FROM ERROR RATES TO EDUCATIONAL MEASURES: BENCHMARKING SPEECH AI FOR TEACHER–CHILD INTERACTION IN PRESCHOOL CLASSROOMS

Anonymous authors
Paper under double-blind review

ABSTRACT

The quality of teacher–child interaction is a central determinant of early development, yet assessing it still depends on trained experts observing classrooms in person, which cannot scale. Speech AI could automate this: transcribe classroom talk, attribute it to the teacher or children, and compute the measures that interaction assessment relies on, such as talk time, turn-taking, and vocabulary. Yet speech systems are evaluated by transcription and diarization error rates, not by whether the measures they produce match those coded by humans. In this paper, we introduce MPCS, to our knowledge the first benchmark that pairs verbatim transcripts with continuous speaker-role timelines over real preschool classroom audio: **227** hours of teacher and child speech (ages 3–5) from **130** classrooms. Evaluating 14 speech recognizers and four speaker-role models, we find that children produce 16% of the characters but 39% of the strongest recognizer's errors, a gap that model scale and in-domain fine-tuning do not close. Lower error rates also do not guarantee better measures: a speaker-role model with 1.4% diarization error still overcounts teacher–child turns by 17%. What matters is the kind of error. Classroom-adapted systems measure how much teachers and children talk to within 1%, but miscount turns and overestimate children's vocabulary. Code is available at this anonymous repository.

1 INTRODUCTION

Language is the primary medium through which teachers and young children interact in preschool classrooms, and the language children hear and use there is linked to their later linguistic, cognitive, and academic development (Whorrall & Cabell, 2016; Weisleder & Fernald, 2013; Gilkerson et al., 2018; Dickinson & Porche, 2011; Côté et al., 2013; Yang et al., 2021). The quality of this interaction, including how often teachers talk with children, whether they ask open questions, and whether they build on what children say, is therefore assessed with standardized observation instruments such as CLASS (Pianta et al., 2008) and SSTEW (Howard et al., 2020). Each assessment, however, requires a trained expert to observe for hours. In a system as large as China's, serving 36 million children in more than 250,000 kindergartens (Li et al., 2026), quality can only be checked occasionally.

Speech AI opens a path toward assessing interaction at scale and giving teachers feedback on their own practice. A recorder worn by the teacher captures an entire morning; automatic speech recognition (ASR) transcribes it, speaker-role diarization labels who spoke when, and measures such as talk time, turn-taking, utterance length, and vocabulary diversity follow automatically (Sun et al., 2024; 2025). These systems are evaluated by character error rate (CER) and diarization error rate (DER), which count how many characters or seconds are wrong and *treat every error as equally costly* (Dutta et al., 2022; Kothalkar et al., 2024; Wang et al., 2003). Educators see it differently. They care about how long children talk, how often teachers and children take turns in the back-and-forth exchanges that extend children's thinking, and how rich the language children hear and use. Each is sensitive to one kind of error and indifferent to the rest. **In other words, not all errors are equal: a lower error rate does not guarantee a better measure.**

The only way to know whether a system's measures can be trusted is to compare them directly with the same measures coded by humans on the same audio. So far, this comparison has been made

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

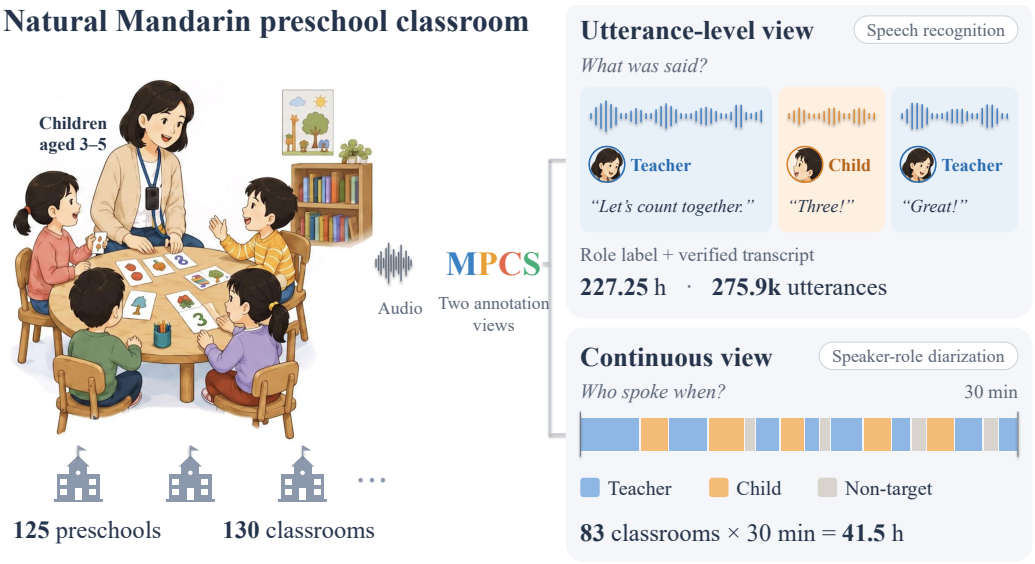

Figure 1: Overview of MPCS. A teacher-worn recorder captures natural Mandarin classroom speech, represented in two forms: segmented teacher and child utterances with verified transcripts and role labels for speech recognition, and a continuous subset with speaker-role timelines for diarization.

in a single preschool classroom (Pelfrey et al., 2024) and in home recordings (Cristia et al., 2021). Answering it more broadly requires data that barely exist: naturalistic recordings of teacher–child interaction in real preschool classrooms, annotated with what was said and who was speaking at every moment. Such recordings require consent from families of three- to five-year-olds and the cooperation of many preschools, and must follow classrooms through whole mornings of instruction, play, and routines. Their annotation is equally demanding, since transcribing young children's speech reliably requires trained listeners, and telling whether a teacher or a child is speaking in noise calls for annotators familiar with early childhood. Existing child-speech corpora therefore mostly record isolated utterances or one-on-one caregiver conversations (Zhou et al., 2025; 2026; Deng & Yip, 2018; Kalanadhabhatta et al., 2024); none pairs verbatim transcripts with speaker-role timelines over classroom audio (Table 1).

To fill this gap, we introduce **MPCS** (Mandarin Preschool Classroom Speech), a benchmark of real preschool classroom interaction (Figure 1). Teachers in 130 classrooms across 125 preschools in two regions of China wore recorders through ordinary mornings, capturing their own speech and that of children aged 3 to 5. Trained annotators with backgrounds in early childhood education produced 227 hours of verbatim, role-labeled transcripts, and a 41.5-hour subset adds continuous teacher and child timelines over the same audio, so that every classroom measure can be computed once from human annotation and once from automatic output; double annotation confirms the reliability of both transcripts and role labels.

We evaluate 14 speech recognizers, including Whisper (Radford et al., 2023), SenseVoice (An et al., 2024), Qwen3-ASR (Shi et al., 2026), and large audio-language models such as Qwen3-Omni (Xu et al., 2025c), together with four speaker-role models, including BabyHuBERT-VTC (Charlot et al., 2025) and WSW 2.0 (Sun et al., 2025). Aggregate error rates hide three findings. *First, recognition fails on children, and scale does not eliminate the teacher–child gap.* Children produce 16% of the characters but 39% of the errors of the best recognizer; a 30-billion-parameter audio model matches that recognizer on teachers but trails it by more than twelve points on children, and more classroom training data widens the ratio between child and teacher error. *Second, near-perfect diarization does not give near-perfect interaction measures.* A speaker-role model with 1.4% DER still reports 17% more teacher–child turns than human annotators, and the off-the-shelf model with the lowest DER reports four times as many. *Third, transcripts too noisy to read can still yield correct measures.* With more than half of children's characters misrecognized, a recognizer still recovers the average length of their utterances to within half a character, while even the best recognizer overestimates how many different words children use. One explanation accounts for all three: recognizers mostly mishear

Table 1: Comparison of child-speech corpora. Ages are in years (y;m: years;months; K: kindergarten; G: grade); –: not reported.

Corpus	Language	Age	Hours	Setting	Continuous	Transcription	Role diarization	Availability
ChildMandarin (Zhou et al., 2025)	Chinese	3–5	41.3	Conversation	No	Yes	No	Yes
ChildTalk (Zhou et al., 2026)	Chinese	2–8	112.5	Child–caregiver	Yes	Yes	No	Yes
CASS CHILD (Gao et al., 2012)	Chinese	1–4	631	Child–parent	Partial	Partial	No	No
SingaKids-Mandarin (Chen et al., 2016)	Chinese	7–12	75	Reading	No	Yes	No	Yes
SLT-CSRC C1 (Yu et al., 2021)	Chinese	7–11	28.6	Reading	No	Yes	No	No
SLT-CSRC C2 (Yu et al., 2021)	Chinese	4–11	29.5	Conversation	No	Yes	No	No
Tong Corpus (Deng & Yip, 2018)	Chinese	1;7–3;4	22	Child–caregiver	Yes	Yes	No	Yes
MyST (Pradhan et al., 2024)	English	G3–G5	393	Virtual tutor	No	Partial	No	Yes
Playlogue (Kalanadhabhatta et al., 2024)	English	3–5	33	Adult–child play	Yes	Yes	Yes	Yes
CSLU Kids (Shobaki et al., 2000)	English	K–G10	–	Prompted + spontaneous	No	Yes	No	Yes
PF-STAR (Batliner et al., 2005)	English	4–14	14.5	Reading	No	Yes	No	No
CHIEDE (Garrote Salazar & Guirao Miras, 2008)	Spanish	3–6	∼8	Classroom/interview	Yes	Yes	No	Yes
Swedish NICE (Bell et al., 2005)	Swedish	8–15	∼6	Child–computer	Yes	Yes	No	No
MPCS (ours)	**Chinese**	**3–5**	**227.25**	**Preschool classroom**	**Yes**	**Yes**	**Yes**	**Yes**

children's words rather than drop them, which leaves length intact but corrupts vocabulary, and speaker-role models mostly insert brief segments of the other role into continuous speech, which leaves talk time intact but inflates turn counts. Controlled perturbations of the human annotation reproduce both effects at a fixed error rate: at 2% DER, splitting the same role confusion into shorter segments raises the turn overcount from 17% to 146%. Which measures can be automated is therefore decided by the kind of error a system makes, not by how many.

Our contributions are as follows:

- **Reframing evaluation for educational use.** We argue that speech AI for classroom assessment should be judged by whether its measures agree with human coding, not by transcript accuracy alone, and operationalize this as measure-level agreement against a paired human reference.
- **A rare paired resource.** To the best of our knowledge, MPCS is the first benchmark to pair verbatim transcripts with continuous speaker-role timelines over real preschool classroom audio, covering children aged 3–5 in 130 classrooms across 125 preschools. It will be released upon acceptance under a data use agreement for research on child speech and scalable classroom assessment.
- **Evidence that changes how systems should be evaluated.** Classroom child speech is a bottleneck that resists model scale, in-domain data, and existing child-speech corpora, and standard error rates rank systems differently from the measures educators use. We map which classroom measures current systems already support, which require calibration, and which await progress on child speech, and recommend reporting error types by speaker role alongside CER and DER.

2 Related Work

Child speech corpora Recent corpora span a growing range of ages, languages, and interaction settings, as summarized in Table 1. Mandarin resources cover segmented child utterances recorded in quiet conditions (Zhou et al., 2025), dyadic child–caregiver conversation (Zhou et al., 2026; Deng & Yip, 2018; Gao et al., 2012), and read or prompted speech from school-age children (Chen et al., 2016; Yu et al., 2021). English and European resources add tutoring, reading, classroom, and play settings (Pradhan et al., 2024; Shobaki et al., 2000; Batliner et al., 2005; Garrote Salazar & Guirao Miras, 2008; Bell et al., 2005), with Playlogue benchmarking both transcription and diarization on adult–child play (Kalanadhabhatta et al., 2024). Apart from Playlogue, these corpora carry either utterance transcripts or continuous speaker segmentation, not both over the same audio; MPCS provides both over identical preschool classroom recordings.

Classroom speech processing Classroom pipelines pair speaker classification with transcription. Dutta et al. (2022) characterized child ASR in naturalistic preschool classrooms across spontaneous speech, activities, and noise, and Sun et al. (2024) combined speaker classification with Whisper transcription to estimate language features from teacher and child speech, which their WSW 2.0 framework scaled with a wav2vec 2.0 classifier (Sun et al., 2025). Speaker diarization for child–adult interaction has progressed from i-Vector systems (Kothalkar et al., 2019) to room-independent ResNets (Kothalkar et al., 2024), speech foundation models (Xu et al., 2024), simulated training conversations (Xu et al., 2025a), and child-centered pretraining (Charlot et al., 2025); whether framed as

child–adult diarization or voice-type classification, these systems assign time intervals to participant categories rather than to individual speakers, the formulation MPCS adopts. Related evaluation sets cover educational speech in English (Lileikyte et al., 2022) and continuous multi-speaker Mandarin audio (Yu et al., 2022). All of this work is reported with character and diarization error rates; whether the measures computed from the output match those computed by hand is not part of the evaluation.

Automatic classroom measurement Classroom language research already derives measures from transcripts: utterance counts, mean length of utterance, lexical diversity, and syntactic complexity characterize children's language experience and development (Chan et al., 2023; Brinchmann et al., 2023), and long-form recordings extend them to all-day estimates of language input linked to children's language ability (Mitsven et al., 2022). ASR-based pipelines have begun to produce such measures directly in preschool classrooms, including mean length of utterance, lexical diversity, question asking, and response rates (Sun et al., 2024; 2025), and classroom discourse research has automated broader teacher-talk features (Wang et al., 2014; Dale et al., 2022; Suresh et al., 2022). Accuracy against ground truth has been checked for thirteen measures in one preschool classroom (Pelfrey et al., 2024) and, for the LENA system, over daylong home recordings (Cristia et al., 2021). What remains open is how recognition errors, speaker-role segmentation errors, and the measures computed from them relate to one another across many systems and many classrooms. MPCS supplies the paired reference and establishes which measures survive the errors current systems make.

3 The MPCS Benchmark

3.1 Data Collection

MPCS was collected from 130 classrooms in 125 preschools across Shenzhen (Guangdong) and Guizhou, China. Class names place 35 classrooms in the junior band, 37 in the middle band, and 58 in the senior band (nominal ages 3, 4, and 5). The recordings cover whole-class instruction, group activities, free play, outdoor activities, and daily routines.

Each teacher wore an iFLYTEK H1 Pro recorder that captured teacher and child speech through the same device across all activities. Teacher speech is captured close to the microphone, while child speech arrives from the varying distances of ordinary classroom interaction. Segmentation, transcription, and quality filtering of the 201 original recordings, 569.97 hours in all, yield 227.25 hours of role-labeled teacher and child speech, which form the segmented ASR corpus; the continuous subset is drawn from the same recordings (Appendix A). Recording was approved by an institutional review board and conducted with informed consent from teachers and from children's parents or legal guardians (Ethics Statement).

The recordings support two complementary benchmark views: a segmented view of utterance-level transcripts and speaker-role labels for speech recognition, and a continuous view of uninterrupted classroom excerpts for speaker-role diarization and role-attributed transcription (Section 3.4). Together they support measure-level evaluation of any recognizer or speaker-role model against a human reference, training and adaptation for child and classroom speech, and calibration of automated classroom measures against human coding; the release is the complete collection, with all partitions, evaluation code, and baseline configurations.

3.2 Annotation

Segmented speech. Annotators assign each utterance a speaker-role label and correct its automatic draft into a verbatim transcript. The transcripts represent intelligible Mandarin speech produced by teachers and children while preserving disfluencies, repetitions, filled pauses, repairs, and spoken numbers (Appendix A.2).

Continuous timelines. The continuous timeline labels every interval as teacher, child, or non-target, with multiple speakers of the same category sharing one role track, matching the level at which classroom measures are defined (Appendix A.3).

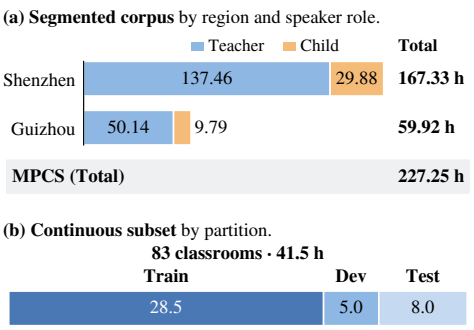

Figure 2: Corpus composition of MPCS.

Table 2: Regional composition of MPCS.

Region	Preschools	Classrooms	Recordings	Raw h	Utterances	Retained h
Shenzhen	97	100	122	414.17	194,461	167.33
Guizhou	28	30	79	155.80	81,405	59.92
Total	**125**	**130**	**201**	**569.97**	**275,866**	**227.25**

Table 3: Partition composition of MPCS.

Partition	Utterances			Duration (h)			Avg. (s)
	Teacher	Child	All	Teacher	Child	All	
Train	157,920	36,964	194,884	134.45	27.60	162.05	2.99
Dev	28,902	7,235	36,137	25.06	5.38	30.44	3.03
Test	35,801	9,044	44,845	28.09	6.68	34.77	2.79
Total	**222,623**	**53,243**	**275,866**	**187.59**	**39.66**	**227.25**	**2.97**

3.3 Quality Control

The dataset was annotated by 22 undergraduate and graduate students with backgrounds in early childhood education, who received unified training. To assess annotation consistency, approximately 5% of the utterances were drawn by role-stratified random sampling and each annotated independently by two annotators without access to the automatic draft. Inter-annotator CER is below 1.3% for both teacher and child speech, and utterance-level speaker-role labels reach a Cohen's κ of 0.83. All disagreements were resolved by a third annotator, whose adjudicated transcript is the released reference. For the continuous timeline, an independent from-scratch re-annotation of the 8-hour test partition yields an inter-annotator DER of 1.2%.

3.4 Corpus Statistics

Figure 2 and Tables 2 and 3 give the regional and partition composition of the 227.25 hours and 275,866 utterances; raw hours include silence and environmental sound, while retained hours cover the final teacher and child utterances.

Two properties shape how the benchmark is used. Teacher speech accounts for 82.5% of the retained duration and child speech for the remaining 17.5%, a share that stays between 17% and 19% in every partition, so scores have to be read by role. Utterances are also short and spontaneous, averaging 3.0 s and 9.3 characters, and child utterances are shorter than teacher utterances on both counts. The continuous subset draws one 30-minute excerpt from each of 83 classrooms and is partitioned by preschool, consistent with the segmented partitions, so that no preschool contributes to more than one partition in either view.

4 Experiments

Our experiments compare each classroom measure computed from system output with the same measure computed from the human annotation. A classroom pipeline can err at two stages: the recognizer can get the words wrong, and the speaker-role model can misplace utterance boundaries or roles. We first evaluate recognition on its own, by speaker role, error type, and training source (Section 4.1), and compute the measures on the human utterance boundaries, where only the transcripts differ from the reference (Section 4.2). We then replace these boundaries with automatic speaker-role timelines, which adds segmentation errors (Section 4.3).

4.1 Recognition Errors Fall on Children

We evaluate 14 systems on the segmented test partition, where reference utterance boundaries isolate recognition from segmentation (Table 4). The systems cover three routes to classroom ASR: Transformer (Vaswani et al., 2017), Conformer (Gulati et al., 2020), and Branchformer (Peng et al., 2022) trained from scratch on MPCS; seven pretrained ASR checkpoints (Zhang et al., 2022; Radford et al., 2023; An et al., 2024; Shi et al., 2026) before and after full-parameter fine-tuning on MPCS; and four large audio-language models (LALMs; Xu et al., 2025b;c; KimiTeam et al., 2025; Wu et al., 2025) under zero-shot inference. MPCS was collected in 2025–2026 and is first released

Table 4: Speech recognition results on the MPCS test partition (CER, %). Gap: child minus teacher CER in the MPCS-trained condition, or zero-shot for models not fine-tuned. ∼: nominal parameter count. Bold: best per column within each condition.

		Zero-shot			MPCS-trained			
System	Params	Overall	Teacher	Child	Overall	Teacher	Child	Gap
From scratch								
Transformer (Vaswani et al., 2017)	29.57 M	–	–	–	20.64	15.75	45.81	30.06
Conformer-31M (Gulati et al., 2020)	31.72 M	–	–	–	20.01	15.05	45.53	30.48
Branchformer (Peng et al., 2022)	29.63 M	–	–	–	20.12	15.22	45.40	30.18
Pretrained ASR								
WenetSpeech-Conformer (Zhang et al., 2022)	135.15 M	67.99	63.07	93.32	13.41	9.43	33.90	24.47
Whisper-small (Radford et al., 2023)	∼244 M	39.83	33.42	72.90	15.75	11.18	39.30	28.12
Whisper-medium (Radford et al., 2023)	∼769 M	29.83	23.93	60.23	13.24	9.15	34.30	25.15
Whisper-large-v3 (Radford et al., 2023)	∼1.55 B	28.31	22.59	57.81	12.80	8.78	33.53	24.75
SenseVoice-Small (An et al., 2024)	∼234 M	19.74	15.38	42.17	11.69	8.05	30.44	22.39
Qwen3-ASR-0.6B (Shi et al., 2026)	782.43 M	13.88	10.19	32.91	11.77	8.10	30.73	22.63
Qwen3-ASR-1.7B (Shi et al., 2026)	2.04 B	**11.80**	8.60	**28.32**	**10.61**	**7.19**	**28.22**	21.03
Large audio-language models (zero-shot)								
Qwen2.5-Omni-7B (Xu et al., 2025b)	∼7 B	15.57	10.68	40.77	–	–	–	30.09
Qwen3-Omni-30B (Xu et al., 2025c)	∼30 B	13.99	**8.53**	42.10	–	–	–	33.57
Kimi-Audio-7B-Instruct (KimiTeam et al., 2025)	∼7 B	35.45	27.56	76.11	–	–	–	48.55
Step-Audio 2 mini (Wu et al., 2025)	∼8 B	18.75	12.90	48.88	–	–	–	35.98

with this paper, so zero-shot scores measure generalization to classroom speech unseen during pretraining. CER is computed at the corpus level over all, teacher, and child utterances; training recipes and hyperparameters are given in Appendix B.

Overall error rate is dominated by teacher speech. Teachers contribute 83.8% of the reference characters, so the overall score moves five times as much with teacher CER as with child CER. For the strongest zero-shot system, Qwen3-ASR-1.7B, children account for 39.0% of all character errors while producing 16.2% of the characters. The two scores can also move apart: fine-tuning Qwen3-ASR-1.7B lowers overall error from 11.80% to 10.61% while child error moves only from 28.32% to 28.22%, and every LALM trails it by more than twelve points on children, including Qwen3-Omni-30B, which matches it on teachers (8.53% against 8.60%). This means that an aggregate score largely reports how well a system recognizes teachers, and that progress on classroom ASR has to be read by role.

In-domain supervision closes the gap between systems but not the gap between roles. Fine-tuning on MPCS improves all seven pretrained checkpoints and lowers child CER by 24–34 points across the three Whisper scales; the first 5% of the training partition, about eight hours, already yields four fifths of that gain (Appendix C.4). What supervision does not move is the floor: child CER across the seven pretrained systems spans 65 points before fine-tuning and 11 points after it, every system lands between 28% and 39% on the test utterances, and the teacher–child gap stays between 21 and 28 points. More classroom data widens the ratio between child and teacher error, from 1.5–3.3 before adaptation to 3.5–3.9 at full supervision.

The teacher–child gap holds across training source, signal-to-noise ratio, and age band. The three models in Table 4 trained on classroom speech alone separate teachers from children by 30 points, more than any fine-tuned pretrained model. Reweighting child utterances to the teacher signal-to-noise distribution leaves the gap unchanged (Appendix C.2), and even for the oldest children it remains at least 20 points (Appendix C.5). What characterizes the gap is its error profile: substitutions account for 70–78% of it in nine of the ten dedicated systems.

We compare classroom supervision with the closest existing Mandarin child-speech supervision, the child–caregiver conversations of ChildTalk (Zhou et al., 2026), under a matched training budget: each of the seven pretrained ASR models is fine-tuned on the 21.8-hour official ChildTalk training set and on a 21.8-hour subset of the MPCS training partition sampled by region and role, and evaluated on child speech from both test sets (Table 5).

Classroom recordings cannot be replaced by the closest existing child-speech corpus. Transfer is asymmetric. Models fine-tuned on classroom speech stay close to setting-matched training on conversational child speech, all seven within 1.8 to 5.7 points of it, and five of the seven improve on their zero-shot scores. In the other direction six of the seven fall 14.7 to 22.5 points behind their

Table 5: Cross-setting transfer: child CER (%) on the ChildTalk and MPCS test sets for zero-shot checkpoints and after fine-tuning on 21.8 hours of the test corpus (Matched) or of the other corpus (Cross). Δ: Cross minus Matched. Lower is better. Bold: best per column.

	ChildTalk test set				MPCS test set			
System	Zero-shot	Matched	Cross	Δ	Zero-shot	Matched	Cross	Δ
WenetSpeech-Conformer (Zhang et al., 2022)	81.82	27.98	33.70	5.72	93.32	40.17	54.86	14.69
Whisper-small (Radford et al., 2023)	50.74	28.82	33.12	4.30	72.90	44.04	66.36	22.32
Whisper-medium (Radford et al., 2023)	41.28	23.93	28.12	4.19	60.23	37.68	58.06	20.38
Whisper-large-v3 (Radford et al., 2023)	37.57	22.62	27.30	4.68	57.81	37.13	59.62	22.49
SenseVoice-Small (An et al., 2024)	23.08	18.50	20.33	1.83	42.17	31.88	**36.16**	4.28
Qwen3-ASR-0.6B (Shi et al., 2026)	19.73	20.71	22.50	1.79	32.91	31.98	53.62	21.64
Qwen3-ASR-1.7B (Shi et al., 2026)	**17.04**	**17.82**	**19.99**	2.17	**28.32**	**29.05**	49.91	20.86

Table 6: Language measures from human and automatic transcripts on the continuous test partition. Cells give classroom means of MLU and MATTR-50, with each system's bias relative to the human transcript in color (characters for MLU, percentage points for MATTR-50). ICC(A,1) and confidence intervals are given in Table 18.

		Child		Teacher	
Transcript	Child CER	MLU	MATTR-50	MLU	MATTR-50
Human	–	9.84	66.6	14.48	69.7
Zero-shot					
Qwen3-ASR-1.7B (Shi et al., 2026)	20.95	9.73 −0.10	69.9 +3.3	14.48 +0.00	70.5 +0.8
SenseVoice-Small (An et al., 2024)	29.54	9.98 +0.15	70.5 +3.9	14.81 +0.32	71.0 +1.3
Whisper-large-v3 (Radford et al., 2023)	45.50	9.37 −0.46	76.6 +10.0	13.94 −0.54	73.2 +3.5
Whisper-small (Radford et al., 2023)	57.19	9.48 −0.35	77.4 +10.7	13.82 −0.67	74.9 +5.2
MPCS fine-tuned					
Qwen3-ASR-1.7B (Shi et al., 2026)	21.29	9.56 −0.27	68.9 +2.3	14.41 −0.08	70.3 +0.6
SenseVoice-Small (An et al., 2024)	22.60	9.51 −0.32	71.5 +4.9	14.42 −0.06	71.0 +1.3
Whisper-large-v3 (Radford et al., 2023)	26.17	9.60 −0.23	69.8 +3.2	14.41 −0.07	70.6 +0.9
Whisper-small (Radford et al., 2023)	31.94	9.64 −0.20	72.1 +5.5	14.41 −0.08	71.0 +1.3

MPCS-matched counterparts on classroom children, and both Qwen3-ASR models end up worse than they had been before any fine-tuning. The pattern holds across regions within MPCS and with the full MPCS training partition (Appendices D.1 and D.2).

4.2 Which Measures Survive Recognition Errors

For each classroom of the continuous test partition (16 classrooms from 12 preschools), the same measure is computed from the human transcript and from a system transcript over the same human-annotated utterances and roles, so that only the text differs. Classrooms are the unit of analysis, and we report classroom means and the mean system-minus-human bias, with the absolute-agreement intraclass correlation ICC(A,1) and 95% confidence intervals from a preschool-clustered bootstrap in Appendix E.2. Two measures stand for two families used in developmental research: mean length of utterance (MLU, characters per utterance), an index of language production, and the moving-average type–token ratio over 50-word windows (MATTR-50), an index of lexical diversity. Table 6 reports both for eight transcription conditions spanning child CER from 21% to 57%.

Table 7: Child MLU by class band (nominal ages 3, 4, 5) on the segmented test partition, human and MPCS-fine-tuned transcripts.

Transcript	Junior	Middle	Senior	Sr.−Jr.
Human	6.09	7.00	7.12	1.02
WenetSpeech-Conf. (Zhang et al., 2022)	5.83	6.73	6.83	1.01
Whisper-small (Radford et al., 2023)	6.02	7.02	7.05	1.04
Whisper-medium (Radford et al., 2023)	6.03	6.99	7.03	1.00
Whisper-large-v3 (Radford et al., 2023)	6.04	6.93	7.03	0.99
SenseVoice-Small (An et al., 2024)	5.90	6.89	6.97	1.07
Qwen3-ASR-0.6B (Shi et al., 2026)	5.82	6.96	7.01	1.19
Qwen3-ASR-1.7B (Shi et al., 2026)	6.01	6.96	7.01	1.00

Measures of utterance length survive recognition error. The main result in Table 6 is that child mean length of utterance stays within half a character of the human value under all eight transcription conditions, and in every condition the transcript improves on a duration-only estimate (Appendix E.6). What survives is a developmental quantity, not only an average, since human transcripts show child utterances growing by one character from junior to senior classes, and Table 7 shows all seven fine-tuned systems across four model families reproducing that difference. Current

Table 8: Role timelines on the continuous test partition scored as diarization (DER), role-attributed transcription (mtCER, role-conditioned CER), and classroom measurement: teacher and child speaking-time share (%), teacher–child turns per minute, and child response rate to teacher utterances (%), as classroom means with relative bias to the human timeline in color.

			CER		Speaking-time share			
Role timeline	**DER**	**mtCER**	**Teacher**	**Child**	**Teacher**	**Child**	**Turns/min**	**Response**
Human annotation	0.00	12.98	4.86	20.95	50.0	14.5	3.45	21.2
Child–Adult Diar. (Xu et al., 2025a)	12.36	27.47	9.27	37.88	42.9 −14.3	19.9 +37.5	14.00 +305.7	60.4 +184.9
+ MPCS adaptation	2.33	16.80	5.37	24.34	50.0 −0.1	14.5 +0.1	4.64 +34.5	27.3 +28.6
BabyHuBERT-VTC (Charlot et al., 2025)	22.19	30.21	14.89	34.99	40.9 −18.2	19.5 +34.8	9.85 +185.4	49.3 +132.5
+ MPCS adaptation	1.41	14.89	5.05	22.25	50.0 −0.1	14.6 +0.6	4.05 +17.3	24.2 +14.1
WSW 2.0 (Sun et al., 2025)	15.97	23.28	7.40	28.09	52.3 +4.5	15.2 +4.9	4.90 +42.0	32.0 +50.8
ResNet (Kothalkar et al., 2024)	17.03	24.14	8.35	28.27	49.1 −1.8	16.7 +15.3	5.32 +54.3	34.0 +60.4

recognition therefore already supports comparisons of children's language production across age groups, even from transcripts too noisy to read.

Measures of which words children use do not. Every system in Table 6 overestimates it, and the overestimate grows with child error: the strongest system adds about 1.6 distinct words to every 50-word window of child speech, and Whisper-small about 5.4, while the teacher side is overestimated by a quarter to a half as much. Yet a lower child CER does not guarantee a better estimate: fine-tuning lowers SenseVoice-Small's child CER from 29.5% to 22.6% while its bias rises from +3.9 to +4.9 points. Errors on child speech are mostly substitutions, which leave an utterance the same length while usually introducing a word that is new to its window; where the error profile differs, utterance length moves with it: at almost the same child CER (42.1% and 40.8%), the insertions of Qwen3-Omni inflate child utterance length by 12.6% while Qwen2.5-Omni stays within 0.1%; the deletions of WenetSpeech-Conformer collapse it by 75%, and every substitution-dominated system stays within 4.3% (Table 12). Taken together, length-based measures are usable from transcripts whose errors are substitutions, while lexical measures wait on fewer child substitutions.

4.3 Which Measures Survive Segmentation Errors

We evaluate four role models on the continuous test partition: two released child–adult classifiers, Child–Adult Diarization (Xu et al., 2025a) and BabyHuBERT-VTC (Charlot et al., 2025), before and after adaptation on the continuous training partition, and two classroom-oriented architectures trained on the same partition, the wav2vec 2.0 role classifier of WSW 2.0 (Sun et al., 2025) and the room-independent ResNet of Kothalkar et al. (2024). Table 8 scores each timeline in three ways. For speaker-role diarization, DER is computed over fixed teacher and child labels. For role-attributed transcription, the zero-shot Qwen3-ASR-1.7B checkpoint transcribes the speech units of every timeline, so that rows differ only in the timeline, and mtCER charges substitutions, deletions, insertions, and misattributed characters to the reference role (Xu et al., 2026). For classroom measurement, teacher and child speaking-time shares, teacher–child turn frequency, and the rate at which a child utterance follows a teacher utterance are computed from the human timeline and from each system timeline. Turns alternate within 5 s and responses follow within 2.5 s (scoring rules in Appendix B.3, sensitivity in Appendix E.1). Adapted and trained models are averaged over three seeds.

Classroom adaptation brings role timelines close to human segmentation. As shown in Table 8, adapting the two released checkpoints on MPCS cuts diarization error from 22.2% to 1.4% for BabyHuBERT-VTC, close to the 1.2% inter-annotator DER. The transcripts follow: with the recognizer held fixed, the adapted BabyHuBERT timeline yields 14.9% role-attributed error against 13.0% on the human timeline, where the released checkpoint had given 30.2%. Taken together, classroom supervision makes role-attributed transcripts of continuous audio nearly as good as those built on human segmentation.

Accurate timelines give accurate talk time but inflate turn counts. The two adapted models place teacher and child speaking-time shares within 1% of the human timeline, yet report far more interaction: BabyHuBERT overcounts teacher–child turns by 17% and Child–Adult Diarization by 35%, consistently across seeds. By comparison, an independent human re-annotation of these classrooms differs from the reference in turn frequency by only 2.7% on average. DER ranks these timelines differently from classroom measurement: of the two released models, Child–Adult Diarization has

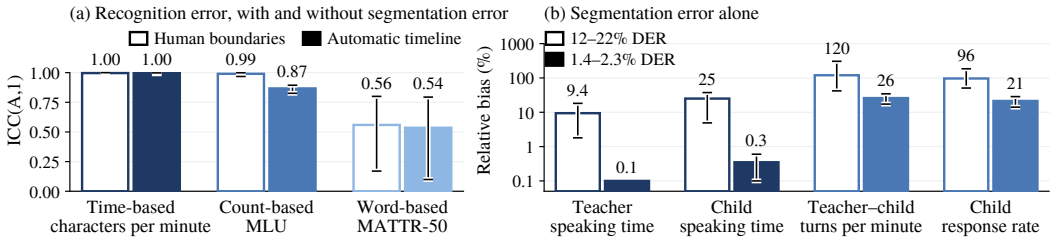

Figure 3: Which measures survive which errors. (a) Agreement of child measures with the human reference on human boundaries (hollow) and on the adapted BabyHuBERT timeline (solid): mean and range over the eight transcription conditions. (b) Relative bias of interaction measures for the two released and two trained role timelines at 12–22% DER (hollow) and the two adapted timelines at 1.4% and 2.3% DER (solid): median and range.

the lower DER (12.4% against 22.2%) yet the larger turn-frequency error (306% against 185%). After adaptation, the remaining errors are brief: each segment of one role inside the other's speech adds two turns and almost no time. The adapted BabyHuBERT timeline still tracks human turn frequencies across classrooms at a correlation of 0.99 (Appendix E.5), so interaction counts can be calibrated against the paired reference where their absolute values cannot be taken directly.

4.4 Error Type Decides Which Measures Survive

Recognition errors and segmentation errors damage different measures. Recomputing the measures of Section 4.2 on the adapted BabyHuBERT timeline, where boundaries and roles are automatic as well, separates the two kinds of error (Figure 3; Appendix E.4). Children's characters per minute are unaffected, agreeing with the human reference at an intraclass correlation between 0.990 and 0.998 across all eight transcription conditions, because fragmentation changes neither the characters spoken nor the time they occupy. Mean length of utterance falls from an agreement between 0.968 and 0.999 on human boundaries to between 0.826 and 0.894, because it divides those characters by a number of utterances that fragmentation inflates. This division follows one principle: an error rate weighs each error by the characters or seconds it occupies, while a measure registers it in its own unit. A substitution occupies one character but can add a word type, and a brief segment of the wrong role occupies a fraction of a second but adds two turns; CER therefore overstates what substitutions do to utterance length, and DER understates what brief segments do to turn frequency.

At a fixed error rate, error type alone reproduces the damage. Injected into the human annotation at 20% CER, substitutions leave children's MLU exact, while deletions and insertions shift it by 20% in opposite directions. At 2% DER, the same role confusion inflates turns by 17% as 2.6-second segments and by 146% as 0.3-second segments, and 1% DER in 0.3-second segments still inflates them by 74% (Appendix E.7). The adapted BabyHuBERT timeline matches the 2.6-second condition: 8.6 extra child units per classroom and 17% more turns in both.

Error types tell which measures a system supports. The principle gives a rule that no single error rate can: report substitution, deletion, and insertion rates by speaker role alongside CER and DER, take time-based measures from any low-DER system, calibrate count-based measures against a paired reference, and expect word-based measures to follow once child substitutions fall.

5 Conclusion

MPCS pairs verbatim transcripts with continuous speaker-role timelines over naturalistic Mandarin preschool classroom recordings, giving every automatic classroom measure a human counterpart. Against this reference, recognition errors concentrate on children and are not removed by in-domain supervision, role timelines at near-zero diarization error still inflate interaction counts, and the two kinds of error damage different measures: substitutions reach word identity, fragmented boundaries reach utterance counts, and measures defined on time survive both. Which classroom measures can be automated is therefore decided by the kind of error a system makes, not by how many; error types by speaker role make that decision visible.

AI USE STATEMENT

Generative AI tools were used to assist with literature retrieval, manuscript revision, and language polishing. The authors reviewed the suggested references and AI-assisted text, verified their relevance and accuracy, and revised the material as needed. The authors take full responsibility for the final manuscript.

ETHICS STATEMENT

MPCS contains speech from teachers and from children aged 3–5 recorded during ordinary classroom activities. The study was approved by the institutional review board of the authors' institution; the approval number is withheld for anonymous review. Before data collection, informed consent was obtained from all teachers and from the parents or legal guardians of all participating children, who were informed of the purpose of the study, the recording procedure, and the intended academic use of the data. Participation was voluntary and unpaid, and participants may withdraw at any time, in which case their recordings are removed from subsequent releases. Only audio was recorded. Preschools, classrooms, teachers, and children appear in the release solely under anonymized identifiers, and segments containing directly identifying information were removed from both the segmented corpus and the continuous subset. The continuous subset carries speech only from consented participants, with audio outside the annotated teacher and child speech replaced by ambient noise from the same recording. MPCS is distributed through controlled access for non-commercial research on classroom language, role-aware speech technology, and tools that support teachers' professional development. Prospective users sign a data use agreement that prohibits re-identifying any child, teacher, or preschool, keeps the audio and transcripts within the requesting research group, requires institutional ethics approval where applicable, and excludes covert monitoring, profiling, and high-stakes decisions about individual teachers or children; recipients delete any recordings that participants withdraw.

REPRODUCIBILITY STATEMENT

Appendix A describes how MPCS was collected and annotated, and Appendix B gives the settings of every evaluated system together with the scoring procedures. Code is available at this anonymous repository. The dataset will be released upon acceptance through controlled access under the data use agreement described in the Ethics Statement.

REFERENCES

Keyu An, Qian Chen, Chong Deng, et al. FunAudioLLM: Voice understanding and generation foundation models for natural interaction between humans and LLMs. Technical report, arXiv, 2024. URL https://arxiv.org/abs/2407.04051.

Anton Batliner, Mats Blomberg, Shona D'Arcy, Daniel Elenius, Diego Giuliani, Matteo Gerosa, Christian Hacker, Martin Russell, Stefan Steidl, and Michael Wong. The PF-STAR children's speech corpus. In *Interspeech 2005*, pp. 2761–2764. ISCA, 2005. doi: 10.21437/Interspeech.2005-705. URL https://doi.org/10.21437/Interspeech.2005-705.

Linda Bell, Johan Boye, Joakim Gustafson, Mattias Heldner, Anders Lindström, and Mats Wirén. The Swedish NICE corpus – spoken dialogues between children and embodied characters in a computer game scenario. In *Interspeech 2005*, pp. 2765–2768. ISCA, 2005. doi: 10.21437/Interspeech.2005-706. URL https://doi.org/10.21437/Interspeech.2005-706.

Ellen Irén Brinchmann, Hanne Røe-Indregård, Jannicke Karlsen, Stefan Kilian Schauber, and Bente Eriksen Hagtvet. The linguistic complexity of adult and child contextualized and decontextualized talk. *Child Development*, 94(5):1368–1380, 2023. doi: 10.1111/cdev.13932. URL https://doi.org/10.1111/cdev.13932.

Kin Chung Jacky Chan, Padraic Monaghan, and Marije Michel. Adapting to children's individual language proficiency: An observational study of preschool teacher talk addressing monolinguals and children learning English as an additional language. *Journal of Child Language*, 50(2):

365–390, 2023. doi: 10.1017/S0305000921000854. URL https://doi.org/10.1017/S0305000921000854.

Théo Charlot, Tarek Kunze, Maxime Poli, Alejandrina Cristia, Emmanuel Dupoux, and Marvin Lavechin. BabyHuBERT: Multilingual self-supervised learning for segmenting speakers in child-centered long-form recordings. Technical report, arXiv, 2025. URL https://arxiv.org/abs/2509.15001.

Nancy F. Chen, Rong Tong, Darren Wee, Peixuan Lee, Bin Ma, and Haizhou Li. SingaKids-Mandarin: Speech corpus of Singaporean children speaking Mandarin Chinese. In *Interspeech 2016*, pp. 1545–1549. ISCA, 2016. doi: 10.21437/Interspeech.2016-139. URL https://doi.org/10.21437/Interspeech.2016-139.

Sylvana M. Côté, Chantal Mongeau, Christa Japel, Qian Xu, Jean R. Séguin, and Richard E. Tremblay. Child care quality and cognitive development: Trajectories leading to better preacademic skills. *Child Development*, 84(2):752–766, 2013. doi: 10.1111/cdev.12007. URL https://doi.org/10.1111/cdev.12007.

Alejandrina Cristia, Marvin Lavechin, Camila Scaff, Melanie Soderstrom, Caroline Rowland, Okko Räsänen, John Bunce, and Elika Bergelson. A thorough evaluation of the Language Environment Analysis (LENA) system. *Behavior Research Methods*, 53(2):467–486, 2021. doi: 10.3758/s13428-020-01393-5. URL https://doi.org/10.3758/s13428-020-01393-5.

Meghan E. Dale, Amanda J. Godley, Sarah A. Capello, Patrick J. Donnelly, Sidney K. D'Mello, and Sean P. Kelly. Toward the automated analysis of teacher talk in secondary ELA classrooms. *Teaching and Teacher Education*, 110:103584, 2022. doi: 10.1016/j.tate.2021.103584. URL https://doi.org/10.1016/j.tate.2021.103584.

Xiangjun Deng and Virginia Yip. A multimedia corpus of child Mandarin: The Tong corpus. *Journal of Chinese Linguistics*, 46(1):69–92, 2018. doi: 10.1353/jcl.2018.0002. URL https://doi.org/10.1353/jcl.2018.0002.

David K. Dickinson and Michelle V. Porche. Relation between language experiences in preschool classrooms and children's kindergarten and fourth-grade language and reading abilities. *Child Development*, 82(3):870–886, 2011. doi: 10.1111/j.1467-8624.2011.01576.x. URL https://doi.org/10.1111/j.1467-8624.2011.01576.x.

Satwik Dutta, Dwight Irvin, Jay Buzhardt, and John H. L. Hansen. Activity focused speech recognition of preschool children in early childhood classrooms. In *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022)*, pp. 92–100, Seattle, Washington, 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.bea-1.13. URL https://aclanthology.org/2022.bea-1.13/.

Jun Gao, Aijun Li, and Ziyu Xiong. Mandarin multimedia child speech corpus: CASS_Child. In *2012 International Conference on Speech Database and Assessments*, pp. 7–12. IEEE, 2012. doi: 10.1109/ICSDA.2012.6422462. URL https://doi.org/10.1109/ICSDA.2012.6422462.

Marta Garrote Salazar and José María Guirao Miras. CHIEDE, corpus de habla infantil espontánea del español. *Procesamiento del Lenguaje Natural*, 41:297–298, 2008. URL http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/2580.

Jill Gilkerson, Jeffrey A. Richards, Steven F. Warren, D. Kimbrough Oller, Rosemary Russo, and Betty Vohr. Language experience in the second year of life and language outcomes in late childhood. *Pediatrics*, 142(4):e20174276, 2018. doi: 10.1542/peds.2017-4276. URL https://doi.org/10.1542/peds.2017-4276.

Anmol Gulati, James Qin, Chung-Cheng Chiu, et al. Conformer: Convolution-augmented transformer for speech recognition. In *Interspeech 2020*, pp. 5036–5040. ISCA, 2020. doi: 10.21437/Interspeech.2020-3015. URL https://doi.org/10.21437/Interspeech.2020-3015.

John H. L. Hansen, Satwik Dutta, and Ellen Grand. Best practices and considerations for child speech corpus collection and curation in educational, clinical, and forensic scenarios. In *10th Workshop on Speech and Language Technology in Education (SLaTE 2025)*, pp. 123–127. ISCA, 2025. doi: 10.21437/SLaTE.2025-25. URL https://doi.org/10.21437/SLaTE.2025-25.

Steven J. Howard, Iram Siraj, Edward C. Melhuish, Denise Kingston, Cathrine Neilsen-Hewett, Marc de Rosnay, Elisabeth Duursma, and Betty Luu. Measuring interactional quality in preschool settings: Introduction and validation of the Sustained Shared Thinking and Emotional Wellbeing (SSTEW) scale. *Early Child Development and Care*, 190(7):1017–1030, 2020. doi: 10.1080/03004430.2018.1511549. URL https://doi.org/10.1080/03004430.2018.1511549.

Manasa Kalanadhabhatta, Mohammad Mehdi Rastikerdar, Tauhidur Rahman, Adam S. Grabell, and Deepak Ganesan. Playlogue: Dataset and benchmarks for analyzing adult–child conversations during play. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 8(4):173:1–173:34, 2024. doi: 10.1145/3699775. URL https://doi.org/10.1145/3699775.

KimiTeam, Ding Ding, Zeqian Ju, et al. Kimi-Audio technical report. Technical report, arXiv, 2025. URL https://arxiv.org/abs/2504.18425.

Tom Ko, Vijayaditya Peddinti, Daniel Povey, and Sanjeev Khudanpur. Audio augmentation for speech recognition. In *Interspeech 2015*, pp. 3586–3589. ISCA, 2015. doi: 10.21437/Interspeech.2015-711. URL https://doi.org/10.21437/Interspeech.2015-711.

Prasanna V. Kothalkar, Dwight Irvin, Ying Luo, Joanne Rojas, John Nash, Beth Rous, and John H. L. Hansen. Tagging child–adult interactions in naturalistic, noisy, daylong school environments using i-vector based diarization system. In *8th ISCA Workshop on Speech and Language Technology in Education (SLaTE 2019)*, pp. 89–93. ISCA, 2019. doi: 10.21437/SLaTE.2019-17. URL https://doi.org/10.21437/SLaTE.2019-17.

Prasanna V. Kothalkar, John H. L. Hansen, Dwight Irvin, and Jay Buzhardt. Child–adult speech diarization in naturalistic conditions of preschool classrooms using room-independent ResNet model and automatic speech recognition-based re-segmentation. *The Journal of the Acoustical Society of America*, 155(2):1198–1215, 2024. doi: 10.1121/10.0024353. URL https://doi.org/10.1121/10.0024353.

Xingming Li, Runke Huang, Yanan Bao, Yuye Jin, Yuru Jiao, and Qingyong Hu. When AI meets early childhood education: Large language models as assessment teammates in Chinese preschools. In *Proceedings of the 27th International Conference on Artificial Intelligence in Education (AIED)*, Lecture Notes in Computer Science, pp. 263–278, Seoul, South Korea, 2026. Springer. doi: 10.1007/978-3-032-29744-0_18. URL https://doi.org/10.1007/978-3-032-29744-0_18.

Rasa Lileikyte, Dwight Irvin, and John H. L. Hansen. Assessing child communication engagement and statistical speech patterns for American English via speech recognition in naturalistic active learning spaces. *Speech Communication*, 140:98–108, 2022. doi: 10.1016/j.specom.2022.01.006. URL https://doi.org/10.1016/j.specom.2022.01.006.

Samantha G. Mitsven, Lynn K. Perry, Yudong Tao, Batya E. Elbaum, Neil F. Johnson, and Daniel S. Messinger. Objectively measured teacher and preschooler vocalizations: Phonemic diversity is associated with language abilities. *Developmental Science*, 25(2):e13177, 2022. doi: 10.1111/desc.13177. URL https://doi.org/10.1111/desc.13177.

Daniel S. Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D. Cubuk, and Quoc V. Le. SpecAugment: A simple data augmentation method for automatic speech recognition. In *Interspeech 2019*, pp. 2613–2617. ISCA, 2019. doi: 10.21437/Interspeech.2019-2680. URL https://doi.org/10.21437/Interspeech.2019-2680.

G. Logan Pelfrey, Laura M. Justice, Hugo Gonzalez Villasanti, and Tiffany J. Foster. Evaluating the accuracy of automated processing of child and adult language production in preschool classrooms.

Frontiers in Psychology, 15:1322665, 2024. doi: 10.3389/fpsyg.2024.1322665. URL https://doi.org/10.3389/fpsyg.2024.1322665.

Yifan Peng, Siddharth Dalmia, Ian Lane, and Shinji Watanabe. Branchformer: Parallel MLP-attention architectures to capture local and global context for speech recognition and understanding. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 17627–17643. PMLR, 2022. URL https://proceedings.mlr.press/v162/peng22a.html.

Robert C. Pianta, Karen M. La Paro, and Bridget K. Hamre. *Classroom Assessment Scoring System (CLASS) Manual, Pre-K*. Paul H. Brookes Publishing Co., Baltimore, MD, 2008. ISBN 978-1-55766-941-4.

Sameer S. Pradhan, Ronald A. Cole, and Wayne 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)*, pp. 12040–12045, Torino, Italia, 2024. ELRA and ICCL. URL https://aclanthology.org/2024.lrec-main.1052/.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 28492–28518. PMLR, 2023. URL https://proceedings.mlr.press/v202/radford23a.html.

Xian Shi, Xiong Wang, Zhifang Guo, et al. Qwen3-ASR technical report. Technical report, arXiv, 2026. URL https://arxiv.org/abs/2601.21337.

Khaldoun Shobaki, John-Paul Hosom, and Ronald A. Cole. The OGI kids' speech corpus and recognizers. In *6th International Conference on Spoken Language Processing (ICSLP 2000)*, volume 4, pp. 258–261. ISCA, 2000. doi: 10.21437/ICSLP.2000-800. URL https://doi.org/10.21437/ICSLP.2000-800.

Melanie Soderstrom, Marisa Casillas, Elika Bergelson, Celia Rosemberg, Florencia Alam, Anne S. Warlaumont, and John Bunce. Developing a cross-cultural annotation system and MetaCorpus for studying infants' real world language experience. *Collabra: Psychology*, 7(1):23445, 2021. doi: 10.1525/collabra.23445. URL https://doi.org/10.1525/collabra.23445.

Anchen Sun, Juan J. Londono, Batya Elbaum, Luis Estrada, Roberto Jose Lazo, Laura Vitale, Hugo Gonzalez Villasanti, Riccardo Fusaroli, Lynn K. Perry, and Daniel S. Messinger. Who said what? An automated approach to analyzing speech in preschool classrooms. In *2024 IEEE International Conference on Development and Learning (ICDL)*, pp. 1–8. IEEE, 2024. doi: 10.1109/ICDL61372.2024.10644508. URL https://doi.org/10.1109/ICDL61372.2024.10644508.

Anchen Sun, Tiantian Feng, Gabriela Gutierrez, Juan J. Londono, Anfeng Xu, Batya Elbaum, Shrikanth Narayanan, Lynn K. Perry, and Daniel S. Messinger. Who said what (WSW 2.0)? Enhanced automated analysis of preschool classroom speech. In *2025 IEEE International Conference on Development and Learning (ICDL)*, pp. 1–6. IEEE, 2025. doi: 10.1109/ICDL63968.2025.11204438. URL https://doi.org/10.1109/ICDL63968.2025.11204438.

Abhijit Suresh, Jennifer Jacobs, Charis Harty, Margaret Perkoff, James H. Martin, and Tamara Sumner. The TalkMoves dataset: K-12 mathematics lesson transcripts annotated for teacher and student discursive moves. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pp. 4654–4662, Marseille, France, 2022. European Language Resources Association. URL https://aclanthology.org/2022.lrec-1.497/.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, pp. 5998–6008. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

Ye-Yi Wang, Alex Acero, and Ciprian Chelba. Is word error rate a good indicator for spoken language understanding accuracy. In *2003 IEEE Workshop on Automatic Speech Recognition and Understanding*, pp. 577–582. IEEE, 2003. doi: 10.1109/ASRU.2003.1318504. URL https://doi.org/10.1109/ASRU.2003.1318504.

Zuowei Wang, Xingyu Pan, Kevin F. Miller, and Kai S. Cortina. Automatic classification of activities in classroom discourse. *Computers & Education*, 78:115–123, 2014. doi: 10.1016/j.compedu.2014.05.010. URL https://doi.org/10.1016/j.compedu.2014.05.010.

Shinji Watanabe, Takaaki Hori, Suyoun Kim, John R. Hershey, and Tomoki Hayashi. Hybrid CTC/attention architecture for end-to-end speech recognition. *IEEE Journal of Selected Topics in Signal Processing*, 11(8):1240–1253, 2017. doi: 10.1109/JSTSP.2017.2763455. URL https://doi.org/10.1109/JSTSP.2017.2763455.

Shinji Watanabe, Michael Mandel, Jon Barker, et al. CHiME-6 challenge: Tackling multispeaker speech recognition for unsegmented recordings. In *6th International Workshop on Speech Processing in Everyday Environments (CHiME 2020)*, pp. 1–7. ISCA, 2020. doi: 10.21437/CHiME.2020-1. URL https://doi.org/10.21437/CHiME.2020-1.

Adriana Weisleder and Anne Fernald. Talking to children matters: Early language experience strengthens processing and builds vocabulary. *Psychological Science*, 24(11):2143–2152, 2013. doi: 10.1177/0956797613488145. URL https://doi.org/10.1177/0956797613488145.

Jennifer Whorrall and Sonia Q. Cabell. Supporting children's oral language development in the preschool classroom. *Early Childhood Education Journal*, 44(4):335–341, 2016. doi: 10.1007/s10643-015-0719-0. URL https://doi.org/10.1007/s10643-015-0719-0.

Boyong Wu, Chao Yan, Chen Hu, et al. Step-Audio 2 technical report. Technical report, arXiv, 2025. URL https://arxiv.org/abs/2507.16632.

Anfeng Xu, Kevin Huang, Tiantian Feng, Lue Shen, Helen Tager-Flusberg, and Shrikanth Narayanan. Exploring speech foundation models for speaker diarization in child-adult dyadic interactions. In *Interspeech 2024*, pp. 5193–5197. ISCA, 2024. doi: 10.21437/Interspeech.2024-717. URL https://doi.org/10.21437/Interspeech.2024-717.

Anfeng Xu, Tiantian Feng, Helen Tager-Flusberg, Catherine Lord, and Shrikanth Narayanan. Data efficient child-adult speaker diarization with simulated conversations. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2025a. doi: 10.1109/ICASSP49660.2025.10889307. URL https://doi.org/10.1109/ICASSP49660.2025.10889307.

Anfeng Xu, Tiantian Feng, Somer Bishop, Catherine Lord, and Shrikanth Narayanan. End-to-end joint ASR and speaker role diarization with child–adult interactions. *IEEE Transactions on Audio, Speech and Language Processing*, 34:3532–3544, 2026. doi: 10.1109/TASLPRO.2026.3707421. URL https://doi.org/10.1109/TASLPRO.2026.3707421.

Jin Xu, Zhifang Guo, Jinzheng He, et al. Qwen2.5-Omni technical report. Technical report, arXiv, 2025b. URL https://arxiv.org/abs/2503.20215.

Jin Xu, Zhifang Guo, Hangrui Hu, et al. Qwen3-Omni technical report. Technical report, arXiv, 2025c. URL https://arxiv.org/abs/2509.17765.

Ning Yang, Jiuqian Shi, Jinjin Lu, and Yi Huang. Language development in early childhood: Quality of teacher-child interaction and children's receptive vocabulary competency. *Frontiers in Psychology*, 12:649680, 2021. doi: 10.3389/fpsyg.2021.649680. URL https://doi.org/10.3389/fpsyg.2021.649680.

Zhuoyuan Yao, Di Wu, Xiong Wang, Binbin Zhang, Fan Yu, Chao Yang, Zhendong Peng, Xiaoyu Chen, Lei Xie, and Xin Lei. WeNet: Production oriented streaming and non-streaming end-to-end speech recognition toolkit. In *Interspeech 2021*, pp. 4054–4058. ISCA, 2021. doi: 10.21437/Interspeech.2021-1983. URL https://doi.org/10.21437/Interspeech.2021-1983.

Donghua Yu, Zhengyuan Lin, Hanfu Chen, et al. MOSS Transcribe Diarize technical report. Technical report, arXiv, 2026. URL https://arxiv.org/abs/2601.01554.

Fan Yu, Zhuoyuan Yao, Xiong Wang, Keyu An, Lei Xie, Zhijian Ou, Bo Liu, Xiulin Li, and Guanqiong Miao. The SLT 2021 children speech recognition challenge: Open datasets, rules and baselines. In *2021 IEEE Spoken Language Technology Workshop (SLT)*, pp. 1117–1123. IEEE, 2021. doi: 10.1109/SLT48900.2021.9383608. URL https://doi.org/10.1109/SLT48900.2021.9383608.

Fan Yu, Shiliang Zhang, Yihui Fu, et al. M2MeT: The ICASSP 2022 multi-channel multi-party meeting transcription challenge. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6167–6171. IEEE, 2022. doi: 10.1109/ICASSP43922.2022.9746465. URL https://doi.org/10.1109/ICASSP43922.2022.9746465.

Binbin Zhang, Hang Lv, Pengcheng Guo, et al. WenetSpeech: A 10000+ hours multi-domain Mandarin corpus for speech recognition. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6182–6186. IEEE, 2022. doi: 10.1109/ICASSP43922.2022.9746682. URL https://doi.org/10.1109/ICASSP43922.2022.9746682.

Jiaming Zhou, Shiyao Wang, Shiwan Zhao, et al. ChildMandarin: A comprehensive Mandarin speech dataset for young children aged 3–5. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12524–12537, Vienna, Austria, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.acl-long.614. URL https://aclanthology.org/2025.acl-long.614/.

Jiaming Zhou, Yujie Guo, Shiwan Zhao, Yao Lu, Jianye Wang, Haoqin Sun, Hui Wang, and Yong Qin. ChildTalk: A multi-dialect Chinese child speech corpus with full-length child–caregiver conversations for speech recognition. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 5103–5116, San Diego, California, United States, 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.findings-acl.251. URL https://aclanthology.org/2026.findings-acl.251/.

APPENDIX

A Dataset Construction **16**

A.1 Recording and Segmentation 16

A.2 Annotation and Filtering Guidelines 17

A.3 Continuous Subset and Partitions 18

B Experimental Setup **18**

B.1 ASR Training 18

B.2 Role Models and Continuous Transcription 19

B.3 Metrics and Scoring 19

C Speech Recognition: Additional Results **20**

C.1 Decoding Strategies for From-Scratch Models 20

C.2 Signal-to-Noise Ratio and the Teacher–Child Gap 21

C.3 Error Composition of the Teacher–Child Gap 21

C.4 Training Data Scaling 22

C.5 Recognition by Class Band 23

D Domain Transfer: Additional Results **23**

D.1 Cross-Region Transfer 23

D.2 Cross-Setting Transfer with Full Training Sets 24

E Classroom Measurement: Additional Results **24**

E.1 Threshold Sensitivity of the Interaction Measures 24

E.2 Confidence Intervals for Language Measures 24

E.3 Vocabulary Diversity at a Wider Window 25

E.4 Measures on the Automatic Timeline 25

E.5 Agreement of the Interaction Measures 26

E.6 Transcription Against a Duration-Only Estimate 27

E.7 Controlled Perturbations at a Fixed Error Rate 27

A DATASET CONSTRUCTION

A.1 RECORDING AND SEGMENTATION

MPCS was constructed from continuous naturalistic classroom recordings collected in Shenzhen (Guangdong) and Guizhou, China. The 201 original recordings were collected from 125 preschools and 130 classrooms, totaling 569.97 hours; classrooms are identified by region, preschool, and class name.

For each recording session, a researcher delivered a recorder to the head teacher of the target classroom before the morning class began. The recorder was worn by the teacher throughout the session and continuously recorded natural classroom activities for approximately 3.5 hours, typically covering the entire morning schedule from the start of class to the lunch period. Multiple recordings

could be collected from the same classroom across different sessions. All recordings were collected using the same recording setup and were stored as mono, 16-bit, 16-kHz WAV audio.

MOSS-Transcribe-Diarize (Yu et al., 2026) segmented the continuous recordings into candidate utterances at semantically complete boundaries and drafted a transcript for each, and 335,133 candidate segments entered the manual annotation pipeline. Annotation excluded 59,267 segments (17.7%) under the criteria of Appendix A.2, retaining 275,866 segments (82.3%). The final corpus contains 275,866 speech segments totaling 227.25 hours, including 222,623 teacher segments (187.59 hours) and 53,243 child segments (39.66 hours). Table 2 reports the regional composition of the retained speech, and Table 3 reports its partition composition by role.

A.2 Annotation and Filtering Guidelines

The dataset was annotated by 22 undergraduate and graduate students with backgrounds in early childhood education in Shenzhen, who received unified training. As shown in Figure 4, all candidate segments were uploaded to an annotation platform, where annotators assign the speaker role, verify the draft transcript against the audio and correct it into a verbatim transcript, log every correction with its error type, and add auxiliary labels for validity and exclusion. The annotation follows the guidelines below:

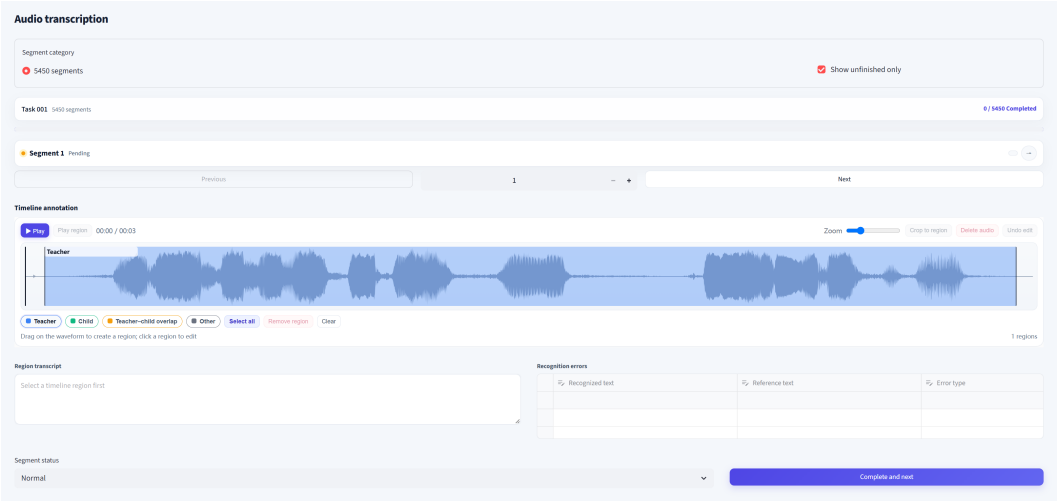

Figure 4: Annotation platform used for candidate-segment annotation.

- **Transcription.** Transcripts represent the intelligible Mandarin speech produced by teachers and children. Natural disfluencies, including repetitions, filled pauses, and self-repairs, are preserved when intelligible. Spoken numbers and other naturally occurring spoken forms are retained in character form.
- **Unintelligible speech.** Unintelligible portions within otherwise transcribable utterances are marked with a designated symbol. When a word or phrase is only partially produced and its intended form cannot be determined reliably, that portion is marked as unintelligible and excluded from the final corpus.
- **Speaker roles.** Roles are assigned as `teacher` or `child` according to the active speaker. When multiple teachers or children speak within a classroom, all speakers of the same category share the corresponding role label.
- **Non-speech events.** Segments consisting entirely of non-speech vocalizations or environmental sounds, including laughter, coughing, throat clearing, crying, music, and recorded media, are excluded from the speech recognition corpus. When such events occur within an otherwise intelligible segment, the spoken material is retained if it can still be reliably transcribed, and the event is not treated as lexical content.
- **Exclusion.** Privacy-sensitive information is time-marked and the segment containing it is withheld from the released corpus, following practice for child speech collections (Hansen et al., 2025).

Segments that cannot be transcribed or attributed to a role reliably, including dialectal and media speech, are labeled and omitted from the speech recognition corpus.

A.3 Continuous Subset and Partitions

Turn frequency and response rate are defined over uninterrupted time, so an excerpt has to be long enough that consecutive teacher and child speech falls inside it rather than across its edges. Long-form protocols that sample many short clips from each recording are built for estimating how much speech a day contains (Soderstrom et al., 2021), which is a different quantity. We therefore annotate one 30-minute excerpt per classroom, the window with the most teacher and child speech, and include the 83 classrooms whose window holds enough speech for these measures. This yields a 41.5-hour continuous subset for speaker-role diarization and role-attributed transcription. The subset is partitioned by preschool into 57 training, 10 development, and 16 test classrooms from 53, 10, and 12 preschools (28.5, 5.0, and 8.0 hours). Each continuous classroom is mapped to the segmented partition of its source recording, and preschools are assigned as a whole, with the test partition taking precedence over development and development over training. The test partition contains 10 Shenzhen and 6 Guizhou classrooms. The timeline extends the segmented annotation of the same audio, and the two views agree by construction: the annotators of Appendix A.2 begin from the transcripts and roles of the segmented utterances in each excerpt and label all remaining intervals on the same platform. Speaker-role annotations are defined at the role level: multiple speakers within the same role are merged into a single role track, and every interval is labeled teacher, child, or non-target, where non-target intervals contain neither teacher nor child speech; simultaneous teacher and child speech is excluded. In the continuous subset, teacher and child speech span 21.46 and 5.19 hours, respectively, and non-target intervals occupy 14.86 hours.

Figure 5 shows the continuous view of the platform, where annotators mark role-labeled regions on the waveform of each 30-minute excerpt and enter their transcripts.

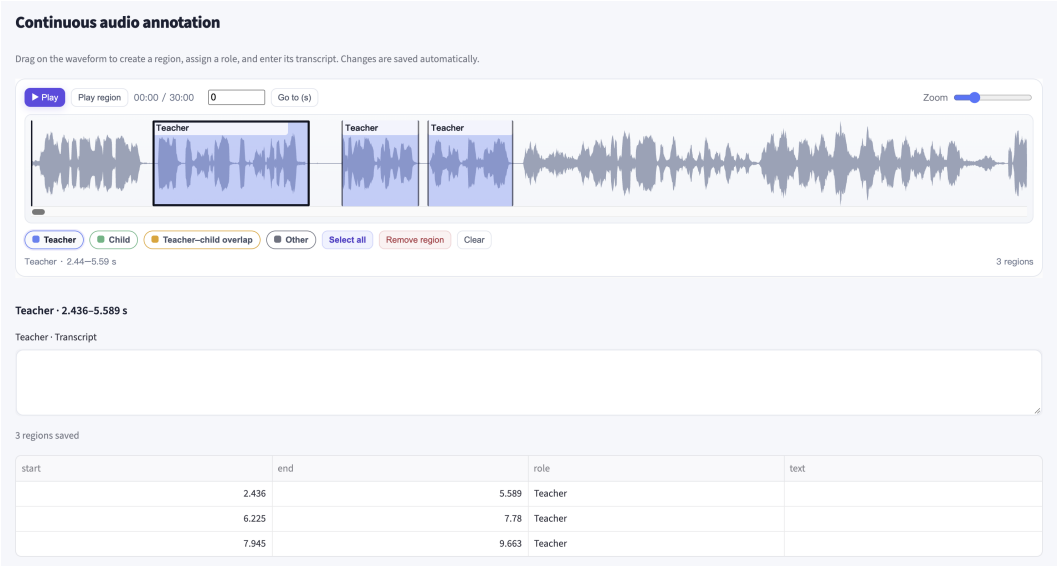

Figure 5: Annotation platform used for continuous timeline annotation.

B Experimental Setup

B.1 ASR Training

Transformer, Conformer-31M, and Branchformer are trained in WeNet (Yao et al., 2021) with 80-dimensional filter-bank features, speed perturbation (Ko et al., 2015), and SpecAugment (Park et al., 2019). All three models use joint CTC/attention training (Watanabe et al., 2017), the Adam optimizer, and the WarmupLR schedule, with seed 777 and no gradient accumulation. Table 9 summarizes their training hyperparameters; decoding uses attention rescoring (Appendix C.1).

Table 9: Training hyperparameters of the from-scratch ASR models. Batch size is per GPU; learning rates are peak values.

Model	Params	Decoder	Batch	LR	Warmup steps	Epochs
Transformer (Vaswani et al., 2017)	29.57 M	CTC + ATT	32	1×10^{-3}	59,670	50
Conformer-31M (Gulati et al., 2020)	31.72 M	CTC + ATT	32	1×10^{-3}	59,670	50
Branchformer (Peng et al., 2022)	29.63 M	CTC + ATT	16	1×10^{-3}	119,340	50

Pretrained ASR models are fine-tuned on MPCS with all parameters updated, retaining their native tokenizers and training objectives. Table 10 summarizes the full-data fine-tuning hyperparameters. Large audio-language models are evaluated zero-shot with a fixed Mandarin transcription instruction.

Table 10: Fine-tuning hyperparameters of the pretrained ASR models.

Model	Params	Batch size	Learning rate	Epochs
WenetSpeech-Conformer (Zhang et al., 2022)	135.15M	16	4×10^{-5}	20
Whisper-small (Radford et al., 2023)	∼244M	16	1×10^{-5}	10
Whisper-medium (Radford et al., 2023)	∼769M	16	1×10^{-5}	10
Whisper-large-v3 (Radford et al., 2023)	∼1.55B	8	1×10^{-5}	10
SenseVoice-Small (An et al., 2024)	∼234M	12,000 tokens	4×10^{-5}	10
Qwen3-ASR-0.6B (Shi et al., 2026)	782.43M	16	2×10^{-5}	3
Qwen3-ASR-1.7B (Shi et al., 2026)	2.04B	4	2×10^{-5}	3

Batch size is the micro-batch per GPU; SenseVoice uses a dynamic token budget. Learning rates are peak values. Params reports total model parameters; $\sim$ denotes nominal sizes. Qwen model names retain their released size labels.

The data-scaling study follows these model-specific recipes on shared nested subsets, with effective batch sizes of 32 utterances for Whisper and 128 for Qwen3-ASR.

B.2 Role Models and Continuous Transcription

The speaker-role diarization comparison includes pretrained Child–Adult Diarization and BabyHuBERT-VTC models, their adapted counterparts, and the classroom-oriented wav2vec 2.0 and ResNet models. Models trained or adapted on MPCS use the continuous training partition: training windows of 10 s (Child–Adult Diarization), 4 s (BabyHuBERT-VTC), and 1 s (wav2vec 2.0 and ResNet) are cut from the continuous audio and labeled teacher, child, or other from the human timelines. Child–Adult Diarization, BabyHuBERT-VTC, and the wav2vec 2.0 classifier start from their released weights, and ResNet is trained from random initialization. Each model is trained with seeds 777, 778, and 779, and its checkpoint, speech threshold, and post-processing parameters are selected on the continuous development partition. Scoring conventions follow Section 4.3 and Appendix B.3.

For continuous transcription, the zero-shot Qwen3-ASR-1.7B checkpoint, without MPCS fine-tuning, transcribes the role timelines produced by each system. Each timeline is organized into speech units: adjacent intervals of the same role separated by less than 0.2 s are merged, units shorter than 0.1 s are discarded, and units are decoded in blocks of at most 25 s, with the recognized text assigned to its unit. The oracle condition applies the same recognizer to the reference timeline, allowing the comparison to isolate the contribution of role attribution to transcription quality.

B.3 Metrics and Scoring

Character-level scoring. Reference and system transcripts pass through the same normalization: Unicode NFKC normalization, conversion of traditional to simplified characters, uppercasing of Latin letters, and removal of characters in Unicode categories P, Z, and C. Filled pauses, repetitions, and self-repairs are kept on both sides and scored verbatim. Character edit counts are computed per utterance, resolving ties in substitution, deletion, and insertion order, and summed within each evaluation group. CER is $100(S + D + I)/N$, where N is the reference character count. Each error type's contribution to the teacher–child gap is its child-minus-teacher rate difference divided by the total CER gap.

Role-attributed transcription. Reference and hypothesis characters are aligned jointly with their role labels. For role s, $\mathrm{mtCER}_s = (S_s + D_s + I_s + A_s)/N_s$, where S_s, D_s, and I_s are the

substitutions, deletions, and insertions of role s, A_s is the number of reference characters of role s attributed to the other role, and N_s is the reference character count of role s; mtCER is the mean of the teacher and child values, following mtWER (Xu et al., 2026). Role-conditioned CER omits the attribution term, $\mathrm{CER}_s = (S_s + D_s + I_s)/N_s$. Roles are fixed, so no speaker permutation is searched, and a misattributed character is charged once to its reference role rather than as a deletion and an insertion across concatenated role streams as in cpWER (Watanabe et al., 2020). These scores are computed on the continuous test partition with its own reference intervals.

Speaker-role diarization. Each time interval of a role timeline is labeled `teacher`, `child`, or `non-target`; roles rather than individual speakers are distinguished, so multiple children form a single role track. DER is $(T_{\mathrm{miss}} + T_{\mathrm{FA}} + T_{\mathrm{conf}})/T_{\mathrm{ref}}$ (Xu et al., 2025a), where the numerator sums missed-speech, false-alarm, and role-confusion durations and T_{ref} is the total reference duration of teacher and child speech. Fixed role labels are scored without permutation, using a 100-ms collar (50 ms on each side of a boundary). False alarms are target roles predicted during non-target intervals, missed speech is non-target predicted during target-role speech, and role confusion is teacher and child interchanged; these components separate speech localization errors from role-assignment errors.

Classroom measures. For each classroom, the same measure function is applied to the human reference and to a system output. Speaking-time shares divide teacher or child speech duration by the evaluated recording duration. Turn frequency counts changes between teacher and child speech separated by at most 5 s, per minute, and the teacher-to-child response rate is the fraction of teacher utterances followed by the onset of a child utterance within 2.5 s. Adjacent same-role intervals separated by less than 0.2 s are merged and units shorter than 0.1 s are discarded, identically for reference and system timelines. Text measures use the utterance units of the human timeline (Section 4.2) or of each system timeline (Section 4.3): MLU is characters per unit, characters per minute divides the character count by the evaluated minutes, and MATTR-50 tokenizes each unit with jieba (0.42.1, precise mode with HMM), concatenates same-role tokens in time order, and averages the type–token ratio over sliding 50-word windows. Classrooms are the unit of analysis; bias is the mean of system minus human, MAE the mean absolute difference, and agreement the absolute-agreement intraclass correlation ICC(A,1). Confidence intervals use 2,000 bootstrap replicates that resample preschools and keep their classrooms.

C Speech Recognition: Additional Results

C.1 Decoding Strategies for From-Scratch Models

Table 11 compares four decoding strategies for the three from-scratch models; Table 4 reports attention rescoring.

Table 11: Results of from-scratch ASR models on the MPCS test partition: overall CER (%) under four decoding strategies and role-conditioned CER (%) with attention rescoring. Gap: child minus teacher CER.

		Overall CER by decoding				Attention rescoring		
System	Params	CTC greedy	CTC prefix	Attention	Att. rescoring	Teacher	Child	Gap
Transformer (Vaswani et al., 2017)	29.57 M	22.36	22.31	22.30	20.64	15.75	45.81	30.06
Conformer-31M (Gulati et al., 2020)	31.72 M	21.07	21.07	21.12	20.01	15.05	45.53	30.48
Branchformer (Peng et al., 2022)	29.63 M	21.56	21.55	20.67	20.12	15.22	45.40	30.18

C.2 Signal-to-Noise Ratio and the Teacher–Child Gap

We estimate the signal-to-noise ratio of every test utterance with WADA-SNR, a blind estimator that is independent of absolute recording level, and pool teacher and child utterances into SNR quintiles. Child speech reaches the teacher-worn microphone at a lower level and SNR than teacher speech: the median energy-gated speech level is 6.46 dB lower and the median SNR 5 dB lower (21 dB versus 26 dB), with a distribution overlap of 0.66 (Figure 6, top).

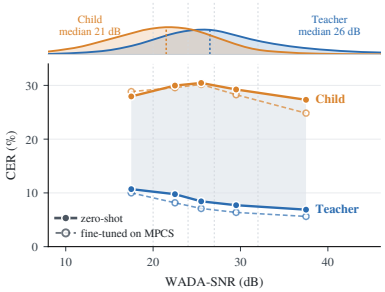

Figure 6: The teacher–child CER gap persists in every SNR quintile (Qwen3-ASR-1.7B). Top: SNR distributions with medians. Bottom: CER by quintile; shading marks the zero-shot gap.

SNR governs teacher errors and leaves child errors unchanged. Teacher CER falls from 10.70% in the lowest quintile to 6.88% in the highest, whereas child CER remains between 27.32% and 30.46% in every quintile; fine-tuning on MPCS lowers both curves without changing this pattern (Figure 6, bottom). Reweighting child utterances to the teacher SNR distribution changes the gap by 0.02 points (95% confidence interval −1.79 to 1.62 over 2,000 preschool-clustered bootstrap replicates), and a character-weighted regression with recording fixed effects, linear and quadratic SNR terms, and log duration retains a child effect of 18.85 points (16.80 to 21.31). The teacher–child gap therefore persists at matched signal-to-noise ratio, in line with the substitution-dominated error profile in Appendix C.3.

C.3 Error Composition of the Teacher–Child Gap

Figure 7 shows the decomposition for representative systems, and Table 12 extends it across model families. Substitution shares span 70.2–77.7% for nine of the ten dedicated ASR systems, while insertion shares span 13.8–40.2% across LALMs. The tenth, WenetSpeech-Conformer, is the one system in a different regime: deletions account for more than the whole gap and substitutions narrow it, because its zero-shot output drops most child speech outright, at a child deletion rate of 74.9% over 4,384 empty predictions. The last column shows what each regime does to a classroom measure. The nine substitution-dominated systems shift child MLU by at most 4.3%, Kimi-Audio and Qwen3-Omni inflate it by 10.5% and 12.6%, and the deletion-dominated system collapses it by 75.2%.

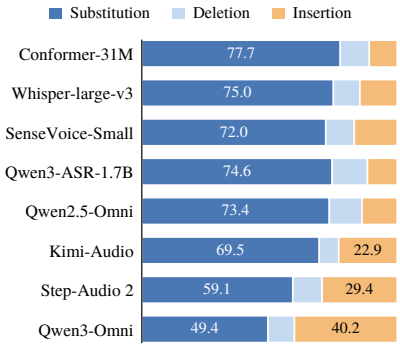

Figure 7: Percentage of the teacher–child CER gap contributed by substitutions, deletions, and insertions for representative systems.

Table 13 reports the duration × role × error-type decomposition for Qwen3-ASR-1.7B (zero-shot). Substitutions contribute 66.8–76.5% of the teacher–child CER gap within the four duration intervals. This consistency supports character recognition as a shared improvement target across utterance lengths.

Table 12: Teacher–child CER gap (percentage points), the share of the gap (%) contributed by each error type, and the child MLU bias each system produces, on the segmented test partition.

System	Gap (pp)	Error contribution (%)			Child MLU bias (%)
		Substitution	Deletion	Insertion	
From-scratch ASR (attention rescoring)					
Transformer (Vaswani et al., 2017)	30.06	76.0	16.3	7.8	−4.0
Branchformer (Peng et al., 2022)	30.18	77.1	13.7	9.2	−1.9
Conformer-31M (Gulati et al., 2020)	30.48	77.7	11.4	10.9	−0.6
Pretrained ASR (zero-shot)					
Whisper-small (Radford et al., 2023)	39.48	70.2	11.5	18.3	−1.8
Whisper-medium (Radford et al., 2023)	36.30	74.3	11.1	14.6	−3.4
Whisper-large-v3 (Radford et al., 2023)	35.22	75.0	10.4	14.6	−3.0
SenseVoice-Small (An et al., 2024)	26.78	72.0	11.2	16.8	+4.3
Qwen3-ASR-0.6B (Shi et al., 2026)	22.71	75.3	14.8	9.8	−1.4
Qwen3-ASR-1.7B (Shi et al., 2026)	19.72	74.6	13.8	11.6	−0.5
WenetSpeech-Conformer (Zhang et al., 2022)	30.25	−55.5	157.1	−1.6	−75.2
Large audio-language models (zero-shot)					
Qwen2.5-Omni (Xu et al., 2025b)	30.09	73.4	12.8	13.8	−0.1
Kimi-Audio (KimiTeam et al., 2025)	48.55	69.5	7.6	22.9	+10.5
Step-Audio 2 (Wu et al., 2025)	35.98	59.1	11.5	29.4	+3.7
Qwen3-Omni (Xu et al., 2025c)	33.57	49.4	10.3	40.2	+12.6

Table 13: Error rates (%) of zero-shot Qwen3-ASR-1.7B by utterance duration and speaker role. Sub, Del, Ins: substitution, deletion, and insertion rates over reference characters. Gap: child minus teacher CER; Sub % and Del %: shares of the gap due to substitutions and deletions.

Duration	Role	CER	Sub	Del	Ins	Gap	Sub %	Del %
<1 s	Teacher	25.93	19.27	3.80	2.86	21.81	66.8	27.5
	Child	47.74	33.85	9.80	4.10			
1–2 s	Teacher	14.78	10.73	2.03	2.02	22.86	76.5	14.8
	Child	37.63	28.21	5.41	4.01			
2–4 s	Teacher	8.01	5.82	0.89	1.30	21.06	75.3	13.0
	Child	29.07	21.67	3.63	3.78			
≥4 s	Teacher	4.86	3.33	0.59	0.94	14.69	72.0	12.8
	Child	19.55	13.91	2.47	3.18			

C.4 Training Data Scaling

MPCS supports controlled measurement of the value of classroom supervision across model families. We evaluate SenseVoice-Small, Whisper-small, Whisper-medium, and Qwen3-ASR-0.6B at eight training proportions (Figure 8). The four models share nested training subsets sampled by region and role with seed 777; development and test sets remain fixed. Proportions are taken over training speech duration, and the full pool contains 194,884 segments and 162.05 hours before model-specific length filtering. Each nonzero setting starts from the original pretrained checkpoint; 0% denotes zero-shot inference, and 100% uses the full-data adaptation result.

A small amount of classroom supervision yields substantial adaptation gains. With 5% of the training data, child CER falls from 42.17% to 32.65% for SenseVoice-Small, from 72.90% to 46.80% for Whisper-small, and from 60.23% to 39.88% for Whisper-medium. Further data improve these results to 30.44%, 39.30%, and 34.30%, respectively. Qwen3-ASR-0.6B starts from a stronger zero-shot child CER of 32.91% and reaches 30.73% with full supervision. These trajectories show how MPCS can guide both annotation-budget allocation and model-specific adaptation.

Classroom supervision widens the ratio between child and teacher error. Before adaptation, child CER is 2.2 to 3.2 times teacher CER across the four models. Every model ends above where it started, and all four converge to between 3.5 and 3.8 at full supervision; Qwen3-ASR-0.6B, which begins at 3.23, ends at 3.79. Classroom audio supplies far more teacher speech than child speech, so the same subsets teach a model more about the speakers it already recognizes well. Closing the remaining gap therefore calls for child speech specifically rather than for more classroom recordings.

Training configurations are summarized in Appendix B.

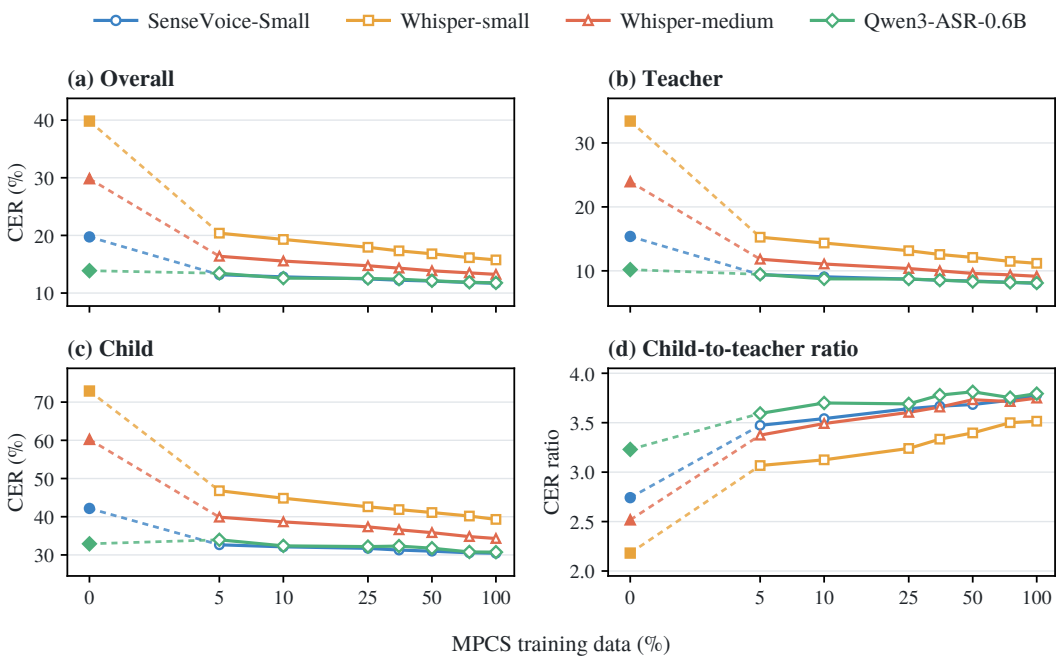

Figure 8: Overall, teacher, and child CER across eight MPCS training proportions for four pretrained ASR models, with the child-to-teacher CER ratio in panel (d). Filled markers denote zero-shot inference; the horizontal axis is logarithmic above 5%.

C.5 Recognition by Class Band

The 23 test classrooms are assigned to junior, middle, and senior bands (nominal ages 3, 4, and 5) from class metadata; the bands contain 6, 5, and 12 classrooms with 2,020, 1,762, and 5,262 child utterances. Table 14 reports classroom-mean child CER by band for the seven fine-tuned pretrained systems. All seven order the bands junior $>$ middle $>$ senior. Even in the senior band, child CER remains at least 20 points above teacher CER.

Table 14: Child CER (%) by class band on the segmented test partition for the seven MPCS-fine-tuned systems, averaged over classrooms.

System	Junior (3)	Middle (4)	Senior (5)
WenetSpeech-Conformer (Zhang et al., 2022)	38.50	36.48	34.84
Whisper-small (Radford et al., 2023)	44.26	42.89	39.79
Whisper-medium (Radford et al., 2023)	39.27	37.25	35.12
Whisper-large-v3 (Radford et al., 2023)	39.24	36.39	34.14
SenseVoice-Small (An et al., 2024)	34.74	33.04	30.96
Qwen3-ASR-0.6B (Shi et al., 2026)	37.60	33.33	31.67
Qwen3-ASR-1.7B (Shi et al., 2026)	34.59	30.78	29.41

D Domain Transfer: Additional Results

D.1 Cross-Region Transfer

Cross-region transfer measures geographic generalization within MPCS. Three representative systems are fine-tuned on the Shenzhen subset, the Guizhou subset, and the full training partition, and every checkpoint is evaluated on the Shenzhen and Guizhou subsets of the test partition (Table 15). Pooled training gives Whisper-large-v3 its lowest overall and child CER in both regions and gives Qwen3-ASR-1.7B its best Guizhou child CER (34.66%, against 38.12% for Guizhou-only and 37.45% for Shenzhen-only training) while staying within 0.6 points of region-matched training on Shenzhen. For each Shenzhen-trained model, evaluation on Guizhou raises child CER more than teacher CER, by 11.21 against 5.45 points for Qwen3-ASR-1.7B.

Table 15: Cross-region transfer. CER (%) on the Shenzhen and Guizhou test classrooms for each system fine-tuned on one region or on both (Pooled). Lower is better.

System	Training region	Shenzhen			Guizhou		
		Overall	Teacher	Child	Overall	Teacher	Child
Whisper-large-v3 (Radford et al., 2023)	Shenzhen	12.02	7.63	32.40	18.85	15.22	45.71
	Guizhou	15.82	10.64	39.88	16.26	12.85	41.43
	Pooled	11.86	7.50	32.12	15.60	12.33	39.77
SenseVoice-Small (An et al., 2024)	Shenzhen	10.82	6.93	28.90	16.36	13.04	40.94
	Guizhou	12.27	8.15	31.43	14.53	11.37	37.85
	Pooled	10.74	6.87	28.74	14.54	11.37	38.01
Qwen3-ASR-1.7B (Shi et al., 2026)	Shenzhen	9.55	5.96	26.24	14.51	11.41	37.45
	Guizhou	12.85	8.18	34.55	14.42	11.22	38.12
	Pooled	9.81	6.16	26.77	12.99	10.06	34.66

D.2 Cross-Setting Transfer with Full Training Sets

Table 16 repeats the cross-setting comparison of Section 4.1 with the full 162-hour MPCS training partition in place of the 21.8-hour subset; the ChildTalk models are those of Table 5. The Mandarin subset of ChildTalk contains 21.77 hours for training, 5.50 hours for development, and 5.64 hours for testing, including 3.64 hours of child speech in the test set. MPCS fine-tuning gives the best classroom child CER for every system and stays within 0.5–3.0 points of ChildTalk-matched fine-tuning on ChildTalk children, 2.1 points on average.

Table 16: Child CER (%) on the ChildTalk and MPCS test sets for zero-shot models and models fine-tuned on the full training set of the test corpus (Matched) or of the other corpus (Cross). Lower is better. Bold: best per column.

System	ChildTalk test set			MPCS test set		
	Zero-shot	Matched	Cross	Zero-shot	Matched	Cross
WenetSpeech-Conformer (Zhang et al., 2022)	81.82	27.98	28.51	93.32	33.90	54.86
Whisper-small (Radford et al., 2023)	50.74	28.82	30.51	72.90	39.30	66.36
Whisper-medium (Radford et al., 2023)	41.28	23.93	26.76	60.23	34.30	58.06
Whisper-large-v3 (Radford et al., 2023)	37.57	22.62	25.59	57.81	33.53	59.62
SenseVoice-Small (An et al., 2024)	23.08	18.50	20.33	42.17	30.44	**36.16**
Qwen3-ASR-0.6B (Shi et al., 2026)	19.73	20.71	22.83	32.91	30.73	53.62
Qwen3-ASR-1.7B (Shi et al., 2026)	**17.04**	**17.82**	**20.30**	**28.32**	**28.22**	49.91

E Classroom Measurement: Additional Results

E.1 Threshold Sensitivity of the Interaction Measures

Turn frequency and response rate depend on two thresholds: the gap below which consecutive teacher and child speech counts as one exchange, and the window within which a child utterance counts as a response. Table 17 varies each on the continuous development partition, holding the other fixed and applying the same value to the human reference and to the system timeline. Both measures are overestimated at every setting, and the turn-frequency overestimate barely moves with its threshold, so the inflation is a property of the timelines rather than of the thresholds.

Table 17: Interaction measures on the continuous development partition under varying thresholds, for two released role models. Turn frequency is per minute and response rate is a percentage; the human value is recomputed at each setting.

Turn frequency			Response rate		
Threshold	BabyHuBERT	Child–Adult	Threshold	BabyHuBERT	Child–Adult
3 s	+5.33	+11.62	1.0 s	+23.80	+35.85
5 s	+5.52	+11.74	2.5 s	+28.04	+41.06
10 s	+5.61	+11.83	5.0 s	+32.50	+44.77

E.2 Confidence Intervals for Language Measures

Table 18 gives the bias and ICC(A,1) behind Table 6, with 95% confidence intervals from the preschool-clustered bootstrap of Appendix B.3. On children the two measures separate without overlap: the lowest MLU agreement interval begins at 0.934 and the highest MATTR-50 interval

ends at 0.916. For the two zero-shot Whisper conditions the MATTR-50 interval lies below 0.32 over its whole range, placing most of the variance in their classroom vocabulary values in measurement error rather than in differences between classrooms.

Table 18: Bias and ICC(A,1) of MLU and MATTR-50 from automatic transcripts against human transcripts on the continuous test partition, with 95% confidence intervals in brackets. Bias is the classroom mean of system minus human, in characters for MLU and percentage points for MATTR-50.

	MLU		MATTR-50	
Transcript	**Bias [95% CI]**	**ICC(A,1) [95% CI]**	**Bias [95% CI]**	**ICC(A,1) [95% CI]**
Child, zero-shot				
Qwen3-ASR-1.7B (Shi et al., 2026)	−0.10 [−0.16, −0.04]	0.999 [0.995, 1.000]	+3.3 [2.3, 4.0]	0.765 [0.425, 0.875]
SenseVoice-Small (An et al., 2024)	+0.15 [−0.03, 0.33]	0.995 [0.977, 0.999]	+3.9 [3.0, 4.6]	0.684 [0.269, 0.818]
Whisper-large-v3 (Radford et al., 2023)	−0.46 [−0.79, −0.19]	0.977 [0.956, 0.985]	+10.0 [8.1, 11.5]	0.182 [0.038, 0.307]
Whisper-small (Radford et al., 2023)	−0.35 [−0.81, −0.04]	0.968 [0.934, 0.989]	+10.7 [8.9, 12.4]	0.170 [−0.008, 0.312]
Child, MPCS fine-tuned				
Qwen3-ASR-1.7B (Shi et al., 2026)	−0.27 [−0.39, −0.19]	0.995 [0.987, 0.998]	+2.3 [0.9, 3.6]	0.800 [0.434, 0.916]
SenseVoice-Small (An et al., 2024)	−0.32 [−0.40, −0.27]	0.995 [0.985, 0.997]	+4.9 [3.7, 6.0]	0.595 [0.234, 0.767]
Whisper-large-v3 (Radford et al., 2023)	−0.23 [−0.28, −0.18]	0.997 [0.990, 0.999]	+3.2 [1.5, 4.5]	0.749 [0.435, 0.856]
Whisper-small (Radford et al., 2023)	−0.20 [−0.29, −0.13]	0.997 [0.992, 0.999]	+5.5 [3.9, 6.8]	0.528 [0.154, 0.723]
Teacher, zero-shot				
Qwen3-ASR-1.7B (Shi et al., 2026)	+0.00 [−0.02, 0.02]	1.000 [1.000, 1.000]	+0.8 [0.4, 1.2]	0.930 [0.808, 0.980]
SenseVoice-Small (An et al., 2024)	+0.32 [0.20, 0.44]	0.994 [0.977, 0.997]	+1.3 [0.9, 1.8]	0.860 [0.647, 0.956]
Whisper-large-v3 (Radford et al., 2023)	−0.54 [−0.65, −0.43]	0.986 [0.959, 0.992]	+3.5 [2.5, 4.5]	0.346 [0.093, 0.601]
Whisper-small (Radford et al., 2023)	−0.67 [−0.85, −0.49]	0.978 [0.926, 0.989]	+5.2 [4.1, 6.4]	0.193 [0.050, 0.378]
Teacher, MPCS fine-tuned				
Qwen3-ASR-1.7B (Shi et al., 2026)	−0.08 [−0.11, −0.05]	1.000 [0.999, 1.000]	+0.6 [0.3, 0.9]	0.963 [0.882, 0.991]
SenseVoice-Small (An et al., 2024)	−0.06 [−0.09, −0.03]	1.000 [0.999, 1.000]	+1.3 [0.9, 1.7]	0.875 [0.680, 0.957]
Whisper-large-v3 (Radford et al., 2023)	−0.07 [−0.11, −0.04]	1.000 [0.999, 1.000]	+0.9 [0.5, 1.3]	0.921 [0.783, 0.977]
Whisper-small (Radford et al., 2023)	−0.08 [−0.11, −0.04]	1.000 [0.999, 1.000]	+1.3 [0.9, 1.8]	0.861 [0.652, 0.952]

E.3 Vocabulary Diversity at a Wider Window

Table 19 recomputes lexical diversity over 100-word windows. Every one of the eight transcription conditions overestimates child diversity by more at this width than at 50 words; on the classrooms common to both widths, zero-shot Qwen3-ASR-1.7B moves from +3.4 to +3.9 percentage points. Child measurement requires 100 reference words per window and is available for 14 of the 16 test classrooms; teacher measurement covers all 16.

Table 19: Bias and ICC(A,1) of MATTR-100 from automatic transcripts against human transcripts on human utterance boundaries, with 95% confidence intervals in brackets. Bias is the classroom mean of system minus human, in percentage points.

	Child		Teacher	
Transcript	**Bias [95% CI]**	**ICC(A,1) [95% CI]**	**Bias [95% CI]**	**ICC(A,1) [95% CI]**
Zero-shot				
Qwen3-ASR-1.7B (Shi et al., 2026)	+3.9 [2.9, 4.6]	0.655 [0.211, 0.814]	+1.0 [0.6, 1.3]	0.923 [0.771, 0.979]
SenseVoice-Small (An et al., 2024)	+4.4 [3.1, 5.5]	0.449 [−0.053, 0.736]	+1.4 [0.9, 1.9]	0.857 [0.617, 0.957]
Whisper-large-v3 (Radford et al., 2023)	+11.0 [9.2, 12.5]	0.160 [−0.009, 0.315]	+4.3 [3.1, 5.5]	0.271 [0.028, 0.554]
Whisper-small (Radford et al., 2023)	+12.1 [9.9, 14.3]	0.046 [−0.037, 0.137]	+6.7 [5.3, 8.0]	0.142 [0.003, 0.337]
MPCS fine-tuned				
Qwen3-ASR-1.7B (Shi et al., 2026)	+2.7 [1.3, 4.1]	0.687 [0.251, 0.878]	+0.6 [0.3, 0.9]	0.965 [0.885, 0.991]
SenseVoice-Small (An et al., 2024)	+5.0 [3.8, 6.1]	0.521 [0.132, 0.705]	+1.4 [1.0, 1.9]	0.862 [0.634, 0.952]
Whisper-large-v3 (Radford et al., 2023)	+4.0 [2.7, 5.1]	0.610 [0.255, 0.772]	+1.0 [0.7, 1.4]	0.913 [0.747, 0.974]
Whisper-small (Radford et al., 2023)	+6.0 [4.3, 7.7]	0.338 [0.033, 0.573]	+1.4 [1.0, 2.0]	0.854 [0.612, 0.950]

E.4 Measures on the Automatic Timeline

Table 20 recomputes the child measures with boundaries and roles taken from the adapted BabyHuBERT timeline rather than from the human annotation, holding everything else fixed. Characters per minute keep the agreement they reach on human boundaries, while mean length of utterance loses it in every condition, because the automatic timeline divides the same characters among more utterances. The two again separate without overlap, the lowest characters-per-minute interval beginning

at 0.974 and the highest MLU interval ending at 0.958. Taking only durations from the automatic timeline and keeping the human utterance count moves the MLU bias to $+0.46$ characters; keeping human durations and taking only the utterance count moves it to -1.31. The shift comes from the count.

Table 20: Bias and ICC(A,1) of child measures computed on the adapted BabyHuBERT timeline against the human reference, with 95% confidence intervals in brackets. Bias is the classroom mean of system minus human, in characters for MLU and characters per minute for the rate.

	Mean length of utterance		Characters per minute	
Transcript	**Bias [95% CI]**	**ICC(A,1) [95% CI]**	**Bias [95% CI]**	**ICC(A,1) [95% CI]**
Zero-shot				
Qwen3-ASR-1.7B (Shi et al., 2026)	−1.34 [−1.88, −0.87]	0.889 [0.635, 0.953]	+0.04 [−0.53, 0.63]	0.998 [0.994, 0.999]
SenseVoice-Small (An et al., 2024)	−1.13 [−1.74, −0.60]	0.894 [0.641, 0.958]	+0.76 [0.01, 1.43]	0.995 [0.988, 0.998]
Whisper-large-v3 (Radford et al., 2023)	−1.65 [−2.28, −1.13]	0.826 [0.582, 0.904]	−1.01 [−2.03, −0.22]	0.992 [0.974, 0.998]
Whisper-small (Radford et al., 2023)	−1.48 [−2.14, −0.96]	0.835 [0.669, 0.903]	−0.58 [−1.89, 0.33]	0.990 [0.976, 0.997]
MPCS fine-tuned				
Qwen3-ASR-1.7B (Shi et al., 2026)	−1.49 [−2.02, −1.03]	0.874 [0.628, 0.939]	−0.49 [−1.11, 0.06]	0.997 [0.992, 0.999]
SenseVoice-Small (An et al., 2024)	−1.58 [−2.14, −1.08]	0.860 [0.581, 0.936]	−0.71 [−1.38, −0.13]	0.996 [0.989, 0.999]
Whisper-large-v3 (Radford et al., 2023)	−1.45 [−1.98, −0.98]	0.877 [0.618, 0.944]	−0.31 [−0.87, 0.19]	0.998 [0.994, 0.999]
Whisper-small (Radford et al., 2023)	−1.38 [−1.91, −0.90]	0.881 [0.642, 0.945]	−0.13 [−0.84, 0.53]	0.997 [0.990, 0.999]

Table 21 gives the agreement behind Figure 3(a). Characters per minute agree on both boundaries, mean length of utterance loses agreement only on the automatic timeline, and lexical diversity agrees to the same degree on both: recognition sets its loss, and segmentation adds little.

Table 21: ICC(A,1) of child measures against the human reference on human boundaries and on the adapted BabyHuBERT timeline, the values summarized in Figure 3(a).

	Characters per minute		Mean length of utterance		MATTR-50	
Transcript	**Human**	**Automatic**	**Human**	**Automatic**	**Human**	**Automatic**
Zero-shot						
Qwen3-ASR-1.7B (Shi et al., 2026)	1.000	0.998	0.999	0.889	0.765	0.744
SenseVoice-Small (An et al., 2024)	0.998	0.995	0.995	0.894	0.684	0.695
Whisper-large-v3 (Radford et al., 2023)	0.992	0.992	0.977	0.826	0.182	0.117
Whisper-small (Radford et al., 2023)	0.988	0.990	0.968	0.835	0.170	0.099
MPCS fine-tuned						
Qwen3-ASR-1.7B (Shi et al., 2026)	0.998	0.997	0.995	0.874	0.800	0.794
SenseVoice-Small (An et al., 2024)	0.998	0.996	0.995	0.860	0.595	0.538
Whisper-large-v3 (Radford et al., 2023)	0.999	0.998	0.997	0.877	0.749	0.743
Whisper-small (Radford et al., 2023)	0.999	0.997	0.997	0.881	0.528	0.555

E.5 Agreement of the Interaction Measures

Table 22 gives the agreement behind the measurement columns of Table 8.

Table 22: Agreement of each role timeline with the human timeline on the continuous test partition, averaged over three seeds. Bias is the classroom mean of system minus human, in the unit of each measure.

Timeline	Bias	Pearson r	ICC(A,1)
Teacher speaking-time share (%), human mean 50.04			
BabyHuBERT-VTC (Charlot et al., 2025)	−0.07	0.998	0.998
Child–Adult Diarization (Xu et al., 2025a)	−0.06	0.995	0.995
WSW 2.0 (Sun et al., 2025)	+2.25	0.940	0.922
ResNet (Kothalkar et al., 2024)	−0.92	0.925	0.921
Child speaking-time share (%), human mean 14.50			
BabyHuBERT-VTC (Charlot et al., 2025)	+0.09	0.996	0.996
Child–Adult Diarization (Xu et al., 2025a)	+0.02	0.993	0.994
WSW 2.0 (Sun et al., 2025)	+0.71	0.939	0.935
ResNet (Kothalkar et al., 2024)	+2.22	0.880	0.847
Teacher–child turns per minute, human mean 3.45			
BabyHuBERT-VTC (Charlot et al., 2025)	+0.60	0.992	0.950
Child–Adult Diarization (Xu et al., 2025a)	+1.19	0.945	0.810
WSW 2.0 (Sun et al., 2025)	+1.45	0.915	0.738
ResNet (Kothalkar et al., 2024)	+1.87	0.747	0.544
Child response rate (%), human mean 21.19			
BabyHuBERT-VTC (Charlot et al., 2025)	+2.98	0.984	0.951
Child–Adult Diarization (Xu et al., 2025a)	+6.06	0.923	0.815
WSW 2.0 (Sun et al., 2025)	+10.77	0.930	0.663
ResNet (Kothalkar et al., 2024)	+12.79	0.800	0.528

E.6 Transcription Against a Duration-Only Estimate

Child utterance length carries timing information. A single speaking rate calibrated on the continuous training partition, 3.12 characters per second, applied to each classroom's child speech duration and utterance count, matches human utterance lengths at an intraclass correlation of 0.951 and human characters per minute at 0.982.

Table 23 reports what a transcript adds over that estimate, as the per-classroom difference in absolute error. All eight conditions improve on it, most clearly when utterance length is counted in words, where the count follows lexical content rather than character totals.

Table 23: Reduction in absolute error from using a transcript in place of the duration-only estimate, per classroom, with 95% confidence intervals from joint resampling of training and test preschools. Positive values favor the transcript.

Transcript	MLU, characters	MLU, words	Characters per minute
Zero-shot			
Qwen3-ASR-1.7B (Shi et al., 2026)	0.70 [0.42, 1.17]	0.56 [0.37, 0.82]	1.89 [1.10, 2.97]
SenseVoice-Small (An et al., 2024)	0.57 [0.33, 0.92]	0.30 [0.12, 0.53]	1.44 [0.77, 2.40]
Whisper-large-v3 (Radford et al., 2023)	0.33 [−0.13, 0.93]	0.33 [0.03, 0.67]	0.80 [−0.41, 2.16]
Whisper-small (Radford et al., 2023)	0.35 [−0.18, 0.95]	0.34 [0.00, 0.70]	0.86 [−0.65, 2.29]
MPCS fine-tuned			
Qwen3-ASR-1.7B (Shi et al., 2026)	0.57 [0.25, 1.05]	0.53 [0.32, 0.80]	1.45 [0.71, 2.47]
SenseVoice-Small (An et al., 2024)	0.52 [0.20, 1.02]	0.54 [0.32, 0.80]	1.29 [0.51, 2.29]
Whisper-large-v3 (Radford et al., 2023)	0.61 [0.30, 1.10]	0.56 [0.33, 0.87]	1.59 [0.78, 2.65]
Whisper-small (Radford et al., 2023)	0.63 [0.31, 1.14]	0.59 [0.36, 0.87]	1.58 [0.76, 2.63]

E.7 Controlled Perturbations at a Fixed Error Rate

The paired reference turns the principle of Section 4.4 into a controlled experiment. Errors of a single kind are injected into the human annotation of the 16 test classrooms at a fixed rate, and every measure is recomputed with the scoring of Appendix B.3 over 50 random seeds, with 95% confidence intervals from 2,000 preschool-clustered bootstrap replicates. Text perturbations spend 20% of each classroom's reference characters per role on pure substitutions, deletions, or insertions, with new characters drawn from the character frequencies of the continuous training partition. Timeline perturbations flip roles inside reference segments at a fixed DER, in equal amounts in each direction, and vary only the segment length; both roles keep their reference talk time.

On the text side (Table 24), one error rate yields three outcomes: substitutions leave mean length of utterance exact and raise lexical diversity the most, while deletions and insertions move utterance length by 20% in opposite directions. Teacher speech follows the same pattern.

Table 24: Child measures under text perturbations at 20% CER on the continuous test partition: bias against the human transcript (characters for MLU, percentage points for MATTR-50) with 95% confidence intervals, and ICC(A,1).

	MLU		MATTR-50	
Error kind	**Bias [95% CI]**	**ICC(A,1)**	**Bias [95% CI]**	**ICC(A,1)**
Substitution	0.00 [0.00, 0.00]	1.000	+11.3 [10.0, 12.5]	0.159
Deletion	−1.97 [−2.35, −1.69]	0.815	+6.7 [5.9, 7.4]	0.431
Insertion	+1.97 [1.69, 2.35]	0.869	+9.2 [8.2, 10.0]	0.266

On the timeline side (Table 25), the same 2% of role confusion raises turn frequency by 17% as 2.6-second segments and by 146% as 0.3-second segments; the added units also lower mean length of utterance by 13% to 52% and lift the child response rate by 3 to 21 points. With half the DER, 1% in 0.3-second segments, turns are still inflated by 74%: the turn error is 4.5 times that of the 2.6-second condition (95% CI 4.37 to 4.57) and larger in all 16 classrooms. The adapted BabyHuBERT timeline matches the 2.6-second condition in added child units, turn inflation, response rate, and utterance length.

Table 25: Interaction and length measures under timeline perturbations of the human reference on the continuous test partition, with the adapted BabyHuBERT timeline of Table 8 for comparison. Units: child speech units added per classroom. Turns per minute with relative bias in color; child response rate as bias in percentage points; child MLU as relative bias (%) with the human text held fixed. Human reference: 3.45 turns per minute, 21.2% response rate, MLU 9.84.

Timeline	**DER (%)**	**Units**	**Turns/min**	**Response**	**MLU**
2.6-second segments	2.0	8.6	4.03 +16.7	+3.2	−12.7
0.95-second segments	2.0	23.5	5.02 +45.4	+7.9	−26.7
0.3-second segments	2.0	75.8	8.50 +146.4	+20.7	−51.5
0.3-second segments	1.0	38.5	6.02 +74.4	+12.0	−36.4
Adapted BabyHuBERT	1.41	8.6	4.05 +17.3	+3.0	−12.7