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

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 classro82.961s all of

model	ISAT	LEVI	MyST	MyST corrected

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:

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.