ChildTalk: A Multi-Dialect Chinese Child Speech Corpus with Full-Length Child–Caregiver Conversations for Speech Recognition

Jiaming Zhou[*], **Yujie Guo**[*], **Shiwan Zhao, Yao Lu, Jianye Wang, Haoqin Sun, Hui Wang, Yong Qin**[†]
College of Computer Science, Nankai University,
Correspondence: zhoujiaming@mail.nankai.edu.cn, qinyong@nankai.edu.cn

Abstract

Automatic speech recognition (ASR) for children remains challenging due to developmental variability and the scarcity of high-quality corpora, especially for Mandarin and its dialects. In this paper, we present *ChildTalk*, a large-scale Chinese child speech corpus designed to address this gap. It contains 112.5 hours of speech from 498 children (aged 2–8) and 500 caregivers, recorded as natural child–caregiver conversations. Unlike prior Mandarin child ASR corpora that mainly release isolated utterances, *ChildTalk* provides full-length dialogues with complete transcriptions, preserving turn-taking and discourse context. To our knowledge, it is the first publicly available Mandarin child speech corpus with full-length dialogues and systematic coverage of standard Mandarin, eight Mandarin dialect subgroups, and two additional dialects (Southern Min and Jin). We benchmark end-to-end models trained from scratch, large pre-trained ASR models fine-tuned on *ChildTalk*, omni-modal LLMs in a zero-shot setting, and commercial speech transcription APIs. Fine-tuning on *ChildTalk* consistently improves both in-domain and cross-domain performance. These results indicate that *ChildTalk* provides a challenging, broad-coverage testbed for Chinese child ASR, dialect robustness, and dialogue-level modeling. The dataset is freely available for all academic purposes on https://github.com/NKU-HLT/ChildTalk.

1 Introduction

Automatic speech recognition (ASR) has made substantial progress in recent years, driven by large-scale pre-trained models (e.g., Whisper (Radford et al., 2023a; An et al., 2024)) and self-supervised learning approaches (e.g., Wav2vec 2.0 (Baevski et al., 2020), HuBERT (Hsu et al., 2021)). However, accurately recognizing child speech remains challenging. Physiological and developmental characteristics such as higher fundamental frequency, shorter vocal tracts, unstable articulation, and large pronunciation variability (Lee et al., 1997; Gerosa et al., 2009; Kennedy et al., 2017) create a substantial mismatch between adult-dominated training corpora and the target domain of child speech, leading to markedly degraded performance on children's voices (Fan et al., 2024).

Progress in child ASR critically depends on high-quality corpora, yet Mandarin resources remain limited, especially for younger speakers, conversational speech, and dialectal variation. ASR-oriented corpora such as CASS CHILD (Gao et al., 2012), SLT-CSRC (Yu et al., 2021), SingaKids-Mandarin (Chen et al., 2016), and ChildMandarin (Zhou et al., 2025) provide valuable data but are either not publicly accessible, dominated by reading-style speech, or released only as segmented utterances in standard Mandarin (Table 1). Meanwhile, several Mandarin child-language corpora within CHILDES[1] (e.g., Tong Corpus (Xiangjun and Yip, 2017), BJCMS (Mai et al., 2024), HKU-70 (Leung, 2018)) offer session-level recordings with transcripts, but involve relatively few speakers, focus on specific elicitation settings, and lack systematic coverage of Mandarin dialect subgroups. As a result, there is still no large-scale, publicly available Mandarin child speech corpus that provides full-length dialogues with broad dialect coverage in an ASR-ready format.

In this paper, we introduce **ChildTalk**, a large-scale Chinese child speech corpus designed to fill this gap. As summarized in Table 1, *ChildTalk* contains 112.5 hours of speech from 498 children (aged 2–8) and 500 caregivers, recorded as natural child–caregiver conversations. Unlike prior Mandarin child ASR corpora that mainly release isolated utterances, *ChildTalk* provides full-length

[*]Equal contribution
[†]Yong Qin is the corresponding author.
[1]https://talkbank.org/childes/access/Chinese/

Findings of the Association for Computational Linguistics: ACL 2026, pages 5103–5116

Corpus	Age range	# Children	Dur. (hrs)	Type	Full dialogue	Dialect	Avail.
CASS CHILD	1–4	23	80.0	Spontaneous speech	P	✗	✗
SLT-CSRC C1	7–11	927	28.6	Reading	✗	✗	✗
SLT-CSRC C2	4–11	54	29.5	Conversation	✗	✗	✗
SingaKids	7–12	255	75.0	Reading	✗	✗	✓
ChildMandarin	3–5	397	41.3	Conversation	✗	✗	✓
Tong Corpus	1;7–3;4	1	22.0	Interactions	✓	✗	✓
BJCMS	3–6;9	48	–	Conversation	✓	✗	✓
HKU-70	2;6–5;6	70	–	Conversation	✓	Cantonese	✓
ChildTalk	2–8	498	112.5	Conversation	✓	Mandarin + 10 varieties[†]	✓

Table 1: Summary of Chinese child ASR datasets: age range, speaker count, duration, interaction type, dialectal coverage, full-dialogue availability, and corpus availability. Dur.: Speech duration. Avail.: public availability. [†]ChildTalk covers standard Mandarin and 10 additional dialect varieties (eight Mandarin dialect subgroups and two non-Mandarin dialects: Southern Min and Jin Chinese).

dialogues with complete transcriptions, preserving turn-taking and discourse-level context. It systematically covers standard Mandarin, eight Mandarin dialect subgroups, and two additional dialects (Southern Min and Jin), enabling studies of dialectal robustness and cross-variety modeling in child ASR.

We benchmark four families of systems: end-to-end models trained from scratch, large pre-trained ASR models fine-tuned on *ChildTalk*, omni-modal LLMs used in a zero-shot setting, and commercial speech transcription APIs. We further show that incorporating short dialogue history as textual prompts yields additional gains, indicating that discourse context is helpful for recognizing spontaneous child speech. Overall, the results suggest that *ChildTalk* is challenging for off-the-shelf systems, while fine-tuning on the corpus and leveraging context provide consistent improvements, highlighting its value as a testbed for child ASR, dialect robustness, and dialogue-level modeling in Chinese.

2 Related Work

2.1 Child Corpora in Chinese

As summarized in Table 1, publicly available Chinese child speech corpora with both full-dialogue recordings and dialectal coverage are very limited. CASS CHILD (Gao et al., 2012) and SLT-CSRC (Yu et al., 2021) are important resources but are not broadly accessible and mainly target standard Mandarin, with data released as segmented utterances rather than full conversational sessions. Among accessible corpora, SingaKids (Chen et al., 2016) focuses on reading tasks from older children, while ChildMandarin (Zhou et al., 2025) provides conversational speech from preschoolers but only in utterance-level form and without systematic dialectal design.

Several linguistic corpora offer full-length conversational sessions, such as the Tong Corpus (Xiangjun and Yip, 2017), BJCMS (Mai et al., 2024), and HKU-70 (Leung, 2018). These resources, however, were primarily constructed for longitudinal acquisition or clinical studies and therefore involve relatively few speakers (e.g., a single child in the Tong Corpus) and limited dialectal coverage; HKU-70 is Cantonese-only, and Mandarin dialects are largely absent. Overall, existing Chinese child speech corpora tend to trade off between scale, natural dialogue structure, and dialectal diversity, leaving few publicly accessible resources that jointly offer large speaker coverage, full-length conversations, and multiple dialect varieties.

2.2 Child Corpora in Other Languages

Child speech corpora in other languages show similar trade-offs. For English, large-scale resources such as the MyST Corpus (Demuth and Tremblay, 2008) and TBALL (Kazemzadeh et al., 2005) include hundreds or thousands of speakers, but typically focus on school-aged children in structured tasks (e.g., tutoring or reading), with limited attention to early preschool speech or native dialectal variation. Other corpora, such as the Non-Native Children's Speech Corpus (Radha and Bansal, 2022), specifically target L2 learners and remain relatively small in duration.

For non-English languages, resources are generally smaller and more sparse. The Dutch JASMIN-CGN (Cucchiarini et al., 2008) and Swedish NICE (Bell et al., 2005) corpora, as well as Spanish datasets like CHIEDE (Garrote and Moreno San-

Relationship	Count	Proportion
Parents	393	77.98%
Grandparents	13	2.58%
Older siblings	52	10.32%
Teacher	35	6.94%
Relative	11	2.18%

Table 2: Statistic of the relationship between children and adults.

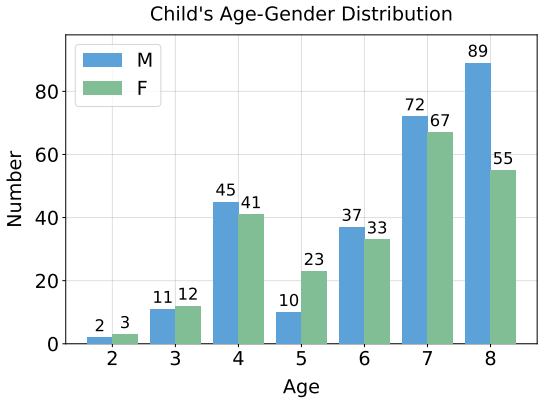


Figure 1: Age and gender distribution of child participants in ChildTalk.

doval, 2008) and IESC-Child (Pérez-Espinosa et al., 2020), provide valuable material but tend to concentrate on older children, narrow age bands, or specific interaction scenarios, and rarely include full-length dialogues with systematic dialectal coverage.

Against this background, *ChildTalk* contributes a complementary type of resource: a publicly available Chinese child speech corpus with full-length child–caregiver conversations, broad age coverage (2–8 years), and a systematic design spanning standard Mandarin and ten additional dialect varieties, aimed at supporting robust and dialect-aware ASR research.

3 Dataset Description

Overall, *ChildTalk* contains 498 child speakers and 500 adult speakers, yielding 88,943 utterances and 112.50 hours of speech across Mandarin and multiple dialects. In this section, we first describe the data collection protocol (§3.1), then present the annotation process (§3.2), and finally provide a detailed analysis of the corpus statistics (§3.3).

3.1 Dataset Collection

We recruit 498 children aged 2–8 and 500 adults from mainland China and record 500 dyadic conversations. Speakers are distributed across standard Mandarin, eight Mandarin dialect subgroups, and two additional dialects (Southern Min and Jin Chinese). The Mandarin and dialect subsets are kept disjoint: no speaker appears in both subsets, and each child participates in at most one variety (either Mandarin or a single dialect).

Prior to participation, we obtain written informed consent from parents or legal guardians. Guardians are fully informed of the purpose of the study, the recording procedures, and the intended academic use of the data. Recordings are conducted in natural conversational settings. Each child is accompanied by a familiar adult such as a parent, grandparent, older sibling, teacher, or other close relative; the distribution of relationship types is summarized in Table 2. The dyads are encouraged to engage in everyday conversations, so that children can speak in a relaxed and familiar atmosphere, eliciting spontaneous and age-appropriate speech.

Dialogue corpora with full-length recordings and aligned annotations are particularly scarce, especially for child speech. Unlike prior Mandarin child ASR corpora that typically release isolated utterances, **ChildTalk** captures complete parent–child dialogues in long-form audio sessions. These recordings preserve natural turn-taking (e.g., parent–child alternation), contextual dependencies between utterances, and broader discourse dynamics such as topic continuation and shifts. All sessions are recorded in quiet indoor environments to minimize background noise while maintaining naturalistic conversational style.

Each session typically lasts several minutes, resulting in continuous long-form speech segments. After collection, the recordings are carefully segmented and transcribed at the utterance level. All data are anonymized by replacing personal identifiers with coded labels. The corpus will be released exclusively for research purposes under a non-commercial license, in accordance with ethical standards for child data collection and use.

3.2 Data Annotation

To ensure high-quality annotation, all recordings are transcribed at the character level by professional annotators following consistent guidelines. The demographic information of 43 annotators, in-

Token	Description
<UNK/>	Unintelligible word/phrase
<LAUGH/>	Laughter
<COUGH/>	Coughing or throat clearing
<NON/>	Non-speech noises (e.g., ring)
<NPS/>	Noise from non-participants
<MUSIC/>	Music segment
<OVERLAP></OVERLAP>	Overlapping speech
<EN></EN>	English words
<PINYIN></PINYIN>	Pinyin sequences
<SING></SING>	Singing content
<SIGH/>	Sigh
<LIP-SMACK/>	Lip-smacking sound
<STA/>	Stationary background noise
#word	Filler words marked with "#"
<S>	Silence part
<Z>	Invalid or discarded part

Table 3: Special tokens used in ChildTalk transcription.

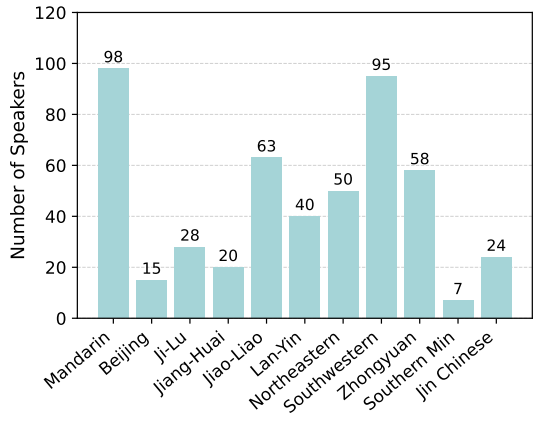


Figure 2: Number of child speakers by dialect in ChildTalk.

cluding gender, age, education and native place, are provided in Table A.1. Transcriptions faithfully capture children's speech, including disfluencies, stutters, repetitions, minor mispronunciations (transcribed as intended words), and age-specific patterns, while unintelligible speech is marked as <UNK/>. Numbers follow their spoken form, license plates are preserved, and segments must contain at least three characters (except common interjections or fillers). English words are enclosed in <EN>. . . </EN>, alphabetic spellings (e.g., DNA) are written with spaces, and Pinyin sequences use <PINYIN>. . . </PINYIN>. Non-speech events such as laughter, coughing, noise, or overlapping speech are marked with dedicated tokens; extended silence is <S>, and discarded segments are <Z>. This standardized annotation protocol ensures high-quality linguistic content while

Split	# Child	# Adult	Utt.	Dur. (hrs)	Avg. (s)
Train	342	344	64370	79.06	4.42
Dev	72	72	10387	14.46	5.01
Test	84	84	14186	18.98	4.82
Sum	498	500	88943	112.50	4.55

Table 4: Summary of data splits. # Child and # Adult denotes the number of childern speakers and adult speakers, respectively.

preserving child-specific characteristics crucial for ASR research. Details of all special tokens in ChildTalk is shown in Table 3.

3.3 Data Analysis

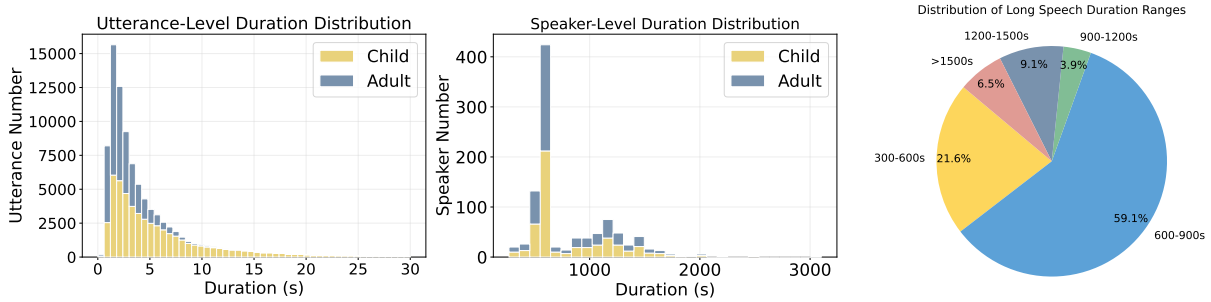
3.3.1 Child Speaker Demographics and Dialect Distribution

Figure 1 illustrates the age–gender distribution of the 498 child participants in *ChildTalk*. The corpus is skewed toward older children: ages 7 and 8 account for more than half of the speakers (139 and 144 children, respectively), while ages 2–3 are relatively sparse, reflecting the practical difficulty of recruiting very young participants. Gender is reasonably balanced across ages, with a slight overall predominance of boys.

Figure 2 shows the dialect distribution of child speakers in *ChildTalk*. Mandarin has the largest number of speakers (98), followed closely by Southwestern Mandarin (95), providing two high-resource varieties for model training and evaluation. A mid-sized group includes Jiao–Liao (63), Zhongyuan (58), Northeastern (50), and Lan–Yin (40), each with sufficient speakers for reliable evaluation and cross-dialect comparison. Several other dialects are smaller in scale, such as Southwest Min (7), Beijing (15), Jiang–Huai (20), Jin Chinese (24), and Ji–Lu (28). These varieties are inherently difficult to recruit for child speech collection, yet *ChildTalk* still achieves meaningful coverage for each of them, offering rare data for studying dialectal robustness and adaptation in child ASR. Further details on the geographic distribution of child participants are provided in Figure A.1.

3.3.2 Duration Statistics

As shown in Figure 3a, we present duration statistics of pure-speech segments in *ChildTalk* at both the utterance level and the speaker level. The left panel illustrates the utterance-level duration distribution for child and adult speech. Almost all short recordings are shorter than 30 s, and the majority



(a) Duration distributions of short pure-speech segments at the utterance and speaker levels.

(b) Duration distribution of long conversations.

Figure 3: Duration Statistics of Utterances, Speakers, and Conversations.

Dialect	# Spk.	Age range	Avg. age	Utt.	Proportion	Duration (hrs)			
						Total	Train	Dev	Test
Mandarin	198	[2, 4]	3.67	38,098	29.24%	32.90	21.76	5.50	5.64
Beijing	30	[6, 8]	6.87	1,777	2.12%	2.38	1.57	0.33	0.48
Ji-Lu	56	[5, 8]	7.29	2,377	4.03%	4.53	3.08	0.61	0.84
Jiang-Huai	40	[4, 8]	6.05	1,760	2.02%	2.27	1.43	0.52	0.31
Jiao-Liao	126	[4, 8]	7.10	8,909	8.66%	9.75	7.02	1.26	1.47
Lan-Yin	80	[4, 8]	6.35	2,354	6.68%	7.52	5.26	1.01	1.25
Northeastern	100	[7, 8]	7.62	6,153	7.15%	8.04	5.64	1.14	1.26
Southwestern	190	[4, 8]	6.35	18,729	27.66%	31.11	23.97	2.05	5.09
Zhongyuan	116	[4, 8]	6.88	6,607	7.99%	8.99	6.06	1.43	1.51
Southern Min	14	[6, 8]	7.00	403	1.00%	1.13	0.68	0.14	0.31
Jin Chinese	48	[5, 8]	7.29	1,776	3.45%	3.88	2.58	0.48	0.82

Table 5: Dialect-wise statistics of child speech in *ChildTalk*. "# Spk." denotes the number of child speakers per subset, "Proportion" is the share of each subset in the corpus, and the last four columns report total and per-split durations in hours.

of utterances are concentrated in the range of 0–15 s, indicating a conversational style dominated by short turns.

The middle panel aggregates utterances at the speaker level. For most speakers, the total duration of pure speech lies below 1500 s, with a large mass between several hundred seconds and around 1500 s. This suggests that each speaker contributes a substantial yet balanced amount of data, which is helpful for building speaker-independent models.

The right panel in Figure 3b further reports the duration distribution of complete conversations. A clear majority of long recordings fall into the 600–900 s (10–15 minutes) range, with about one fifth in the 300–600 s (5–10 minutes) range and only a small proportion exceeding 1500 s (>25 minutes). Overall, the corpus provides many dialogues of moderate length, suitable for both utterance-level and dialogue-level modeling.

3.3.3 Data Splits and Dialect-wise Statistics

We divide the dataset into speaker-independent train, development, and test subsets with a ratio of 0.70:0.15:0.15. The summary of data splits is provided in Table 4. Overall, *ChildTalk* contains 498 child speakers and 500 adult speakers, yielding 88,943 utterances and 112.50 hours of speech. The training, development, and test sets comprise 79.06, 14.46, and 18.98 hours, respectively. The average utterance duration is around 4.5 seconds and is similar across splits (4.42–5.01 seconds), indicating consistent segmentation.

Table 5 presents a detailed breakdown by dialect subset. Mandarin and Southwestern Mandarin account for the largest portions of the corpus (32.9 and 31.11 hours), while Jiao–Liao, Zhongyuan, Northeastern, and Lan–Yin form medium-sized subsets (approximately 4–10 hours). Smaller but still meaningful amounts of data are available for Beijing, Ji–Lu, Jiang–Huai, Jin Chinese, and South-

Model	#Params	CTC greedy search	CTC beam search	Attention	Attention rescoring
Transformer	29.80M	42.09	41.91	49.86	40.46
Branchformer	29.01M	39.02	38.96	49.20	37.79
Conformer	31.94M	**37.08**	**37.04**	**41.16**	**35.94**

Table 6: CER (%) of models trained from scratch on *ChildTalk*.

Model	# Params	Decoding method	In-domain		Cross-domain	
			Zero-shot	Fine-tuned	Zero-shot	Fine-tuned
Whisper-tiny	39M	Autoregressive	75.29	43.47	67.63	34.25
Whisper-base	74M	Autoregressive	64.70	38.18	51.49	30.53
Whisper-small	244M	Autoregressive	51.77	30.49	37.99	21.55
Whisper-medium	769M	Autoregressive	43.94	24.16	28.55	16.05
CW	122M	Attention Rescoring	31.02	21.53	18.05	13.05
SenseVoice-Small	234M	CTC greedy search	**20.63**	**16.30**	**11.79**	**10.19**

Table 7: Performance of zero-shot and finetuned models trained on the ChildTalk training set. "In-domain" denotes results on the ChildTalk test set, while "Cross-domain" denotes results on the ChildMandarin test set, illustrating the robustness of our dataset.

ern Min. The Mandarin subset focuses on younger children (2–4 years, average 3.67), whereas most dialect subsets cover slightly older children (average ages between 6 and 8), providing a range of developmental stages. Each dialect appears in all three splits with non-negligible durations, so that train, development, and test sets all reflect the overall dialectal composition of the corpus.

4 Experiments

4.1 Models

We evaluate models trained from scratch, pre-trained models fine-tuned on *ChildTalk*, omni-modal LLMs used in a zero-shot setting, and commercial speech transcription APIs.

For training from scratch, we consider three end-to-end ASR architectures implemented with Wenet (Yao et al., 2021): Transformer (Dong et al., 2018), Conformer (Gulati et al., 2020), and Branchformer (Peng et al., 2022).

To assess the effect of large-scale pre-training, we further fine-tune several pre-trained models: Whisper (Radford et al., 2023b) (from `tiny` to `medium`), Conformer-WenetSpeech (CW) (Zhang et al., 2022), and SenseVoice-small (An et al., 2024).

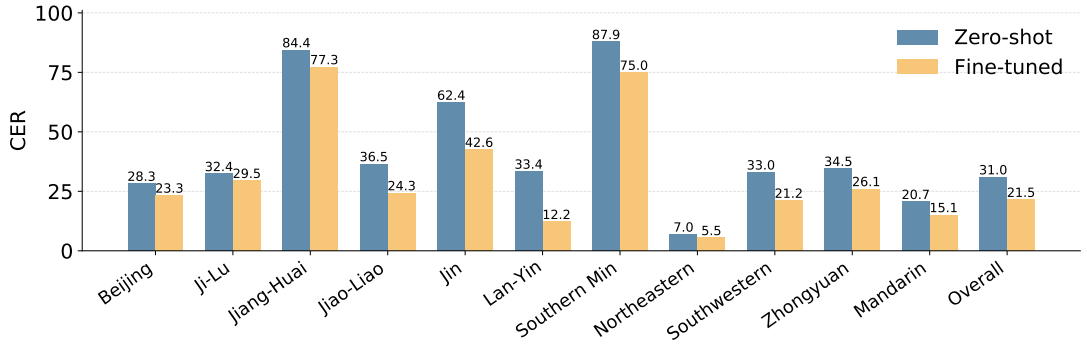
In addition, we evaluate general-purpose omni-modal LLMs in a zero-shot setting, including Qwen2.5-Omni (Xu et al., 2025a), Qwen3-Omni (Xu et al., 2025b), Qwen3-ASR (Shi et al., 2026) and OpenAI's GPT-4o (OpenAI et al., 2024) speech transcription endpoints (`gpt-4o-transcribe` and `gpt-4o-mini-transcribe`). We report these systems together as off-the-shelf zero-shot baselines that are not fine-tuned on *ChildTalk*, to contrast their performance with supervised ASR models trained or adapted on our corpus. More detailed model descriptions and hyperparameters are provided in Appendix B.

4.2 Results of Training from scratch.

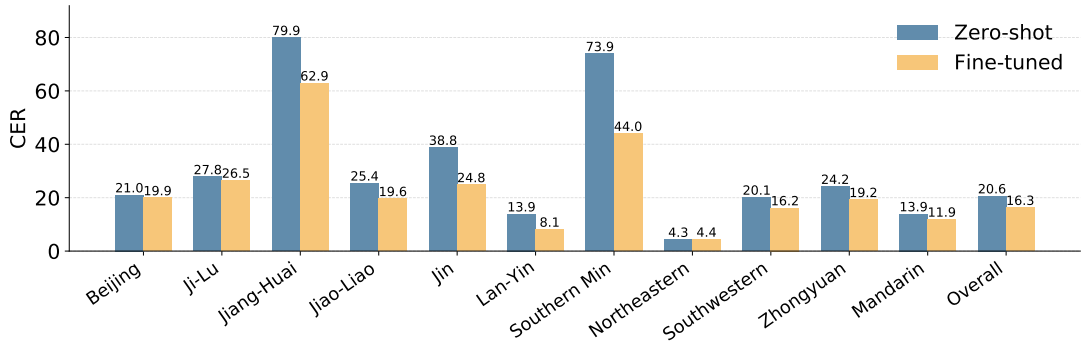
Table 6 summarizes the performance of models trained from scratch on *ChildTalk*. Among the three architectures, Conformer achieves the lowest error rates across all decoding strategies (35.94% with attention rescoring), followed by Branchformer and Transformer. Greedy and beam search perform very similarly for all models, indicating that simple CTC decoding is already reasonably strong in this setting. Attention-only decoding yields noticeably higher error rates, while attention rescoring consistently provides the best performance for each architecture, suggesting that combining CTC with an attention-based decoder offers modest but stable gains on spontaneous child speech. Despite these improvements, the remaining error rates also highlight the inherent difficulty of recognizing spontaneous child speech and leave substantial room for further gains from large-scale pre-training and adaptation.

4.3 Results of Fine-tuning Pre-trained Model

Table 7 reports zero-shot and fine-tuned results for pre-trained models on both the in-



(a) Conformer-Wenetspeech (CW)



(b) SenseVoice-small

Figure 4: Performance breakdown of SenseVoice-small and Conformer-Wenetspeech (CW) on each subdialect.

domain (ChildTalk) and cross-domain (ChildMandarin (Zhou et al., 2025)) test sets. Note that all fine-tuned results are obtained by adapting the models on the *ChildTalk* training set. Within the Whisper family, performance improves steadily with model size in all settings: larger variants obtain lower error rates in zero-shot mode and benefit more from fine-tuning.

Fine-tuning on ChildTalk consistently yields substantial gains across all models and both evaluation domains. For Whisper-medium, fine-tuning reduces CER by 19.78 points in-domain (43.94% $\rightarrow$ 24.16%) and by 12.50 points cross-domain (28.55% $\rightarrow$ 16.05%). Similar, albeit smaller, improvements are observed for the smaller Whisper variants. The two non-autoregressive models, Conformer-WenetSpeech (CW) and SenseVoice-small, achieve lower error rates overall than the Whisper family at comparable or larger parameter scales. SenseVoice-small obtains the best results in all settings, with CERs of 16.30% (in-domain) and 10.19% (cross-domain) after fine-tuning.

Comparing the in-domain and cross-domain columns, CER on ChildMandarin are generally lower than on ChildTalk, which suggests that the ChildMandarin test set is somewhat easier but also that fine-tuning on ChildTalk improves robustness beyond the training domain. All models benefit from ChildTalk-based fine-tuning on the ChildMandarin test set, indicating that the corpus provides useful supervision that transfers to another Mandarin child speech dataset with different speakers and recording conditions.

4.4 Performance Breakdown of Fine-tuned Models

Figure 4 presents the CER breakdown of Conformer-WenetSpeech (CW) and SenseVoice-small on each subdialect of the *ChildTalk* test set, comparing zero-shot performance with models fine-tuned on the *ChildTalk* training set. For CW (Figure 4a), fine-tuning reduces errors for most dialects, with particularly large absolute gains on Lan–Yin (33.4% $\rightarrow$ 12.2%) and Jin (62.4% $\rightarrow$ 42.6%). Nevertheless, some varieties such as Jiang–Huai and Southern Min remain challenging, with CERs above 75% even after adaptation, while Northeastern Mandarin is relatively easier (7.0% $\rightarrow$ 5.5%). Overall, the CW average CER decreases from 31.0% to 21.5%.

SenseVoice-small (Figure 4b) achieves lower CERs than CW for all subdialects in both zero-

shot and fine-tuned settings. Fine-tuning again yields improvements for nearly all dialects, including substantial reductions on Lan–Yin (13.9% → 8.1%), Jin (38.8% → 24.8%), and Southern Min (73.9% → 44.0%). Standard Mandarin attains relatively low CERs (13.9% → 11.9%). The overall CER for SenseVoice-small decreases from 20.6% to 16.3% after fine-tuning, indicating that adaptation on *ChildTalk* provides consistent gains across a wide range of Mandarin subdialects and related varieties.

We further examine these subdialect-wise results in light of the age distribution summarized in Table 5. Northeastern Mandarin has the highest average age among all groups (7.62 years), whereas the Mandarin subset consists mainly of younger children aged 2–4 (average 3.67 years). The relatively low error rates on Northeastern thus likely reflect a combined effect of older speakers and its phonological proximity to standard Mandarin. In contrast, dialects such as Southern Min and Jiang–Huai combine smaller sample sizes, larger phonological differences from standard Mandarin, and mid-range ages, which is in line with their higher CERs. These patterns indicate that both speaker age and dialectal variation are closely related to ASR performance on *ChildTalk*, and they should be taken into account when interpreting subdialect-wise results.

4.5 Results of Zero-shot API-based ASR Systems

Table 8 summarizes the zero-shot performance of off-the-shelf omni-modal models on the *ChildTalk* test set. All systems show noticeable difficulty with spontaneous, multi-dialect child speech, yielding CERs above 25% without any adaptation. Among them, Qwen2.5-Omni obtains the lowest error rate (25.39%), with Qwen3-Omni slightly higher at 28.56%. The GPT-4o transcription endpoints perform substantially worse in this setting, with CERs around 47–49%. Overall, these results indicate that, in a pure zero-shot setup, current general-purpose omni-modal models and API still fall short of the accuracy typically desired for robust child ASR on *ChildTalk*.

4.6 Effect of Dialogue Context length

Table 9 reports CER performance when varying the amount of dialogue history used as a textual prompt, where S, D, and I denote substitution, deletion, and insertion errors, respectively. Without any context ($C = 0$), Whisper-Large-V3 attains a CER of 43.25%. Providing the immediately preceding turn as context ($C = 1$) reduces CER to 35.16%, with most of the gain coming from fewer substitutions (28.33% → 22.00%) and deletions (11.60% → 9.65%), while insertions remain relatively small. Extending the prompt to two and three previous turns ($C = 2$ and $C = 3$) yields further but diminishing improvements, reaching a CER of 31.87% at $C = 3$. Overall, short dialogue history helps the model correct substitution and deletion errors, suggesting that even limited context is useful for recognizing child speech in conversational settings. We further verified this trend on additional Qwen-3-ASR (see Appendix C), where consistent improvements from C=0 to C=3 are observed.

Model	Zero-shot CER (%)
GPT-4o-mini-transcribe	48.65
GPT-4o-transcribe	47.44
Qwen2.5-Omni	**25.39**
Qwen3-Omni	28.56

Table 8: Performance of off-the-shelf zero-shot baselines on ChildTalk testing set.

Context	S (%)	D (%)	I (%)	CER (%)
0	28.33	11.60	**3.33**	43.25
1	22.00	9.65	3.51	35.16
2	20.67	9.15	3.69	33.50
3	**20.29**	**8.06**	3.53	**31.87**

Table 9: Effect of dialogue context length on Whisper-Large-V3. "Context" denotes the number of preceding dialogue turns used as textual prompts ($C = 0$ means no context, $C = 1$ uses the immediately preceding turn, etc.). "S,D,I" denote the error rate (%) of substitution, deletion and insertion, respectively.

5 Conclusion

In this paper, we present *ChildTalk*, a multi-dialect Chinese child speech corpus with full-length child–caregiver conversations, covering 498 children aged 2–8 and 112.5 hours of speech across standard Mandarin and ten additional dialect varieties. The dataset is carefully collected, anonymized, and released for non-commercial research, with utterance-level transcriptions and speaker-independent splits. ChildTalk represents a significant step toward comprehensive, publicly accessible Mandarin child speech resources. Its combination of large scale, broad age coverage,

full-dialogue recordings, and rich dialectal diversity establishes a solid foundation for advancing speech recognition, speaker and dialect modeling, and language acquisition research for young children, with potential impact on educational technology, healthcare, and human–computer interaction.

Limitations

Although *ChildTalk* covers a broad range of dialects and age groups, it is not perfectly balanced. First, the dialect distribution is skewed toward standard Mandarin and Southwestern Mandarin, while several other dialects (e.g., Southwest Min, Jiang–Huai, Jin Chinese) have substantially fewer speakers and shorter durations. This reflects the practical difficulty of recruiting child participants from some dialect communities, but it also means that results may be more reliable for high-resource varieties than for low-resource ones. Second, the age distribution is uneven: older children (7–8 years) are more frequently represented than very young children (2–3 years), for whom data collection is considerably more challenging. Future extensions of the corpus could aim to increase the coverage of underrepresented dialects and younger age ranges to further improve the robustness and representativeness of child ASR evaluations.

Ethical Considerations

All recordings in *ChildTalk* were collected with prior informed consent from parents or legal guardians. Before any recording took place, caregivers were informed about the purpose of the project, the recording setup, the types of data to be collected, and the intended use of the corpus for academic research. Each participating family received a compensation of 300 RMB in recognition of their time and effort, and guardians were free to decline or discontinue participation at any time.

To safeguard privacy, all audio and transcriptions are anonymized before release. Personal names and other directly identifying information are removed or replaced with codes, and we do not distribute fine-grained metadata that could make individual participants easily identifiable. Any demographic attributes reported in this paper or in the released metadata are provided only in aggregated form. Access to *ChildTalk* is restricted to non-commercial research use. Prospective users will be asked to agree to a data use agreement that specifies acceptable use of the corpus. In particular, the agreement will require that:

- Users treat removal requests from participants or their guardians as binding, and delete the corresponding recordings and transcripts from their local copies of the dataset within a reasonable time.
- Users do not attempt to infer the real-world identity of any child, caregiver, or institution, nor link the corpus with external resources for re-identification purposes.
- Users obtain approval from their own institutional ethics boards (e.g., IRB or equivalent), where applicable, and refrain from applications that could harm participants, such as surveillance, profiling, or discriminatory decision making.
- Users do not redistribute the original audio, transcripts, or derived subsets that could expose individual recordings outside their research group, unless they have received explicit permission from the dataset maintainers.

These measures are intended to balance the research value of *ChildTalk* with a strong commitment to the privacy and welfare of participating children and their families.

Acknowledgments

This work has been supported by the National Key R&D Program of China (Grant No.2022ZD0116307) and NSF China (Grant No.62271270).

References

Keyu An, Qian Chen, Chong Deng, Zhihao Du, Changfeng Gao, Zhifu Gao, Yue Gu, Ting He, Hangrui Hu, Kai Hu, Shengpeng Ji, Yabin Li, Zerui Li, Heng Lu, Xiang Lv, Bin Ma, Ziyang Ma, Chongjia Ni, Changhe Song, and 13 others. 2024. Funaudiollm: Voice understanding and generation foundation models for natural interaction between humans and llms. *ArXiv*, abs/2407.04051.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460.

Linda Bell, Johan Boye, Joakim Gustafson, Mattias Heldner, Anders Lindström, and Mats Wirén. 2005. The swedish nice corpus–spoken dialogues between

children and embodied characters in a computer game scenario. In *Interspeech 2005-Eurospeech, 9th European Conference on Speech Communication and Technology, Lisbon, Portugal, September 4-8, 2005*, pages 2765–2768. ISCA.

Nancy F Chen, Rong Tong, Darren Wee, Pei Xuan Lee, Bin Ma, and Haizhou Li. 2016. Singakids-mandarin: Speech corpus of singaporean children speaking mandarin chinese. In *Interspeech*, pages 1545–1549.

Catia Cucchiarini, Joris Driesen, H Van Hamme, and EP Sanders. 2008. Recording speech of children, non-natives and elderly people for hlt applications: the jasmin-cgn corpus.

Katherine Demuth and Annie Tremblay. 2008. Prosodically-conditioned variability in children's production of french determiners. *Journal of child language*, 35(1):99–127.

Linhao Dong, Shuang Xu, and Bo Xu. 2018. Speech-transformer: a no-recurrence sequence-to-sequence model for speech recognition. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5884–5888. IEEE.

Ruchao Fan, Natarajan Balaji Shankar, and Abeer Alwan. 2024. Benchmarking children's asr with supervised and self-supervised speech foundation models. In *Interspeech 2024*, pages 5173–5177.

Jun Gao, Aijun Li, and Ziyu Xiong. 2012. Mandarin multimedia child speech corpus: Cass_child. In *2012 International Conference on Speech Database and Assessments*, pages 7–12. IEEE.

Marta Garrote and A Moreno Sandoval. 2008. Chiede, a spontaneous child language corpus of spanish. In *Proceedings of the 3rd International LABLITA Workshop in Corpus Linguistics*.

Matteo Gerosa, Diego Giuliani, Shrikanth Narayanan, and Alexandros Potamianos. 2009. A review of asr technologies for children's speech. In *Proceedings of the 2nd Workshop on Child, Computer and Interaction*, pages 1–8.

Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, and 1 others. 2020. Conformer: Convolution-augmented transformer for speech recognition. *arXiv preprint arXiv:2005.08100*.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM transactions on audio, speech, and language processing*, 29:3451–3460.

Abe Kazemzadeh, Hong You, Markus Iseli, Barbara Jones, Xiaodong Cui, Margaret Heritage, Patti Price, Elaine Andersen, Shrikanth S Narayanan, and Abeer Alwan. 2005. Tball data collection: the making of a young children's speech corpus. In *Interspeech*, pages 1581–1584.

James Kennedy, Séverin Lemaignan, Caroline Montassier, Pauline Lavalade, Bahar Irfan, Fotios Papadopoulos, Emmanuel Senft, and Tony Belpaeme. 2017. Child speech recognition in human-robot interaction: evaluations and recommendations. In *Proceedings of the 2017 ACM/IEEE international conference on human-robot interaction*, pages 82–90.

Sungbok Lee, Alexandros Potamianos, and Shrikanth Narayanan. 1997. Analysis of children's speech: Duration, pitch and formants. In *Fifth European Conference on Speech Communication and Technology*.

W. S. Leung. 2018. A corpus-based study of the acquisition of post-verbal particles by cantonese-speaking children aged 2;5-5;8. Thesis, The University of Hong Kong, Pokfulam, Hong Kong SAR.

Z. Mai, M. Shang, J. Liu, S. Yan, S. Matthews, and V. Yip. 2024. Acquiring chinese in us, hong kong and beijing: three new corpora and three verbal structures. Paper presented at the XVIth International Congress for the Study of Child Language (IASCL-2024). July 15–19.

OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, and Aidan Clark et al. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

Yifan Peng, Siddharth Dalmia, Ian Lane, and Shinji Watanabe. 2022. Branchformer: Parallel mlp-attention architectures to capture local and global context for speech recognition and understanding. In *International Conference on Machine Learning*, pages 17627–17643. PMLR.

Humberto Pérez-Espinosa, Juan Martínez-Miranda, Ismael Espinosa-Curiel, Josefina Rodríguez-Jacobo, Luis Villaseñor-Pineda, and Himer Avila-George. 2020. Iesc-child: an interactive emotional children's speech corpus. *Computer Speech & Language*, 59:55–74.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023a. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023b. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR.

Kodali Radha and Mohan Bansal. 2022. Audio augmentation for non-native children's speech recognition through discriminative learning. *Entropy*, 24(10):1490.

Xian Shi, Xiong Wang, Zhifang Guo, Yongqi Wang, Pei Zhang, Xinyu Zhang, Zishan Guo, Hongkun Hao, Yu Xi, Baosong Yang, and 1 others. 2026. Qwen3-asr technical report. *arXiv preprint arXiv:2601.21337*.

Deng Xiangjun and Virginia Yip. 2017. A multimedia corpus of child mandarin: The tong corpus. *Journal of Chinese Linguistics*.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, and 1 others. 2025a. Qwen2. 5-omni technical report. *arXiv preprint arXiv:2503.20215*.

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, Yuanjun Lv, Yongqi Wang, Dake Guo, He Wang, Linhan Ma, Pei Zhang, Xinyu Zhang, Hongkun Hao, Zishan Guo, and 19 others. 2025b. Qwen3-omni technical report. *arXiv preprint arXiv:2509.17765*.

Zhuoyuan Yao, Di Wu 0061, Xiong Wang, Binbin Zhang, Fan Yu, Chao Yang, Zhendong Peng, Xiaoyu Chen, Lei Xie, and Xin Lei. 2021. Wenet: Production oriented streaming and non-streaming end-to-end speech recognition toolkit. In *interspeech*, volume 2021, pages 4054–4058.

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

Binbin Zhang, Hang Lv, Pengcheng Guo, Qijie Shao, Chao Yang, Lei Xie, Xin Xu, Hui Bu, Xiaoyu Chen, Chenchen Zeng, and 1 others. 2022. 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)*, pages 6182–6186. IEEE.

Jiaming Zhou, Shiyao Wang, Shiwan Zhao, Jiabei He, Haoqin Sun, Hui Wang, Cheng Liu, Aobo Kong, Yujie Guo, Xi Yang, Yequan Wang, Yonghua Lin, and Yong Qin. 2025. 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)*, pages 12524–12537, Vienna, Austria. Association for Computational Linguistics.

A Data details

Figure A.1 shows the geographic distribution of speakers in *ChildTalk* by province, with child and adult counts plotted separately. Most participants come from a handful of provinces in Mainland China: Sichuan and Shandong contribute the largest numbers of speakers, followed by Liaoning, Henan, Hebei, and Ningxia. Beyond these major contributors, the corpus also includes smaller but non-negligible numbers of speakers from a broad range of provinces and municipalities (e.g., Jiangsu, Fujian, Beijing, Guangdong, Shanghai, Yunnan), yielding a long-tail distribution. The child and adult bars are generally aligned within each province, indicating that caregiver recruitment tracks the regional distribution of children and that *ChildTalk* covers a geographically diverse set of dialect regions.

B Experiment Setup

All open-source models and datasets utilized in this study are distributed under open-source licenses, in compliance with the terms of their respective original distributions. For closed-source commercial models (e.g., GPT-4o), we accessed their speech transcription functionality via official APIs. All experiments are conducted on four NVIDIA RTX 3090, NVIDIA RTX 4090 and NVIDIA A100 GPUs.

B.1 Models

Training from scratch. We consider three end-to-end ASR architectures, implemented using the open-source Wenet (Yao et al., 2021) toolkit:

- **Transformer** (Dong et al., 2018), a sequence-to-sequence baseline with stacked self-attention and feed-forward layers.
- **Conformer** (Gulati et al., 2020), which augments self-attention with convolutions to capture both global and local patterns.
- **Branchformer** (Peng et al., 2022), which employs a branched attention–convolution design to balance long-range dependencies and local feature extraction.

Fine-tuning pre-trained models. To assess the benefits of large-scale pre-training, we further experiment with the following pre-trained models:

- **Whisper** [2] (Radford et al., 2023b), a widely used open-source end-to-end ASR model developed by OpenAI. We employ Whisper models from the `tiny` to `medium` configurations.
- **Conformer-WenetSpeech (CW)**[3], a Conformer model trained on the 10,000-hour

[2] https://github.com/openai/whisper

[3] https://github.com/wenet-e2e/wenet/blob/main/docs/pretrained_models.md

Table A.1: Demographic Information of Annotators.

Dialect	Gender	Age	Education	Native place
Jiao-Liao Mandarin	Male	31	Junior college	Qingdao
	Male	26	Bachelor's degree	Weifang
	Male	23	Bachelor's degree	Binzhou
	Male	37	Junior college	Weifang
Southern Min	Male	35	Bachelor's degree	Shantou
	Male	29	Bachelor's degree	Shantou
Beijing Mandarin	Female	37	Bachelor's degree	Chengde
	Female	37	Bachelor's degree	Chengde
Northeastern Mandarin	Female	35	Bachelor's degree	Shenyang
	Female	31	Bachelor's degree	Fushun
	Male	33	Bachelor's degree	Fushun
	Female	39	Junior college	Shenyang
	Male	31	Junior college	Tonghua
Southwestern Mandarin	Female	29	Bachelor's degree	Chengdu
	Female	35	Junior college	Chengdu
	Female	32	Junior college	Chongqing
	Female	30	Junior college	Mianyang
	Female	30	Junior college	Mianyang
	Female	35	Junior college	Chongqing
	Female	32	Junior college	Chengdu
	Female	31	Bachelor's degree	Yibin
	Male	29	Bachelor's degree	Guang'an
	Female	33	Bachelor's degree	Chongqing
Ji-Lu Mandarin	Female	40	Junior college	Baoding
	Male	32	Bachelor's degree	Baoding
	Male	35	Junior college	Baoding
	Female	29	Junior college	Shijiazhuang
Jin	Female	37	Junior college	Datong
	Female	35	Junior college	Taiyuan
	Male	30	Junior college	Taiyuan
Jiang-Huai Mandarin	Male	28	Bachelor's degree	Changzhou
	Female	34	Junior college	Wuxi
	Male	38	Junior college	Suzhou
	Female	32	Junior college	Shanghai
	Female	34	Junior college	Hefei
Lan-Yin Mandarin	Male	31	Bachelor's degree	Zhongwei
	Female	27	Bachelor's degree	Yinchuan
	Female	32	Junior college	Baiyin
	Female	25	Junior college	Lanzhou
Zhongyuan Mandarin	Female	38	Junior college	Zhumadian
	Female	33	Junior college	Zhumadian
	Female	48	Junior college	Xuchang
	Male	30	Bachelor's degree	Zhoukou
Mandarin	Female	43	Vocational secondary school	Datong
	Female	32	Bachelor's degree	Chengdu
	Female	42	Junior college	Zhengzhou
	Female	36	Vocational secondary school	Ili
	Female	45	High school	Qinhuangdao
	Male	47	High school	Qinhuangdao
	Female	37	Bachelor's degree	Qinhuangdao
	Female	32	Bachelor's degree	Hangzhou

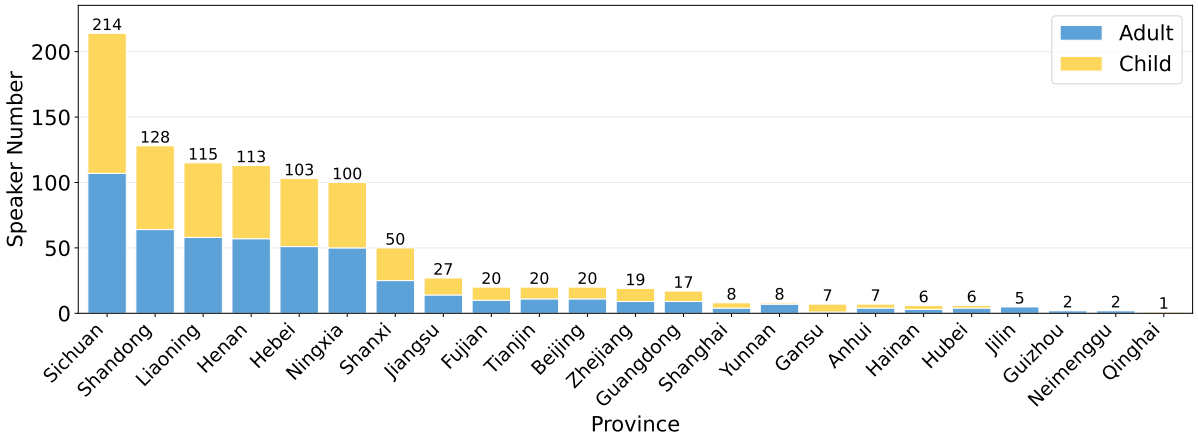


Figure A.1: Geographic distribution of child participants in ChildTalk.

WenetSpeech (Zhang et al., 2022) corpus with joint CTC–attention training.

- **SenseVoice-small** [4] (An et al., 2024), a speech foundation model pre-trained on multiple speech understanding tasks. This model supports multiple Chinese dialect ASR tasks and has demonstrated strong performance.

Off-the-shelf ASR: Qwen2.5-Omni [5] (Xu et al., 2025a) and Qwen3-Omni [6] (Xu et al., 2025b) are end-to-end omni-modal large language models from Alibaba's Qwen series that support unified speech, vision, and text understanding and generation, including high-quality speech transcription and spoken interaction. Moreover, we also include zero-shot API-based systems in our evaluation, namely OpenAI's GPT-4o [7] (OpenAI et al., 2024) transcription endpoints, including `gpt-4o-transcribe` and `gpt-4o-mini-transcribe`.

B.2 Hyperparameters

For models trained from scratch, we use a batch size of 32, a learning rate of 1×10^{-3}, and 100 training epochs for both the Transformer and Conformer architectures, while Branchformer is trained with a smaller batch size of 16 under the same learning rate and number of epochs. For fine-tuning the pre-trained Conformer-WenetSpeech model, we adopt a batch size of 16, a learning rate of 4×10^{-5}, and 20 training epochs. For Whisper, we use a batch size of 16, a learning rate of 1×10^{-5}, and 10 training epochs. For SenseVoice-small, the batch size is defined in terms of total audio duration (120 seconds per batch), and the model is fine-tuned for 10 epochs with a learning rate of 4×10^{-5}.

Model	Batch size	Learning rate	#Epochs
Transformer	32	1×10^{-3}	100
Conformer	32	1×10^{-3}	100
Branchformer	16	1×10^{-3}	100
CW	16	4×10^{-5}	20
Whisper	16	1×10^{-5}	10
SenseVoice-small	120 s/audio†	4×10^{-5}	10

Table B.1: Training hyperparameters for models trained from scratch and fine-tuned (FT) on *ChildTalk*. †Batch size for SenseVoice-small is measured as the total audio duration per batch.

C Context-Length Experiments on Qwen-3-ASR

To further validate the generality of the context-length effect, we evaluate Qwen-3-ASR models (0.6B and 1.7B) under different dialogue context lengths.

Context	0	1	2	3
Qwen3-ASR (0.6B)	19.47	16.89	16.29	16.26
Qwen3-ASR (1.7B)	17.01	14.66	14.34	14.07

Table B.1: WER performance with different dialogue context length on Qwen-3-ASR models.

The results show consistent improvements when introducing short context from C=0 to C=3, which aligns with the observations in Table 9.

[4] https://github.com/FunAudioLLM/SenseVoice
[5] https://github.com/QwenLM/Qwen2.5-Omni
[6] https://github.com/QwenLM/Qwen3-Omni
[7] https://platform.openai.com/docs/guides/speech-to-text

D AI Usage

In this work, AI tools are used exclusively for linguistic refinement of the manuscript, including optimizing the clarity of arguments, improving logical coherence, and correcting grammatical errors. The authors assume complete responsibility for the scientific content, conclusions, and overall integrity of the paper, confirming compliance with academic ethics and the absence of plagiarism.