AUTOMATIC SPEECH RECOGNITION TUNED FOR CHILD SPEECH IN THE CLASSROOM

Rosy Southwell [1], *Wayne Ward* [1], *Viet Anh Trinh* [2], *Charis Clevenger* [1], *Clay Clevenger* [1], *Emily Watts* [1], *Jason Reitman* [1], *Sidney D'Mello* [1], *Jacob Whitehill* [2]

[1] Institute of Cognitive Science, University of Colorado
[2] Worcester Polytechnic Institute

1. ABSTRACT

K-12 school classrooms have proven to be a challenging environment for Automatic Speech Recognition (ASR) systems, both due to background noise and conversation, and differences in linguistic and acoustic properties from adult speech, on which the majority of ASR systems are trained and evaluated. We report on experiments to improve ASR for child speech in the classroom by training and fine-tuning transformer models on public corpora of adult and child speech augmented with classroom background noise. By tuning OpenAI's Whisper model we achieve a 38% relative reduction in word error rate (WER) to 9.2% on the public MyST dataset of child speech – the lowest yet reported – and a 7% relative reduction to reach 54% WER on a more challenging classroom speech dataset (ISAT). We also introduce a novel beam hypothesis rescoring method that incorporates a speed-aware term to capture prior knowledge of human speaking rates, as well as a Large Language Model, to select among hypotheses. We demonstrate the effectiveness of this technique on both publicly-available datasets and a classroom speech dataset.

Index Terms— Automatic Speech Recognition, Child Speech, Language Modeling, Transfer Learning, Transformers

2. INTRODUCTION

The ability to accurately transcribe multiparty speech in school classrooms has the potential to enable automated learning support based on Natural Language Understanding, which can operate unobtrusively during natural dialog, such as a conversational virtual agent to promote collaborative learning among students in the classroom [1,2]. However, prior work has shown that off-the-shelf Automatic Speech Recognition (ASR) systems yield very high error rates (84%) in this context [3]. Factors responsible for the poor ASR performance include the young age of the speakers [4], the high level of background noise [5] and reverb, and multiparty speech. In this paper, we explore techniques to address two of these issues: background noise and children's speech, by fine-tuning or training from-scratch custom ASR models, and using an in-domain tuned large language model (LLM) for rescoring ASR hypotheses.

In order to improve accuracy of ASR on child speech, prior work has targeted the acoustic differences in children's voices, e.g., by compensating for the shorter vocal tract length [6], and augmenting adult speech to more closely resemble child speech for use as training data [7]. Yet with these approaches a domain mismatch remains in terms of the *linguistic* properties of the training data and the application to child speech. ASR models have been trained and evaluated on child speech, e.g. [8, 9], and while such models often improve on adult-trained recognizers, we consider whether accuracy can be improved further by utilizing recent advancements in transformer-based acoustic and language models, including the availability of publicly-available models pretrained on vast amounts of data [10–12].

In this paper we experiment with training and fine-tuning end-to-end ASR models on transcribed recordings of children's speech in the hope of obtaining models that perform better on classroom speech owing to a closer match between training data and intended use. The corpora we use incorporate both the acoustic and linguistic properties characteristic of children's speech. Our approach is the following: **(1)** We use end-to-end transformer models. These have recently shown promise in speech applications [13], especially for shorter utterances [9], and even without an external language model applied during decoding; this indicates the capability of the model to learn acoustic and language representations using a single training objective. **(2)** Since the total sum of transcribed child speech is still low, and transformer-based models extract a better language model the larger the training dataset [13], we apply fine-tuning on child speech following initial training on a larger corpus of adult speech (either high-quality transcriptions of clean speech [14], or a larger but noisier corpus of audio scraped from videos on the web [10]). **(3)** Augmentation of the training data with acoustic noise has been shown to improve deep neural ASR performance by acting as regularization to prevent over-fitting, and specifically improves performance for test audio with matched noise properties by enabling the model to implicitly learn representations of the specific noises [15]. For this reason we incorporate noise augmentation with real classroom background noise, including babble, which is particularly challenging as it occurs in the same frequency bands as the target speech. Finally, **(4)** we explore to what extent optimization of the selection of the beam hypothesis can improve ASR accuracy. To this end, we explore a novel speed-based hypothesis selection approach.

Research questions: (1) How much can child speech recognition accuracy be improved by (a) fine-tuning on child speech and/or (b) data augmentation with realistic background noise? (2) How does the model size impact accuracy, and how does it interact with fine-tuning and/or data augmentation methods? (3) How much can ASR accuracy be improved by intelligently choosing the beam hypothesis?

3. DATA

We used multiple public datasets of transcribed children's speech recordings: MyST, CUKids, and CSLUKids. In addition to these public datasets, we use two corpora of real-world conversational speech among middle school students (age approx. 10-13) in educational settings: ISAT and LEVI. All are described in more detail below. We also make use of TalkMoves [16], and TSCC-v2 [17], two text-only corpora, for fine-tuning the language model.

Database	# Speakers	Hours	Utterances
ISAT (train)	166	2.66	6.5k
MyST (train+dev)	575	139	62.1k
CSLU	498	0.65	644
CU	844	60.3	51.5k

Table 1: Data used for Training/Tuning

Database	# Speakers	Hours	Utterances
ISAT (test)	80	1.26	3066
MyST (test)	91	21	9700
LEVI	33	0.19	429

Table 2: Data used for Evaluation

3.1. Training Datasets

Three public corpora of child speech from US schools were used for training and tuning the ASR systems. **MyST** comprises spontaneous speech of children in grade 3-5 (age approx. 8-11); **CUKids** [18] contains scripted and spontaneous speech from grades K-5 (age 5-11); and **CSLUKids** consists of scripted and spontaneous speech from grades K-10 (OGI corpus; age 5-16 [19]). For MyST, the *train* and *development* partitions were used for training/tuning, whereas all of CSLU and CU corpora were used for training.

ISAT contains working together on science problems while sitting in groups of 4-5 around tables recorded using a table-top microphone. Each classroom contained multiple groups working concurrently. A total of 129 5-minute clips of small-group interaction were transcribed. Recordings were divided into train, development and test sets in approximately 60/10/30% proportion, such that no recording appeared in more than one set and the word error rate (WER) from Whisper [10] was roughly equal across partitions. Utterances were extracted using the transcript timings, and filtered to contain only student speech.

For pretraining on adult speech, we used **LibriSpeech**, a corpus of approximately 1000 hours of read English speech [14]. Finally, for our experiments on speed-aware hypothesis selection, we used **TalkMoves** [16] and **TSCC-v2** [17], with 1.5M/0.2M and 0.2M/0.03M words in train/development sets respectively.

Noise augmentation: Because many public speech corpora – including MyST, CUKids, and CSLUKids – are recorded in quiet environments with headset microphones, we experimented with noise augmentation. In particular, a noise dataset (ISAT noise) was created by extracting portions of the audio files in ISAT that were not contained in transcribed utterances, and concatenating segments from the same source recording. ISAT noise was mixed with all clean-speech training instances (i.e. except for those from the ISAT corpus), with a randomly-chosen segment from a randomly-chosen recording additively mixed with each utterance at a signal-to-noise ratio chosen uniformly between -5 and 20 dB. This range overlaps with noise levels reported in classrooms [5] and signal levels found to be effective for augmenting ASR training data for noisy applications [15]. We tuned separate models with and without noise augmentation.

Preprocessing: All training corpora were filtered to remove utterances shorter than 0.25 sec (the duration of the mel spectrogram generated from the feature extractor) and longer than 30 sec (the input size of the Whisper model). We also removed instances with more than 448 tokens in the transcript, as this is the model's maximum sequence length [10]. Finally, repeated transcripts were filtered out by removing at random all but two utterances with each unique transcript text, as pilot experiments revealed a tendency of Whisper to overfit to the repeated utterances. In total, training/tuning data consisted of 116k utterances totaling 200 hours (Table 1).

3.2. Testing Datasets

Models were evaluated on multiple test sets of children's speech recorded under different conditions (Table 2): MyST test (clean speech, single speaker), ISAT (multi-speaker, noisy speech) and LEVI (headset microphones). The **LEVI** dataset comprises tutoring sessions between a tutor and small groups (max 3) of students, recorded from an online tutoring platform [20]. The audio is intermediate between MyST and ISAT in terms of background noise. The full MyST test set was used for evaluation, although we found errors in 147 utterances of the test partition of the published MyST corpus, so we report results relative to both the original and corrected test transcripts. We will also publish a corrected version of the test set to enable the research community to more accurately evaluate child speech ASR models.

4. EXPERIMENTS

We performed three experiments: (1) Fine-tuning Whisper models; (2) training custom transformer models from scratch; and (3) Speed-aware rescoring with a language model.

Evaluation: We use corpus-level word error rate (WER) for evaluation, i.e. tallying substitution, deletion and insertion counts over utterances, then dividing by the total word count in the corpus. We normalized reference and hypothesis text to minimize the impact of minor formatting differences on WER: removing non-spoken annotations, lower-casing, spelling numbers and contractions in full, and removing punctuation.

4.1. Fine-tuning Whisper on Child speech

As a state-of-the-art (SOTA) starting point for fine-tuning, we used Whisper [10], which is a transformer-based encoder-decoder ASR model trained on 680k hours of multilingual transcribed audio scraped from the web, trained on multiple tasks simultaneously: transcription, translation, voice activity and language detection; determined by prefixing special tokens to the input. As transcriptions in the training data were of variable quality, the training approach is termed "weakly supervised". Nevertheless, the sheer size and diversity of the training dataset results in ASR performance comparable to other SOTA systems, and better robustness to noisy speech. This is promising for transcribing classroom audio, so we evaluated performance of Whisper on various children's speech test corpora, as well as fine-tuned it using the training splits of these corpora, most of which are public.

We fine-tuned Whisper using children's speech corpora in Table 1. We used the implementation of Whisper from the *huggingface Transformers* library [21], specifically the *large-v2* version, with 32 layers of width 1280 and 20 attention heads. This model contains 1.5 billion parameters, so we used low-rank adaptation (LoRA [22]) to reduce the trainable parameters, using the PEFT library [23] with rank $r = 32$, scaling factor $alpha = 64$, and $dropout = 0.05$. LoRA entails learning an adapter model with approximately 15 million parameters for the largest Whisper model. To further reduce compu-

model	ISAT	LEVI	MyST	MyST corrected
large-v2				
pretrained	*57.9*	*46.5*	*15.0*	*13.5*
tuned	**54.0**	38.7	9.4	7.7
tuned (beam search)	61.2	**36.5**	**9.2**	**7.6**
tuned (+aug +noise)	56.0	39.1	9.7	8.1
base				
pretrained	*81.8*	*61.8*	*19.4*	*17.9*
tuned (int8 & LoRA)	85.5	55.5	16.4	14.8
tuned (full)	75.1	64.0	13.7	12.1

Table 3: Word error rates (%) of Whisper models before and after fine-tuning. +aug: augmentation with SpecAugment, +noise: mixed with ISAT noise.

tational requirements, we used int8 quantization [24]. Whisper base, at 1/20th the size of large-v2 (74M parameters) was also tuned with and without LoRA and int8 quantization, both to estimate the impact of these modifications on accuracy, and to compare a similar-size Whisper model to the custom transformer.

Whisper was tuned for a single epoch with a batch size of 32 and learning rate of 1e-4 (3e-6 for *base* withuot LoRA/int8), 50 warmup steps and a linear decay. Tuning Whisper large-v2 for one epoch took approximately 6 hours on an NVIDIA GeForce RTX 3090. We used 112 tokens for the maximum output length of the decoder as the intended application of spontaneous speech in the classroom generally produces very short utterances (ISAT test utterances were on average 1.48 seconds and 21 characters).

Results: By tuning Whisper on clean recordings of kids' speech, relative WER reduction was 7% on ISAT, 38% on MyST, 43% on MyST (corrected), and 17% on LEVI relative to the original *large-v2* model (Table 3). For comparison, the best previously reported test performance [9] on MyST was 16%. Tuning with data augmentation did not improve WER on any test set, and in fact worsened it.

The smaller *base* model configuration had a higher WER on all test sets both before and after tuning, yet tuning still provided an improvement in WER (8% on ISAT, 10% on LEVI, 29% on MyST, 32% on MyST (corrected)). Tuning the model in full was more successful at reducing WER on ISAT and MyST, whereas using int8 quantization and LoRA led to greater improvements in WER on the LEVI test set. For MyST, the tuned *base* model outperformed the pretrained *large-v2* model, suggesting that in relatively clean recordings, fewer parameters are required to model child speech. In contrast, on the noisy classroom recordings (ISAT) the *large-v2* model was clearly superior to the *base* model, even without tuning.

Whisper's decoder functions as a language model, with each output token conditioned on both the encoder's hidden states and prior decoded tokens. With greedy decoding, the token with the highest probability is selected at each time point; in contrast, beam search tracks multiple alternative sequence continuations at each step, and the highest-probability sequence is chosen at the end. We compared greedy and beam-search decoding methods for the model tuned on clean child speech. Beam search decoding gave superior results for LEVI and MyST, with further reductions in WER of 6% and 2% respectively. For ISAT, the greedy decoding strategy was better, and the higher WER of 61.2% for beam-search decoding was particularly due to the insertion rate which increased from 7.8% to 17.2%; inspection of the output revealed a lot of repetition in a small fraction of the hypothesis transcripts, for example – *reference:* What'd you do? *ASR:* Do do do . . . (109 times).

model	ISAT	LEVI	MyST	MyST corrected
tuned	79.6	69.4	14.5	12.9
+aug	80.8	62.2	11.7	10.1
+aug +noise	73.7	65.7	17.8	16.3

Table 4: Word error rates for Custom transformer models pretrained on adult speech and tuned on child speech. +aug: augmentation with SpecAugment, +noise: mixed with ISAT noise

4.2. Training transformers from scratch

Potential downsides of using Whisper as a starting point for a fine-tuned model are that (1) the training dataset is closed, and (2) the training data are noisy, both acoustically and in terms of the quality of transcripts. In a recent comparison of different deep architectures for ASR on childrens' speech, the best performing system on MyST was a Transformer + CTC architecture that was pre-trained on adult speech and tuned with child speech [9]. Building on these results, we experimented with training transformers from scratch using the SpeechBrain Toolkit [25]. In particular, Transformer Acoustic (AM) and Language (LM) models were each trained on a vocabulary of 5000 tokens (as in the pre-trained LibriSpeech models). Models were first pre-trained on the LibriSpeech corpus (as in [9]) and then tuned on MyST, CSLU, and CU corpora (Table 1). The models had 12 encoder and 6 decoder layers of width 512 and 4 attention heads, totalling 71.5M parameters. The LM is applied during decoding by weighted interpolation with the acoustic log-probability (shallow fusion). We further explored data augmentation using SpecAugment [26] and environmental corruption with ISAT Noise.

Results: The Word Error Rate for each custom transformer model is shown in Table 4. The tuned model (without data augmentation) outperformed the pretrained Whisper *large-v2* model on MyST (though not on LEVI or ISAT), thus illustrating how a relatively modest quantity of domain-specific fine-tuning data can, at least sometimes, outperform a more general model trained on much larger datasets. We did not find a consistent benefit of augmentation: SpecAugment improved performance on MyST and LEVI, but increased error rate on ISAT . Adding ISAT noise hurt performance for the MyST and LEVI data, but improved results for the ISAT data.

4.3. Speed-aware rescoring of beam hypotheses

When examining specific utterances that the fine-tuned Whisper models mistranscribed, we found that the set of beam hypotheses for a given input sometimes contained a better hypothesis (w.r.t. ground-truth) than the one that was actually selected. This raises the question: how might we choose a beam hypothesis more effectively, and by how much could this reduce the WER? To explore this question, we developed a novel method for rescoring the best hypotheses with a novel *speech speed-aware* score, along with ASR and LM log probabilities. This approach addresses two common issues in speech recognition: (i) repetitive word output, and (ii) short ASR hypotheses compared to the ground truth, i.e. with a high deletion rate. ASR and language models often assign higher scores to shorter sentences. One solution to mitigate this issue could be to apply a length penalty [27], encouraging the model to favor longer hypotheses to reduce deletion errors; however, this approach could prioritize long but incorrect hypotheses, including those with repetitive words. Our method, however, tackles both deletion and insertion errors, reducing Whisper's hallucination.

Our proposed rescoring scheme strikes a balance between excessively short and long sentences to achieve optimal word counts in

Model	Data	ISAT	LEVI	MyST	MyST-c
pretrained		*57.9*	*46.5*	*15.0*	*13.5*
LP [27]	ID+OD	*56.3*	*44.3*	*15.7*	*14.2*
GPT-2	ID+OD	**56.0**	**43.0**	**13.6**	**12.0**
LLAMA 2	ID+OD	56.1	43.2	13.7	12.0
GPT-2	OD	56.0	43.7	13.7	12.1
LLAMA 2	OD	56.0	43.1	13.7	12.1
oracle		40.0	28.8	10.3	8.7

Table 5: WER using speed-aware rescoring. In-domain (ID) datasets: ISAT, LEVI, MyST. Out-of-domain (OD): TalkMoves, TSCC-v2. LP stands for length penalty method in [27]

utterances. We base our approach on the common human speech rate of 3-4 words per second. By using a parabolic scoring function, we assign higher scores when the ASR hypothesis matches the expected word-per-second rate and lower scores when it deviates. In particular, we define the score

$$s(\boldsymbol{y}|\boldsymbol{x}) = \alpha \frac{\log p_{ASR}(\boldsymbol{y}|\boldsymbol{x})}{|\boldsymbol{y}|} + \beta \log p_{LM} - \gamma(wps_{hyp} - c)^2 \quad (1)$$

where $p_{ASR}(\boldsymbol{y}|\boldsymbol{x})$, $\log p_{LM}$ represent the log probabilities of the ASR and LM respectively. wps_{hyp} denotes the words per second (WPS) of the hypothesis while c serves as a constant representing the typical words per second rate. Additionally, α, β, γ are the weights for the ASR, language model, and speed-aware scores, respectively. We used either GPT-2 [11] or LLAMA-2 [12] as our language model and fine-tuned it on the out-of-domain (OD) datasets, i.e. TalkMoves and The Teacher-Student Chatroom Corpus version 2 (TSCC-v2) [17] as well as in-domain (ID) datasets (ISAT, LEVI and MyST).

Rescoring procedure: Whisper large-v2 (no fine-tuning) was used to generate multiple ASR hypotheses: one via greedy decoding and a set of m=16 hypotheses via beam search decoding. The rescoring candidate is the one with the highest score among the $m+1$ candidates, as determined by Equation 1.

We determine the weights, namely α, β, γ, through a hyperparameter search conducted with Adaptive Experimentation Platform Ax. This search uses a subset of the training set for ISAT, and a development set for LEVI and MyST. As the WER for ISAT is so high, we narrow our focus to a specific portion of the ISAT training utterances with lower WER values. We have selected the value of c to be 3.5, close to the reported average number in [28].For domains with different speech rate characteristics (such as long pauses), we could optimize the parameters (vertex, steepness) of the parabola in the speed-aware rescoring based on the optimal word count from the target domain. We also utilize Ax to identify the optimal values for the α, β parameters in the length penalty method [27].

Results: Our method was effective in reducing WER (compared to the pretrained Whisper *large-v2* model) and the length penalty approach [27] across all datasets (Table 5). The best model is GPT-2 fine-tuned on a combination of OD and ID datasets. The best model's results are as follows: on MyST the WER was reduced from 15.0% to 13.6% and 13.5% to 12.0%, respectively for MyST and corrected MyST. For LEVI, the WER was lowered from 46.5% to 43.0%. Furthermore, even in the challenging far-field recording setting in noisy classroom environments with multiple speakers as in ISAT, the proposed rescoring method reduced the error rate from 57.9% to 56.0%. The best model's hyperparameter (α, β, γ) values are $(0.403, 0.119, 0.946)$, $(1.0, 0.009, 0.094)$, $(1.0, 0.003, 0.90)$ for ISAT, LEVI, and MyST respectively. Additionally, for ISAT, given the diminished quality of beam hypotheses in noisy conditions, we only select beam hypotheses for rescoring when the greedy hypothesis exhibits an ASR probability below 0.5 or a word-per-second rate lower than 1.5 (based on results in [28]). While fine-tuning solely on OD data, LLAMA achieved a lower WER (43.1%) than GPT-2 (43.7%) on the LEVI dataset.

Finally, we also provide the oracle WER, i.e. WER of the hypothesis with lowest WER, which represents the accuracy of the best possible rescorer. On both ISAT and LEVI, oracle WER is substantially lower than any rescoring methods, suggesting that further research into such methods may be fruitful.

5. DISCUSSION

We examined how to improve ASR performance on children's speech in the presence of substantial background noise in the classroom. We used transformer-based encoder-decoder models, both training from scratch and pre-training on adult speech. **Main findings**: (1) Fine-tuning ASR models (pre-trained on adult speech) using children's speech improved ASR performance on both clean and noisy children's speech. (2) Data augmentation with classroom background noise only helped the transformer model accuracy on noisy classroom speech at the cost of performance on clean speech. (3) Increasing the contribution of language models to the decoding process can improve accuracy on child speech, whether by rescoring with an external language model or by using beam search with the fine-tuned Whisper decoder. (4) We attained the best WER reported at time of writing on the full MyST test dataset (but see [29, 30] who report similar WER on cleaned subsets of MyST test).

The accuracy improvements from tuning transformers on child speech could be due to improved modeling of acoustics, linguistic properties of child speech, or both. The end-to-end training of both Whisper and the custom transformer-based acoustic model means that language structure can be implicitly learned by the decoder. By applying a novel rescoring method designed to reduce the generation of excessively short and long hypotheses, we demonstrate that improvements can be made with a tuned language model. Indeed, the lower bound on WER given oracle selection from the greedy hypothesis and top 16 beam hypotheses indicates substantial improvements would be possible on noisy and/or multiparty child speech merely with improved rescoring, highlighting directions for further research.

Real-time performance in the classroom: While we optimized the present models for accuracy, speed is also important. These models require a GPU for inference which is an obstacle for practical use. An interactive educational agent that converses with students would need to provide responses in near-real-time, which is a challenge for transformer-based speech recognition, where the entire input audio is encoded and decoded at once. By extracting utterances in sufficiently short segments, transformer-based ASR can be used in applications where a response on the timescale of a few speaker turns is sufficient. That greedy decoding provides adequate ASR accuracy can be considered an advantage: greedy decoding is faster. Although the best performing models were based on the largest Whisper model, our trained-from-scratch custom transformer outperformed the equivalent-sized Whisper-base, using less than 1% of the training data.

6. ACKNOWLEDGEMENTS

This research was supported by the NSF National AI Institute for Student-AI Teaming (ISAT) under grant DRL 2019805, and also by NSF #2046505. The opinions expressed are those of the authors and do not represent views of NSF.

7. REFERENCES

[1] J. Cao, A. Ganesh, J. Cai, et al., "A Comparative Analysis of Automatic Speech Recognition Errors in Small Group Classroom Discourse," in *Proceedings of the 31st ACM Conference on User Modeling, Adaptation and Personalization*, June 2023, pp. 250–262.

[2] A. Latham, "Conversational Intelligent Tutoring Systems: The State of the Art," in *Women in Computational Intelligence: Key Advances and Perspectives on Emerging Topics*, A. E. Smith, Ed., Women in Engineering and Science, pp. 77–101. Springer International Publishing, Cham, 2022.

[3] R. Southwell, S. Pugh, E. M. Perkoff, et al., "Challenges and Feasibility of Automatic Speech Recognition for Modeling Student Collaborative Discourse in Classrooms," in *International Conference on Educational Data Mining*, 2022.

[4] M. Gerosa, D. Giuliani, S. Narayanan, and A. Potamianos, "A review of ASR technologies for children's speech," *Workshop on Child, Computer and Interaction '09*, p. 7, 2009.

[5] C. S. Howard, K. J. Munro, and C. J. Plack, "Listening effort at signal-to-noise ratios that are typical of the school classroom," *International Journal of Audiology*, vol. 49, pp. 928–932, 2010.

[6] H. K. Kathania, S. R. Kadiri, P. Alku, and M. Kurimo, "A formant modification method for improved ASR of children's speech," *Speech Communication*, vol. 136, pp. 98–106, 2022.

[7] J. Fainberg, P. Bell, M. Lincoln, and S. Renals, "Improving Children's Speech Recognition Through Out-of-Domain Data Augmentation," in *Interspeech*, 2016, pp. 1598–1602.

[8] H. Liao, G. Pundak, O. Siohan, et al., "Large vocabulary automatic speech recognition for children," in *Interspeech*, 2015.

[9] P. G. Shivakumar and S. Narayanan, "End-to-end neural systems for automatic children speech recognition: An empirical study," *Computer Speech & Language*, p. 101289, 2022.

[10] 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*, 2023.

[11] A. Radford, J. Wu, R. Child, et al., "Language models are unsupervised multitask learners," *OpenAI blog*, vol. 1, no. 8, pp. 9, 2019.

[12] H. Touvron, L. Martin, K. Stone, et al., "Llama 2: Open foundation and fine-tuned chat models," *arXiv preprint arXiv:2307.09288*, 2023.

[13] G. Synnaeve, Q. Xu, J. Kahn, et al., "End-to-end asr: from supervised to semi-supervised learning with modern architectures," in *ICML 2020 Workshop on Self-supervision in Audio and Speech*, 2020.

[14] 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)*, South Brisbane, Queensland, Australia, Apr 2015, p. 5206–5210, IEEE.

[15] S. Yin, C. Liu, Z. Zhang, et al., "Noisy training for deep neural networks in speech recognition," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2015, no. 1, pp. 2, 2015.

[16] A. Suresh, J. Jacobs, M. Perkoff, J. H. Martin, and T. Sumner, "Fine-tuning transformers with additional context to classify discursive moves in mathematics classrooms," in *Workshop on Innovative Use of NLP for Building Educ. Applications*, 2022.

[17] A. Caines, H. Yannakoudakis, H. Allen, P. Pérez-Paredes, B. Byrne, and P. Buttery, "The teacher-student chatroom corpus version 2: more lessons, new annotation, automatic detection of sequence shifts," in *Swedish Language Technology Conference and NLP4CALL*, 2022, pp. 23–35.

[18] A. Hagen, B. L. Pellom, and R. A. Cole, "Children's speech recognition with application to interactive books and tutors," *2003 IEEE Workshop on Automatic Speech Recognition and Understanding*, pp. 186–191, 2003.

[19] Shobaki, Khaldoun, Hosom, John-Paul, and Cole, Ronald Allan, "CSLU: Kids' Speech Version 1.1," Nov. 2007.

[20] A. Nickow, P. Oreopoulos, and V. Quan, "The impressive effects of tutoring on prek-12 learning: A systematic review and meta-analysis of the experimental evidence," Working Paper 27476, National Bureau of Economic Research, July 2020.

[21] T. Wolf, L. Debut, V. Sanh, et al., "Transformers: State-of-the-art natural language processing," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, Online, Oct. 2020, pp. 38–45.

[22] E. J. Hu, Y. Shen, P. Wallis, et al., "Lora: Low-rank adaptation of large language models," *arXiv preprint arXiv:2106.09685*, 2021.

[23] S. Mangrulkar, S. Gugger, L. Debut, Y. Belkada, and S. Paul, "Peft: State-of-the-art parameter-efficient fine-tuning methods," `https://github.com/huggingface/peft`, 2022.

[24] T. Dettmers, M. Lewis, Y. Belkada, and L. Zettlemoyer, "LLM.int8(): 8-bit Matrix Multiplication for Transformers at Scale," *arXiv preprint arXiv:2208.07339*, 2022.

[25] M. Ravanelli, T. Parcollet, P. Plantinga, et al., "SpeechBrain: A general-purpose speech toolkit," 2021, arXiv:2106.04624.

[26] D. S. Park, W. Chan, Y. Zhang, et al., "SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition," in *Interspeech 2019*, Sept. 2019, pp. 2613–2617.

[27] H. Futami, H. Inaguma, M. Mimura, S. Sakai, and T. Kawahara, "Asr rescoring and confidence estimation with electra," in *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2021, pp. 380–387.

[28] J. Yuan, M. Liberman, and C. Cieri, "Towards an integrated understanding of speaking rate in conversation," in *Ninth International Conference on Spoken Language Processing*, 2006.

[29] R. Jain, A. Barcovschi, M. Yiwere, P. Corcoran, and H. Cucu, "Adaptation of Whisper models to child speech recognition," in *Proc. INTERSPEECH 2023*, 2023, pp. 5242–5246.

[30] 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," *arXiv preprint arXiv:2309.07927*, 2023.