Classroom-ASR: A Benchmark Dataset and Evaluation of Modern Speech Recognition Systems for Young Children's Speech

Zhongdi Wu*, Yihao Wang*, Eric C. Larson*, Nastaran Ghorbani*, Doris Luft Baker†, Holland P. Kowalkowski†
* Department of Computer Science, Lyle School of Engineering
Southern Methodist University, Dallas, TX, USA
† Department of Special Education and Department of Curriculum and Instruction
College of Education, The University of Texas at Austin, Austin, TX, USA

***Abstract*—Automatic Speech Recognition (ASR) supports scalable assessment and personalized feedback in education, yet current systems perform poorly in authentic classrooms, especially for young children producing spontaneous speech. We introduce a dataset of 16,409 utterances from Grade 2–3 students collected during vocabulary assessments in real classrooms. Recognition is challenging due to developmental variation in pronunciation, non-standard grammar, code-switching, and background noise. Models trained primarily on clean adult speech show high Word Error Rate (WER) in this setting. We establish comprehensive benchmarks spanning conventional ASR systems, commercial APIs and recent speech foundation models, including supervised fine-tuning of Whisper and Qwen3-ASR. Fine-tuned Qwen3-ASR achieves the best performance with a word error rate of 19.7%. Furthermore, we introduce a richer annotation protocol that distinguishes lexical content from non-speech behavior and various types of noises. Nevertheless, spontaneous classroom speech remains considerably more challenging than standard ASR benchmarks.**

***Index Terms*—Automatic speech recognition, Children's speech recognition, speech corpora**

I. Introduction

Automatic speech recognition (ASR) plays a pivotal role in educational settings, enabling a range of downstream applications such as automated assessment of oral reading fluency [1]–[3], vocabulary knowledge evaluation [4], [5], real-time feedback for language learning [6], and large-scale analysis of classroom discourse [7]. Such tools have the potential to support teachers in formative assessment, provide individualized feedback to students, and enable large-scale educational research. Moreover, studies show that students' spoken answers are higher quality and better retained when spoken versus written [8], [9].

Despite its importance, the majority of existing child speech datasets have been collected under reading tasks, prompted dialogues, or tutoring systems [10]–[16]. On these datasets, it is well documented that speech recognition is more difficult, with numerous works reporting word error rates (WERs) from 6% up to 77% ($\mu = 39\%$). However, fine-tuning these models brings the WER to a more manageable range, 1.7%-34% ($\mu = 11.1\%$) [17]. While these resources and fine-tuning have been instrumental in advancing child speech recognition research, they generally capture only subsets of the variability encountered in authentic classroom speech. Task-driven, open-ended student responses contain natural hesitations, self-corrections, code-switching, and occasionally off-topic responses. Moreover, classroom recordings additionally contain background noise, overlapping speakers, and variation in recording hardware.. To bridge this gap, we introduce a new dataset of 16,409 open-ended student responses (roughly 40 hours) collected from vocabulary assessments administered in classroom settings across multiple schools in North America. Unlike prior read-speech corpora, students were prompted to define target words and produce sentences using them, resulting in free-form, task-driven speech. Responses were transcribed through a multi-stage human-in-the-loop process. The dataset was split into five balanced folds, enabling robust training and evaluation for ASR models using traditional pipelines. The initial dataset was annotated using lexical transcriptions (i.e., only the final text) which limits the ability to fully characterize student speech. To address this limitation, we further refined and standardized the transcripts to construct a reliable benchmark for speech recognition research.

Recent speech foundation models have substantially advanced ASR through large-scale pretraining and supervised fine-tuning, achieving strong performance across diverse speech domains. These advances have also improved recognition of children's speech when adapted to age-specific datasets. However, current evaluations are largely limited to read-aloud or otherwise constrained speech, leaving open the question of how well these models generalize to spontaneous classroom interactions characterized by disfluencies, non-standard grammar, code-switching, and environmental noise. We therefore establish a comprehensive benchmark spanning recent speech foundation models, evaluating both zero-shot and supervised fine-tuning performance for systems including Whisper [18]

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305A200521.Code, model checkpoints, dataset access instructions, and supporting artifacts are available at https://github.com/WuZhongdi/classroom-asr.

and Qwen3-ASR [19].

While collecting realistic classroom speech is essential, benchmark quality also depends on transcript quality. Transcripts generated during educational assessment may contain inconsistent normalization, omitted speech, or inserted lexical content, resulting in difficulty in producing reliable transcripts [20]–[22]. Constructing evaluation-quality transcripts is therefore an important but often overlooked step in developing reliable ASR benchmarks. Beyond benchmark construction, understanding why recognition remains difficult requires information beyond lexical transcripts. We therefore annotate an additional collection of classroom recordings with speech event labels that distinguish lexical speech from non-lexical vocalizations and environmental acoustic events. Unlike conventional transcripts, these annotations enable systematic analysis of speech phenomena that contribute to recognition errors, providing complementary insight beyond aggregate WER metrics. These annotations provide insight into characteristics of spontaneous classroom speech that are not captured by conventional WER evaluation. However, because of the expense and time required to annotate recordings in this way, we do not use these data for model training–they are only used for evaluation on a subset of the collected data.

In summary, we make the following contributions:

– We introduce a curated benchmark of 16,409 spontaneous classroom utterances collected from Grade 2–3 vocabulary assessments, together with carefully refined transcripts and standardized five-fold data splits for reproducible ASR evaluation.

– We provide a comprehensive benchmark of modern speech foundation models, including Whisper and Qwen3-ASR under both zero-shot and supervised fine-tuning settings, establishing strong baselines for spontaneous classroom child speech recognition.

– We present a companion fine-grained annotation resource of classroom speech that distinguishes lexical speech, non-lexical vocalizations, and environmental acoustic events. Using this resource, we characterize key properties of spontaneous classroom speech.

II. Dataset

A variety of child speech corpora have been developed to support speech recognition research. Early datasets such as OGI Kids [10], CU Kids [12], and PF-STAR [13] primarily consist of read or prompted speech collected under controlled recording conditions. More recent resources have expanded to include tutoring interactions, language-learning scenarios, and prompted conversational speech [14], [15]. However, the majority of existing corpora focus on constrained speaking tasks in which children read predefined text, repeat prompts, or respond within structured dialogues.

A. Children Free-form Speech Dataset

We originally designed and collected this dataset to support automatic vocabulary acquisition assessment for young English learners. Students completed the assessment online in classroom settings. Each test consisted of either 12 or 24 target words. For each word, students were required to complete two tasks: (1) provide a free-form definition of the word and (2) produce a sentence using the word. All data collection was approved by University IRB(Institutional Review Board), school districts, and the testing procedure was standardized across schools. Researchers either directly supervised the audio collection in the classroom or trained the instructors to guide students through the process. First, a scientific vocabulary word appeared on the screen, accompanied by a text-to-speech (TTS) prompt asking the student to define the word. The student could then record their spoken response by clicking a button on the screen, with the option to pause and listen to their recording before submission. After completing the definition task, the interface presented an image related to the target word along with a second TTS prompt requesting the student to produce a sentence containing the word. The same recording and playback process was followed for this sentence-production task. The test then advanced to the next word until all problems were completed. The method of recording each student was non-standardized and influenced by the available equipment in the classroom—some classes used headsets, while others used commodity ambient microphones. The audio collection was cloud-based, automatically uploading the audio with relevant meta-data.

Student responses were intended to be scored according to the Depth of Knowledge (DOK) framework [23], which measures the goodness of understanding demonstrated in a response. To this end, we collected audio samples of free-form speech from classrooms across multiple schools with the aim of assessing DOK for each student. In total, 38 classrooms of second- and third-grade students underwent data collection, comprising 548 different student speakers. However, during the development of automated scoring pipelines, the accuracy of ASR quickly emerged as a critical deficiency. An online platform was designed and deployed to enable trained human annotators to assess students' knowledge based on ASR-generated transcripts. Annotators consistently reported that commercial ASR systems produced unstable and error-prone transcripts, severely limiting their utility.

In response, we implemented a multi-stage correction process to ensure transcription quality. Human annotators were asked to review and correct ASR-generated transcripts using an interactive interface. A third annotator subsequently performed verification, resolving any discrepancies between the audio and text to produce a final gold-standard transcript. During this process, audio responses that have severe issues such as: overwhelmingly noisy, mixed language or majority Spanish, and multiple students speaking on top of each other are excluded.

In this manner, 16,409 utterances were made available for model training and evaluation (additional measures of the dataset are available in Table I). In the dataset we also store additional metadata about the educational context of the word, including: (1) the golden standard definition for the word, (2) the sentence question prompt, and (3) teacher-written

TABLE I
SUMMARY STATISTICS OF THE ANNOTATED DATASETS. CONTINUOUS VALUES ARE REPORTED AS MEDIAN [5TH, 95TH PERCENTILE]

Statistic	Main Dataset	Rich Annotations
Total hours	39.81	2.29
Audio duration (sec)	5.0 [1.5, 23.9]	2.4 [1.1, 6.0]
Total recordings	16,409	2,717
Number of speakers	548	292
Recordings per speaker	37 [1, 48]	5 [1, 37]
Duration per speaker (sec)	209.3 [4.7, 667.8]	10.3 [1.9, 136.3]

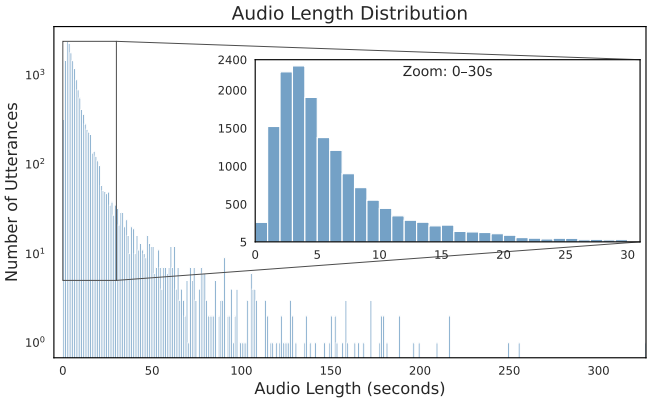


Fig. 1. Length distribution of utterances

example sentences of the word being used. The dataset was split into five stratified folds. The split was designed to keep all utterances from the same student within a single fold, maintain a similar balance of definition and sentence-task utterances across folds, and keep the total number of utterances in each fold approximately equal. It is also worth noting that given these constraints, the word distribution across folds is difficult to perfectly enforce. In this study, fold 0 was kept as held-out set while fold 1-4 were used for training. Figure 1 shows the length distribution of utterances. Overall, the dataset comprises roughly 40 hours of spontaneous classroom speech. IRB protections for student speech restrict the release of the raw audio, but pre-processed, anonymized versions are available for non-commercial research use.

We also adopted a symmetric text normalization procedure applied to both reference (label) and hypothesis (model output). First, text was lower cased. We then removed content inside brackets and parentheses as well as deleted annotation-only tokens that do not belong to the spoken transcript, including unintelligible, inaudible, incomprehensible, clipped, clipped audio, and background noise. Common filler tokens (hmm, mm, mhm, mmm, uh, um) were removed. Afterwards, all characters outside the allowed set of lowercase letters, digits, whitespace, and the punctuation marks . , ' - were stripped. Repeated periods were collapsed, extra whitespace was normalized, and punctuation spacing was standardized.

B. Rich Annotation Subset

While the full dataset provides a large set of clean text transcriptions, we also aim to have a more detailed coding and timing annotation provided. Due to the time and cost of this annotation, an additional set of utterances(with no overlap with the Main set) were annotated in this manner. Each utterance in the Rich Annotation Set is annotated using two complementary layers: **segments** and **masks**. The segment layer represents transcript-bearing child speech and is used to construct the reference transcript. Segment spans are temporally ordered and non-overlapping, and may correspond to either lexical content or predefined non-lexical child vocalizations such as *humming* and *gibberish*. In contrast, the mask layer captures contextual acoustic conditions that are not considered part of the transcript, including background noise, overlapping speakers, adult speech, and Spanish speech. Unlike segments, mask spans may overlap to reflect the simultaneous occurrence of multiple acoustic conditions.

This two-layer annotation scheme explicitly separates what the student says from the acoustic environment in which it is spoken. Child-produced non-lexical vocalizations remain part of the transcript-bearing segment layer when explicitly annotated (e.g., humming or gibberish), whereas environmental interference and competing speech are represented independently through masks. This design enables the Rich Annotation Set to support both transcript generation and future analyses of condition-specific recognition performance, noise-robust training, and speech event characterization.

To ensure annotation quality, annotators first complete a training stage using a curated set of 50 recordings covering all segment categories, mask types, and representative edge cases. Annotation quality is evaluated using a custom agreement metric based on the temporal intersection-over-union of annotated spans weighted by Levenshtein distance between corresponding labels. Only annotators who achieve the required agreement threshold proceed to independently annotate the remaining recordings.

III. METHODS AND EXPERIMENT SETUP

A. Evaluation

All experiments follow the predefined five-fold partition of the benchmark, using folds 1–4 for training and fold 0 exclusively for evaluation. The benchmark contains 16,409 utterances in total, including 3,275 utterances in the held-out evaluation fold. All supervised models were trained using the refined transcript convention described in Section II-A, and identical text normalization was applied to both reference transcripts and model predictions during evaluation. We report both punctuation-aware and punctuation-free WER (Table II), with punctuation-aware WER serving as the primary comparison metric throughout the paper.

B. Context information

Because some additional meta-data about the utterance is available, we also investigate whether the inclusion of additional word context can improve ASR performance. Each recording in the dataset is associated with item-level classroom metadata describing the vocabulary assessment prompt presented to the student. We derive contextual information

from three fields: the target vocabulary word, the task type (definition or sentence), and the corresponding question text. For definition tasks, the question consists of the prompt asking the student to define the target word, while for sentence tasks it contains the prompt requesting the student to use the target word in a sentence. These metadata reflect the information available to the student at the time of recording and therefore provide a realistic source of contextual information for context-aware ASR.

To evaluate inference-time contextual prompting, we consider two prompt formats. The first, question-only, contains only the question text associated with the recording. The second, word+task+question, concatenates the target vocabulary word, task type, and question into a structured prompt. Both prompting strategies are applied only during inference. Importantly, all models are fine-tuned without contextual prompts, allowing these experiments to isolate the effect of inference-time contextual guidance from that of context-aware model training. Unless otherwise specified, baseline systems operate without contextual information and receive only the input audio.

C. Supervised Fine-tune

Modern speech foundation models were adapted to the classroom-asr benchmark using supervised fine-tuning (SFT) on paired audio-transcript data. For all experiments, the supervision signal consisted of the normalized reference transcript, and model parameters were optimized using each model's native auto-regressive sequence objective.

D. Benchmarked ASR Systems

1) Commercial ASR: We evaluate several commercial ASR systems as reference baselines, including OpenAI's GPT-4o and GPT-4o-mini transcription models. All systems were evaluated using their default inference settings without additional adaptation to the classroom domain.

2) HuBERT Transfer Learning Baseline: As a representative conventional ASR baseline, we use HuBERT [24] as the acoustic encoder, followed by Transformer layers and a CTC output layer trained with the CTC objective [25]. This baseline is pretrained on the English subset of Multilingual LibriSpeech (MLS) [26]. Optionally, contextual information from the target vocabulary prompt is incorporated through frozen LLaMA-3.2-1B embeddings [27] using cross-attention, while decoding is performed with beam search and a KenLM language model [28]. The model is first pretrained on MLS and subsequently fine-tuned on the proposed classroom corpus, serving as a strong transfer-learning baseline for comparison with modern speech foundation models.

3) Whisper: We evaluate Whisper Large-v3 Turbo under both zero-shot and supervised fine-tuning settings. Zero-shot inference uses the pretrained checkpoint without adaptation. For supervised adaptation, the entire model is fine-tuned on the training split. A limited hyperparameter sweep over learning rate and training epochs was performed, with the final configuration selected based on evaluation performance on the held-out fold (i.e., best case tuning).

4) Qwen3-ASR: We evaluate Qwen3-ASR-1.7B under both zero-shot and supervised fine-tuning settings. Similar to Whisper, supervised adaptation is performed by fine-tuning the pretrained checkpoint on the training split. A limited hyperparameter search over learning rate and gradient accumulation was conducted, and the final configuration was selected according to held-out evaluation performance.

IV. Results and Analysis

A. Benchmark results

Table II summarizes the benchmark performance of commercial ASR systems, conventional transfer-learning approaches, and recent speech foundation models on the evaluation set. Overall, modern speech foundation models substantially outperform both commercial APIs and conventional CTC-based systems, demonstrating the significant impact of large-scale pretraining and supervised adaptation for spontaneous classroom child speech recognition.

Among the conventional systems, transfer learning improves the HuBERT-CTC baseline from 47.61% to 44.24% WER, while incorporating contextual information further reduces WER to 43.22%. Although these approaches outperform commercial APIs, they remain considerably less accurate than recent speech foundation models.

Whisper Turbo achieves 31.75% WER under zero-shot inference and improves to 26.47% following supervised fine-tuning, confirming that even modest domain adaptation substantially benefits spontaneous child speech recognition. Qwen3-ASR further improves performance, reducing zero-shot WER from 28.66% to 21.71% after supervised fine-tuning. This result highlights the effectiveness of instruction-capable speech foundation models for educational ASR.

Inference-time contextual prompting consistently provides additional gains for Qwen3-ASR without requiring context-aware training. For the zero-shot model, supplying classroom prompts reduces WER from 28.66% to 26.31%, while the fine-tuned model achieves the best overall performance of **21.45%** WER using the question-only prompt. Interestingly, incorporating the target vocabulary word and task type in addition to the question does not yield further improvements, suggesting that the classroom question alone provides sufficient contextual guidance for the model.

Despite these substantial improvements, spontaneous classroom child speech remains considerably more challenging than conventional speech recognition benchmarks. Even the best-performing system exhibits a WER of 21.45%, motivating further analysis of the remaining recognition errors as well as understanding the difficulties of ASR on classroom children speech. That is, these baselines confirm that classroom speech recognition remains an open research problem.

B. Rich Annotation Analysis

1) Mask segment analysis: While the difficulty of the problem is established through our baseline modeling, we use the rich annotation dataset to further elucidate why ASR is so challenging in this domain. As summarized in Table I,

TABLE II

BENCHMARK RESULTS ON THE EVALUATION SET. RESULTS ARE GROUPED BY ASR PARADIGM. WORD ERROR RATE (WER) WITH PUNCTUATION IS REPORTED AS THE PRIMARY EVALUATION METRIC, WHILE PUNCTUATION-FREE WER IS INCLUDED FOR REFERENCE.

Category	Model	Condition	Context	WER (%) Punct.	WER (%) No Punct.
Commercial	GPT-4o Mini	Zero-shot	None	42.45	42.28
	GPT-4o	Zero-shot	None	42.88	43.19
	GPT-4o	Zero-shot	Question	39.13	37.47
Conventional ASR	HuBERT-CTC	From scratch	None	47.61	48.60
	HuBERT-CTC	Transfer learning	None	44.24	44.75
	HuBERT-CTC-Cross Attention	Transfer learning	Question	43.22	43.58
Whisper	Whisper Turbo	Zero-shot	None	31.75	29.23
	Whisper Turbo	Fine-tuned	None	26.47	24.09
Qwen3-ASR	1.7B	Zero-shot	None	28.66	23.56
		Zero-shot	Question	26.62	21.89
		Zero-shot	Word + Task + Question	26.31	21.74
		Fine-tuned	None	21.71	20.04
		Fine-tuned	Question	**21.45**	**19.72**
		Fine-tuned	Word + Task + Question	21.65	19.89

the companion Rich Annotation Set contains 2,717 usable classroom recordings (2.29 hours). Each recording contains an average of 2.80 transcript-bearing segments and 2.35 contextual masks, illustrating that spontaneous classroom responses frequently contain multiple speech events or acoustic conditions within a single utterance.

A notable characteristic of the corpus is the presence of non-lexical child vocalizations. Approximately 28.9% of recordings contain at least one special segment token, with *gibberish* appearing in 24.1% of recordings and *humming* in 7.2%. These phenomena are treated as transcript-bearing child speech rather than environmental interference, allowing them to be analyzed independently or potentially trained to recognize in the future.

Environmental interference is pervasive throughout the corpus. Only 13.1% of recordings contain transcript segments without any accompanying masks, whereas 54.5% contain both *segments* and *masks*. Across all annotations, background noise is the dominant masking condition, followed by mixed speakers, adult speech, and Spanish speech. Figure 2 shows the exact count of different mask sources. Furthermore, 17.1% of recordings contain overlapping mask regions, demonstrating that multiple sources of interference frequently co-occur within the same utterance.

Coverage statistics further illustrate the difficulty of spontaneous classroom ASR. On average, transcript-bearing segments occupy only 33.7% of each recording, while masked acoustic conditions cover approximately 56.2% of the audio. These observations suggest that a substantial portion of classroom recordings consists of environmental interference, competing speech, or other contextual events rather than clean lexical speech, helping explain why spontaneous classroom speech remains considerably more challenging than conventional ASR benchmarks.

2) ASR Performance Across Acoustic Conditions: To investigate the influence of individual acoustic conditions, we

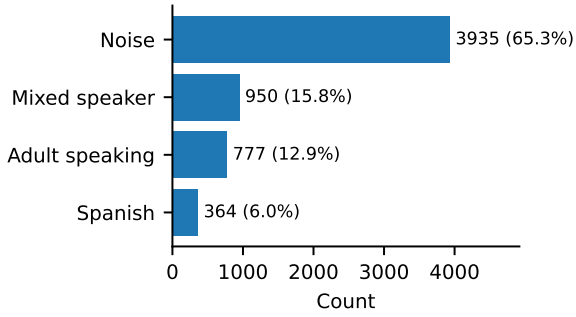


Fig. 2. Distribution of mask annotations in the Rich Annotation Set.

evaluated the best performing fine-tuned Qwen3-ASR model on the Rich Annotation Set using only transcript-bearing lexical segments. That is, the model was fine-tuned **only** on Main Set and evaluated on Rich Annotation Set. Non-lexical child vocalizations (gibberish and humming) were excluded from the reference transcripts, and recordings whose remaining lexical transcript became empty after this filtering were omitted from evaluation. Consequently, only 1,512 of the 2,717 usable recordings contributed to this analysis. Table III summarizes recognition performance grouped by annotated mask conditions.

Recordings without any contextual mask achieved WER at 19.5%, closely matching the overall benchmark performance. Background noise substantially degrades recognition accuracy, increasing WER to 28.1%, while adult speech produces an even larger degradation (34.0%). Interestingly, recordings labeled only as mixed speaker exhibit performance comparable to clean recordings (18.6%), suggesting that limited overlapping speech does not necessarily interfere with recognition when the target child's speech remains dominant. However, combinations involving both adult speech and noise consistently yield the highest error rates (38–39% WER), indicating

TABLE III
QWEN3-ASR PERFORMANCE ON THE RICH ANNOTATION SET GROUPED BY MASK CONDITION. COMBINATIONS OF CONTEXTUAL MASKS WITH FEWER THAN 5 RECORDINGS ARE NOT SHOWN IN THE TABLE.

Mask condition	N	WER No Punct. (%)
None	310	**19.5**
Noise	489	28.1
Mixed speaker	134	18.6
Adult speaking	60	34.0
Adult + Noise	179	38.9
Mixed + Noise	236	30.8
Adult + Mixed + Noise	53	38.3

TABLE IV
NEGATIVE BINOMIAL REGRESSION RELATING ACOUSTIC MASK COUNTS TO ASR EDIT COUNTS.

Predictor	IRR	95% CI (Confidence Interval)	p
Noise	1.381	[1.136, 1.679]	0.001
Mixed speaker	1.028	[0.790, 1.338]	0.836
Adult speaking	1.421	[1.117, 1.809]	0.004

that multiple simultaneous acoustic disturbances compound recognition difficulty.

While the descriptive analysis in Table III compares recognition performance across individual mask combinations, these conditions frequently co-occur in classroom recordings. To estimate the contribution of each acoustic condition while accounting for this overlap, we fit a negative binomial generalized linear model with edit count as the response variable, the log number of reference tokens as an offset, and log-transformed mask counts as predictors to account for variation in audio length. The corresponding Poisson model exhibited substantial overdispersion (variance ratio = 3.79), motivating the use of the negative binomial model. Standard errors were clustered at the speaker level to account for repeated recordings from the same child. Pairwise correlations among predictors were uniformly small (Table V), suggesting that multicollinearity is unlikely to substantially affect the coefficient estimates.

Table IV shows that background noise and adult speech are associated with increased recognition errors after adjusting for transcript length and the presence of other annotated acoustic conditions. Specifically, each unit increase in log-transformed noise count is associated with a 38.1% increase in expected edit count (IRR (Inter Rater Reliability) = 1.38, $p = 0.001$), while adult speech is associated with a 42.1% increase (IRR = 1.42, $p = 0.004$). In contrast, mixed-speaker annotations do not show a statistically significant association with recognition performance (IRR = 1.03, $p = 0.836$). These findings suggest that environmental noise and adult speech are the strongest observed contributors to recognition difficulty among the annotated acoustic conditions.

TABLE V
PEARSON CORRELATIONS BETWEEN LOG-TRANSFORMED MASK COUNTS.

	Noise	Mixed	Adult
Noise	–	0.068	0.198
Mixed		–	-0.039
Adult			–

V. DISCUSSION AND CONCLUSION

Next-generation educational environments are often envisioned as classrooms with pervasive assessment and automation capabilities embedded throughout the learning experience. However, many of these systems implicitly assume that children's speech can be reliably understood under realistic classroom conditions. Our findings reveal a significant challenge to this assumption. While modern speech foundation models substantially outperform conventional and commercial ASR systems, the best-performing model still produces more than one recognition error for every five words spoken, corresponding to a WER exceeding 20%. For educational assessment tasks, where transcription errors can obscure evidence of student understanding or misrepresent vocabulary knowledge, this level of performance remains insufficient for dependable deployment. Analysis of our rich annotation subset further illustrates the difficulty of the problem: nearly 90% of recordings contain at least one contextual acoustic condition, while transcript-bearing speech occupies only one-third of the recording duration. These findings suggest that classroom speech recognition remains substantially more challenging than ASR in conventional read-speech settings, due to the acoustic complexity of classroom environments during assessment. Finally, our results demonstrate both the promise of modern speech foundation models and the considerable gap that remains before speech recognition can serve as a trustworthy foundation for classroom-scale assessment and feedback systems. We hope that the benchmark, dataset, and annotation resources introduced in this work will help spark the development of the next generation of ASR systems capable of meeting the demands of real-world educational environments.

ACKNOWLEDGMENT

The authors thank the participating teachers and MELVA-S team for manually reviewing student responses, Nathan Gage for developing the assessment interface, and the SMU student annotators for annotating the classroom audio. Rich annotations were created using *Clip Annotate*[1], primarily developed by Yihao Wang with contributions from Zhongdi Wu.

OpenAI ChatGPT and Codex were used to assist with coding the rich-annotation interface and coding/debugging the Whisper and Qwen3-ASR fine-tuning pipelines; they were not used for the HuBERT+CTC baseline. ChatGPT was also used for manuscript language editing and rephrasing. All AI-assisted code and text were reviewed and verified by the authors.

[1]https://github.com/YihaoJW/clip-annotate

REFERENCES

[1] A. Loukina, B. B. Klebanov, P. Lange, Y. Qian, B. Gyawali, N. Madnani, A. Misra, K. Zechner, Z. Wang, and J. Sabatini, "Automated estimation of oral reading fluency during summer camp e-book reading with myturntoread," in *Interspeech*, 2019, pp. 21–25.

[2] B. Molenaar, C. Tejedor-Garcia, C. Cucchiarini, and H. Strik, "Automatic assessment of oral reading accuracy for reading diagnostics," in *Interspeech*, 2023, pp. 5232–5236.

[3] Y. Wang, Z. Wu, J. Nese, A. Kamata, V. Nilabh, and E. C. Larson, "Improving oral reading fluency assessment through sub-sequence matching of acoustic word embeddings," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 10 766–10 770.

[4] K. Knill, M. Gales, K. Kyriakopoulos, A. Malinin, A. Ragni, Y. Wang, and A. Caines, "Impact of asr performance on free speaking language assessment," in *Interspeech*, 2018, pp. 1641–1645.

[5] Z. Wu, E. Larson, M. Sano, D. Baker, N. Gage, and A. Kamata, "Towards scalable vocabulary acquisition assessment with bert," in *Proceedings of the tenth ACM conference on learning@ Scale*, 2023, pp. 272–276.

[6] R. H. Abbasi and S. R. Kaukab, "Digital feedback modalities: Real-time vs. delayed feedback in esl reading platforms," *Pakistan Languages and Humanities Review*, vol. 9, no. 3, pp. 181–195, 2025.

[7] R. Southwell, S. Pugh, E. M. Perkoff, C. Clevenger, J. Bush, R. Lieber, W. Ward, P. Foltz, and S. D'Mello, "Challenges and feasibility of automatic speech recognition for modeling student collaborative discourse in classrooms," in *Proceedings of the 15th International Conference on Educational Data Mining*, Durham, United Kingdom, July 2022, pp. 302–315.

[8] F. Luft, A. Miyamoto, and M. Nückles, "Writing or speaking? the role of the production medium in the effectiveness of learning journals," *Learning and Instruction*, vol. 101, p. 102250, 2026. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0959475225001744

[9] C. Hankeln, U. Kroehne, S. Hahn, and S. Prediger, "Fifth and sixth graders speak more richly than they write: Benefits of speech input for digitally assessing conceptual understanding," *Technology, Knowledge and Learning*, pp. 1–27, 2025.

[10] K. Shobaki, J.-P. Hosom, and R. A. Cole, "The ogi kids² speech corpus and recognizers," in *6th International Conference on Spoken Language Processing (ICSLP 2000)*, 2000, pp. vol. 4, 258–261.

[11] A. Kazemzadeh, H. You, M. Iseli, B. Jones, X. Cui, M. Heritage, P. Price, E. Andersen, S. S. Narayanan, and A. Alwan, "Tball data collection: The making of a young children's speech corpus," in *Proceedings of InterSpeech*, Lisbon, Portugal, oct 2005, pp. 1581–1584.

[12] F. Claus, H. G. Rosales, R. Petrick, H.-U. Hain, and R. Hoffmann, "A survey about databases of children's speech," in *Interspeech*, 2013, pp. 2410–2414.

[13] R. Gretter, M. Matassoni, D. Falavigna, K. Evanini, and C. W. Leong, "Overview of the interspeech tlt2020 shared task on asr for non-native children's speech," in *Interspeech*, 2020, pp. 245–249.

[14] R. Gretter, M. Matassoni, D. Falavigna, A. Misra, C. Leong, K. Knill, and L. Wang, "Etlt 2021: Shared task on automatic speech recognition for non-native children's speech," in *Interspeech*, 2021, pp. 3845–3849.

[15] L. Wagner, S. Alghowinhem, A. Alwan, K. Bowdrie, C. Breazeal, C. G. Clopper, E. Fosler-Lussier, I. A. Jamsek, D. Lander, R. Ramnath, and J. Ross, "The ohio child speech corpus," *Speech Communication*, vol. 170, p. 103206, 2025. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167639325000214

[16] S. Pradhan, R. A. Cole, and W. H. Ward, "My science tutor (MyST)–a large corpus of children's conversational speech," in *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, and N. Xue, Eds. Torino, Italia: ELRA and ICCL, May 2024, pp. 12 040–12 045. [Online]. Available: https://aclanthology.org/2024.lrec-main.1052/

[17] R. Jain, A. Barcovschi, M. Y. Yiwere, P. Corcoran, and H. Cucu, "Exploring native and non-native english child speech recognition with whisper," *IEEE Access*, vol. 12, pp. 41 601–41 610, 2024.

[18] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *Proceedings of the 40th International Conference on Machine Learning*, ser. ICML'23. JMLR.org, 2023.

[19] X. Shi, X. Wang, Z. Guo, Y. Wang, P. Zhang, X. Zhang, Z. Guo, H. Hao, Y. Xi, B. Yang, J. Xu, J. Zhou, and J. Lin, "Qwen3-asr technical report," *arXiv preprint arXiv:2601.21337*, 2026.

[20] S. Bates, J. Titterington, and U. {Child Speech Disorder Research Network}, "Good practice guidelines for the analysis of child speech (2nd edition)," Northern Ireland, Oct. 2021.

[21] Z. Yue, M. Barberis, T. Patel, J. Dineley, W. Doedens, L. Stipdonk, Y. Zhang, E. de Witte, E. Loweimi, H. Van hamme, D. Satoer, M. Ruiter, L. M. Velazquez, N. Cummins, and O. Scharenborg, "Challenges and practical guidelines for atypical speech data collection, annotation, usage and sharing: A multi-project perspective," in *Interspeech 2025*, 2025, pp. 3943–3947.

[22] H. Ehlert, E. Beaulac, M. Wallbaum, C. Gebauer, L. Rumberg, J. Ostermann, and U. Lüdtke, "Collecting and annotating natural child speech data – challenges and interdisciplinary perspectives," in *Studientexte zur Sprachkommunikation: Elektronische Sprachsignalverarbeitung 2023*, C. Draxler, Ed. TUDpress, Dresden, Mar. 2023, pp. 72–78.

[23] S. Baker, L. Santoro, D. Chard, H. Fien, Y. Park, and J. Otterstedt, "An evaluation of an explicit read aloud intervention taught in whole-classroom formats in first grade," *The Elementary School Journal*, vol. 113, pp. 331–358, 03 2013.

[24] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, vol. 29, p. 3451–3460, Oct. 2021. [Online]. Available: https://doi.org/10.1109/TASLP.2021.3122291

[25] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *Proceedings of the 23rd International Conference on Machine Learning*, ser. ICML '06. New York, NY, USA: ACM, 2006, p. 369–376. [Online]. Available: https://doi.org/10.1145/1143844.1143891

[26] V. Pratap, Q. Xu, A. Sriram, G. Synnaeve, and R. Collobert, "Mls: A large-scale multilingual dataset for speech research," *ArXiv*, vol. abs/2012.03411, 2020.

[27] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, and G. Lample, "Llama: Open and efficient foundation language models," 2023. [Online]. Available: https://arxiv.org/abs/2302.13971

[28] K. Heafield, "Kenlm: faster and smaller language model queries," in *Proceedings of the Sixth Workshop on Statistical Machine Translation*, ser. WMT '11. USA: Association for Computational Linguistics, 2011, p. 187–197.