, "Cheap, fast and good enough: Automatic speech recognition with non-expert transcription," in *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 2010, pp. 207–215.
- [4] Z.-H. Zhou, "A brief introduction to weakly supervised learning," *National science review*, vol. 5, no. 1, pp. 44–53, 2018.
- [5] B. Frénay and M. Verleysen, "Classification in the presence of label noise: a survey," *IEEE transactions on neural networks and learning systems*, vol. 25, no. 5, pp. 845–869, 2013.
- [6] D. C. Brabham, "Crowdsourcing as a model for problem solving: An introduction and cases," *Convergence*, vol. 14, no. 1, pp. 75–90, 2008.
- [7] X. Chang, Y. Qian, K. Yu, and S. Watanabe, "End-to-end monaural multi-speaker asr system without pretraining," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6256–6260.
- [8] X. Chang, W. Zhang, Y. Qian, J. Le Roux, and S. Watanabe, "End-to-end multi-speaker speech recognition with transformer," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6134–6138.
- [9] C. Simic and T. Bocklet, "Self-supervised adaptive av fusion module for pre-trained asr models," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 12 787–12 791.
- [10] D. Demszky, J. Liu, H. C. Hill, S. Sanghi, and A. Chung, "Improving teachers' questioning quality through automated feedback: A mixed-methods randomized controlled trial in brick-and-mortar classrooms," *EdWorkingPapers*, 2023.
- [11] J. Jacobs, K. Scornavacco, C. Harty, A. Suresh, V. Lai, and T. Sumner, "Promoting rich discussions in mathematics classrooms: Using personalized, automated feedback to support reflection and instructional change," *Teaching and Teacher Education*, vol. 112, p. 103631, 2022.
- [12] J. Jacobs, K. Scornavacco, C. Clevenger, A. Suresh, and T. Sumner, "Automated feedback on discourse moves: teachers' perceived utility of a professional learning tool," *Educational technology research and development*, pp. 1–23, 2024.
- [13] A. A. Attia, J. Liu, W. Ai, D. Demszky, and C. Espy-Wilson, "Kid-whisper: Towards bridging the performance gap in automatic speech recognition for children vs. adults," in *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, vol. 7, 2024, pp. 74–80.
- [14] R. Fan, N. Balaji Shankar, and A. Alwan, "Benchmarking children's asr with supervised and self-supervised speech foundation models," in *Interspeech 2024*, 2024, pp. 5173–5177.
- [15] Federal Trade Commission, "Children's online privacy protection rule (coppa)," 1998, accessed: 2024-10-12. [Online]. Available: <https://www.ftc.gov/legal-library/browse/rules/childrens-online-privacy-protection-rule-coppa>
- [16] J. Cao, A. Ganesh, J. Cai, R. Southwell, E. M. Perkoff, M. Regan, K. Kann, J. H. Martin, M. Palmer, and S. D'Mello, "A comparative analysis of automatic speech recognition errors in small group classroom discourse," pp. 250–262, 2023.
- [17] D. Demszky and H. Hill, "The NCTE transcripts: A dataset of elementary math classroom transcripts," in *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, Jul. 2023, pp. 528–538.
- [18] A. A. Attia, D. Demszky, T. Ògúnremí, J. Liu, and C. Espy-Wilson, "Cpt-boosted wav2vec2.0: Towards noise robust speech recognition for classroom environments," in *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2025, pp. 1–5.
- [19] F. Hernandez, V. Nguyen, S. Ghannay, N. Tomashenko, and Y. Esteve, "Ted-lium 3: Twice as much data and corpus repartition for experiments on speaker adaptation," in *Speech and Computer: 20th International Conference, SPECOM 2018, Leipzig, Germany, September 18–22, 2018, Proceedings 20*. Springer, 2018, pp. 198–208.
- [20] 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.
- [21] M. Ashraf, "Mms-300m-1130 forced aligner," <https://huggingface.co/MahmoudAshraf/mms-300m-1130-forced-aligner>, 2025, accessed: 2025-02-19.
- [22] T. e. a. Wolf, "Transformers: State-of-the-art natural language processing," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Association for Computational Linguistics, Oct. 2020, pp. 38–45.
- [23] C. Wang, Y. Tang, X. Ma, A. Wu, D. Okhonko, and J. Pino, "Fairseq S2T: Fast speech-to-text modeling with fairseq," in *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing: System Demonstrations*. Suzhou, China: Association for Computational Linguistics, Dec. 2020, pp. 33–39.
- [24] W.-N. Hsu, A. Sriram, A. Baevski, T. Likhomanenko, Q. Xu, V. Pratap, J. Kahn, A. Lee, R. Collobert, G. Synnaeve, and M. Auli, "Robust wav2vec 2.0: Analyzing domain shift in self-supervised pre-training," in *Interspeech 2021*, 2021, pp. 721–725.
- [25] Datamuse, "Datamuse api," <https://www.datamuse.com/api/>, 2025, accessed: 2025-02-19.
- [26] J.-B. Grill, F. Strub, F. Althé, C. Tallec, P. Richemond, E. Buchatskaya, C. Doersch, B. Avila Pires, Z. Guo, M. Gheshlaghi Azar *et al.*, "Bootstrap your own latent-a new approach to self-supervised learning," *Advances in neural information processing systems*, vol. 33, pp. 21 271–21 284, 2020.
- [27] M.-R. Amini, V. Feofanov, L. Pauletto, L. Hadjadj, Émilie Devijver, and Y. Maximov, "Self-training: A survey," *Neurocomputing*, vol. 616, p. 128904, 2025. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0925231224016758>
- [28] C. B. Fox, M. Israelsen-Augenstein, S. Jones, and S. L. Gillam, "An evaluation of expedited transcription methods for school-age children's narrative language: Automatic speech recognition and real-time transcription," *Journal of Speech, Language, and Hearing Research*, vol. 64, no. 9, pp. 3533–3548, 2021. [Online]. Available: https://pubs.asha.org/doi/abs/10.1044/2021_JSLHR-21-00096
- [29] C. Cieri, D. Miller, and K. Walker, "The fisher corpus: A resource for the next generations of speech-to-text," in *LREC*, vol. 4, 2004, pp. 69–71.
- [30] R. Southwell, W. Ward, V. A. Trinh, C. Clevenger, C. Clevenger, E. Watts, J. Reitman, S. D'Mello, and J. Whitehill, "Automatic speech recognition tuned for child speech in the classroom," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 12 291–12 295.