ChildMandarin: A Comprehensive Mandarin Speech Dataset for Young Children Aged 3-5

Jiaming Zhou[1], Shiyao Wang[1], Shiwan Zhao[1], Jiabei He[1], Haoqin Sun[1], Hui Wang[1], Cheng Liu[1], Aobo Kong[1], Yujie Guo[1], Xi Yang[2], Yequan Wang[2], Yonghua Lin[2], Yong Qin[1]*

[1] College of Computer Science, Nankai University,
[2] Beijing Academy of Artificial Intelligence, Beijing, China,
Correspondence: zhoujiaming@mail.nankai.edu.cn, qinyong@nankai.edu.cn

Abstract

Automatic speech recognition (ASR) systems have advanced significantly with models like Whisper, Conformer, and self-supervised frameworks such as Wav2vec 2.0 and HuBERT. However, developing robust ASR models for young children's speech remains challenging due to differences in pronunciation, tone, and pace compared to adult speech. In this paper, we introduce a new Mandarin speech dataset focused on children aged 3 to 5, addressing the scarcity of resources in this area. The dataset comprises 41.25 hours of speech with carefully crafted manual transcriptions, collected from 397 speakers across various provinces in China, with balanced gender representation. We provide a comprehensive analysis of speaker demographics, speech duration distribution and geographic coverage. Additionally, we evaluate ASR performance on models trained from scratch, such as Conformer, as well as fine-tuned pre-trained models like HuBERT and Whisper, where fine-tuning demonstrates significant performance improvements. Furthermore, we assess speaker verification (SV) on our dataset, showing that, despite the challenges posed by the unique vocal characteristics of young children, the dataset effectively supports both ASR and SV tasks. This dataset is a valuable contribution to Mandarin child speech research. The dataset is now open-source and freely available for all academic purposes on https://github.com/flageval-baai/ChildMandarin.

1 Introduction

Automatic Speech Recognition (ASR) technology has become increasingly prevalent across various applications, ranging from virtual assistants and educational tools to accessibility services for individuals with disabilities (Kennedy et al., 2017). In particular, child speech recognition holds great potential in educational settings, such as language learning applications, reading tutors, and interactive systems. However, despite the rapid advancements in ASR technology, the performance of most systems—whether state-of-the-art or commercial—remains suboptimal when applied to children's speech (Fan et al., 2024).

ASR systems are predominantly trained on adult speech (Zhou et al., 2024), making them highly effective for everyday interactions but ill-suited for children due to physiological differences in vocal tract development, higher pitch, and inconsistent pronunciation (Lee et al., 1997; Gerosa et al., 2009). Children's speech also exhibits considerable variability in articulation, speech patterns, and vocabulary, further complicating the recognition process (Benzeghiba et al., 2007; Bhardwaj et al., 2022). These challenges are compounded by the lack of sufficient child-specific training data, which is crucial for developing ASR systems that can accurately and reliably understand children's speech across different age groups. However, datasets focused on young children are extremely rare (Graave et al., 2024). Most existing speech datasets either concentrate on adult speakers or cover older children, overlooking the unique linguistic and developmental characteristics of younger children (Zhou et al., 2023). This gap is critical, as the scarcity of training data limits the ability of ASR systems to perform well on speech from 3-5 age group.

Although there are a few open-source Mandarin speech datasets for children (Xiangjun and Yip, 2017; Gao et al., 2012; Yu et al., 2021; Chen et al., 2016), they are often limited in scope. For instance, the Tong Corpus (Xiangjun and Yip, 2017) records the speech of a single child from ages 1;7 to 3;4 (i.e., 1 year and 7 months to 3 years and 4 months), which is useful for certain research areas, but insufficient for ASR development due to the lack of speaker diversity. Similarly, while the

*Yong Qin is the corresponding author.

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 12524–12537

Corpus	Age range	# Speakers	Dur. (hrs)	Style	Year	Trans.	Avail.
Tong Corpus	1;7-3;4	1	22	Interactions	2018	Y	Y
CASS CHILD	1-4	23	631	Spontaneous speech	2012	P	N
SLT-CSRC C1	7-11	927	28.6	Reading	2021	Y	N
SLT-CSRC C2	4-11	54	29.5	Conversation	2021	Y	N
SingaKids	7-12	255	75	Reading	2016	Y	Y
Ours	3-5	397	41.3	Conversation	2024	Y	Y

Table 1: Summary of Chinese child speech datasets: age range, speaker count, duration, and availability. Dur.: duration. Trans.: transcriptions (P: partial). Avail.: availability.

CASS CHILD corpus (Gao et al., 2012) includes data from 23 children aged 1 to 4 years, a portion of 80 hours is transcribed, it is not publicly available, restricting its use in ASR research. Children's speech poses unique challenges, with frequent mispronunciations, ungrammatical expressions, and child-specific vocabulary. To address these issues, it is essential to collect data from a large number of speakers, ensuring substantial amounts of data per speaker to capture linguistic variability and improve the generalization of ASR models. Existing datasets, such as the SingaKids-Mandarin (Chen et al., 2016) and SLT-CSRC (Yu et al., 2021), primarily focus on older children (aged 7-12), leaving a gap for younger age groups.

Constructing a dedicated speech dataset for young children is crucial. It addresses a significant gap in existing resources and provides a foundation for developing ASR systems specifically tailored to young children. In this paper, we introduce a Mandarin speech dataset designed for children aged 3 to 5, comprising 41.25 hours of speech from 397 speakers across 22 of China's 34 provincial-level administrative divisions. Our evaluations of ASR models and speaker verification (SV) tasks demonstrate substantial improvements, underscoring the dataset's effectiveness in advancing technology for children's speech. This dataset bridges the gap in age-specific speech data by incorporating a wide range of speakers and extensive regional diversity. It represents a valuable contribution to Mandarin child speech research and holds significant potential for applications in educational technology and child-computer interaction.

2 Related Work

2.1 Child Speech Recognition Corpora in Mandarin Chinese

Publicly available child speech corpora for Mandarin Chinese are highly limited, particularly for younger age groups, as shown in Table 1. The few existing datasets are either too small in terms of speakers or lack accessibility, which restricts their utility for developing robust ASR systems.

The Tong Corpus (Xiangjun and Yip, 2017) is a longitudinal dataset that records the speech of a single child, Tong, with one hour of recordings per week from ages 1;7 to 3;4. Although this corpus is valuable for research on language acquisition, its use in ASR development is limited by its single-speaker nature, which cannot provide the diversity needed for model generalization.

Gao et al. (2012) collected the CASS CHILD dataset, which contains 631 hours of speech from 23 children aged 1 to 4 years. However, only about 80 hours of this dataset are labeled with transcriptions, and, critically, the dataset is not publicly accessible. This restricts its use in ASR experiments and highlights the difficulty of obtaining child speech corpora in Mandarin.

The SingaKids-Mandarin Corpus (Chen et al., 2016) contains 75 hours of speech data from 255 children aged 7 to 12, which is suitable for ASR training. This corpus encompasses diverse linguistic contexts. However, it focuses exclusively on children aged 7 to 12 and does not address the speech of younger children, which represents a significant gap in Mandarin ASR research.

Another important dataset is SLT-CSRC (Yu et al., 2021), which consists of two collections: SLT-CSRC C1 and C2. The former includes 28.6 hours of reading-style speech from 927 children aged 7 to 11, while the latter consists of 29.5 hours of conversational speech from 54 children aged 4 to 11. Although these datasets provide valuable speech data for Mandarin ASR, they were only available for participants of the SLT 2021 challenge and are no longer publicly accessible.

In summary, for Mandarin child speech, only the Tong Corpus and SingaKids-Mandarin datasets are available upon request, and both are limited in terms of speaker diversity and age range cover-

Corpus	Language	Age range	# Speakers	Dur.(hrs)	Year
Providence Corpus (Demuth et al., 2006)	English	1-3	6	363	2006
Lyon Corpus (Demuth and Tremblay, 2008)	English	1-3	4	185	2008
TBALL (Kazemzadeh et al., 2005)	English	K - G4	256	40	2005
CU Children's Read and Prompted Speech Corpus (Hagen et al., 2003)	English	K - G5	663	-	2003
CSLU Kids' Speech Corpus (Shobaki et al., 2007)	English	K-G10	1,100	-	2007
CU Story Corpus (Hagen et al., 2003)	English	G3-G5	106	40	2003
MyST Corpus (Pradhan et al., 2024)	English	G3-G5	1,371	393	2024
PF-STAR Children's Speech Corpus (Batliner et al., 2005)	English	4-14	158	14.5	2005
The CMU Kids Corpus (Eskenazi et al., 1997)	English	6-11	76	-	1997
TIDIGITS (Leonard and Doddington, 1993)	English	6-15	101	-	1993
CID children's speech corpus (Lee et al., 1999)	English	5-18	436	-	1999
Speechocean762 (Zhang et al., 2021)	English	5-18	125	6	2021
Non-Native children's speech corpus (Radha and Bansal, 2022)	English	7-12	20	3.3	2022
Demuth Sesotho Corpus (Demuth, 1992)	Sesotho	2-4	59	98	1992
CHIEDE (Garrote and Moreno Sandoval, 2008)	Spanish	3-6	59	∼8	2008
IESC-Child (Pérez-Espinosa et al., 2020)	Spanish	6-11	174	∼35	2020
JASMIN-CGN Corpus (Cucchiarini et al., 2008)	Dutch	7-16	-	∼64	2008
SANACS (Kruyt et al., 2024)	Slovak	6-12	67	∼15	2024
CFSC (Pascual and Guevara, 2012)	Filipino	6-11	57	∼8	2012
Swedish NICE Corpus (Bell et al., 2005)	Swedish	8-15	5,580	∼6	2005

Table 2: Summary of child speech datasets in other languages, where K denotes kindergarten while G denotes grade.

age. This lack of publicly accessible child speech corpora, particularly for younger children, continues to be a significant challenge in Mandarin ASR development.

2.2 Child Speech Corpora in Other Languages

In other languages, especially English, a wider variety of child speech corpora exists, as shown in Table 2. These corpora differ significantly in size, age range, and speaker diversity, reflecting various research priorities. However, many still lack sufficient coverage for younger children, a crucial age group for advancing ASR development.

English corpora, in particular, are among the most well-represented. For example, the Providence (Demuth et al., 2006) and Lyon Corpora (Demuth and Tremblay, 2008) focus on early childhood speech (ages 1-3), offering 363 and 185 hours of recordings, respectively. Despite their extensive durations, these datasets are limited in the number of speakers, with only 6 and 4 children represented, respectively. On the other hand, larger datasets such as the MyST Corpus (Pradhan et al., 2024) offer 393 hours of conversational speech from virtual tutoring sessions in elementary school science, collected from 1,371 children in grades 3 to 5. This broader speaker diversity is highly advantageous for training robust ASR systems.

Other notable English datasets include the CSLU Kids' Speech Corpus (Shobaki et al., 2007), which features reading recordings from over 1,100 children from kindergarten through grade 10 including simple words,digits and sentences, and the TBALL Corpus (Kazemzadeh et al., 2005), which contains speech from 256 children in kindergarten through grade 4. These datasets contribute valuable resources for developing ASR systems for various childhood age ranges and linguistic styles.

Child speech datasets in other languages are less common and typically smaller. For example, the Demuth Sesotho Corpus (Demuth, 1992) offers 98 hours of speech from 59 children aged 2 to 4, focusing on a non-Indo-European language, while the CHIEDE corpus (Garrote and Moreno Sandoval, 2008) contains around 8 hours of speech from 59 Spanish-speaking children aged 3 to 6. The IESC-Child Corpus (Pérez-Espinosa et al., 2020) provides about 35 hours of Spanish speech from 174 children aged 6 to 11.

For European languages, the JASMIN-CGN Corpus (Cucchiarini et al., 2008) offers 64 hours of Dutch speech from children aged 7 to 16, and the Swedish NICE Corpus (Bell et al., 2005) features data from 5,580 children aged 8 to 15. Although the NICE Corpus stands out for its large number of speakers, the total duration of recordings is relatively short, and similar limitations regarding younger children persist across these corpora.

Split	# Spk.	# Utt.	Dur. (hrs)	Avg. (s)
Train	317	32,658	33.35	3.68
Dev	39	4,057	3.78	3.35
Test	41	4,198	4.12	3.53
Sum	397	40,913	41.25	3.52

Table 3: Summary of dataset splits, including the number of speakers (# Spk.) and utterances (# Utt.), total duration (Dur.), and average utterance length (Avg.).

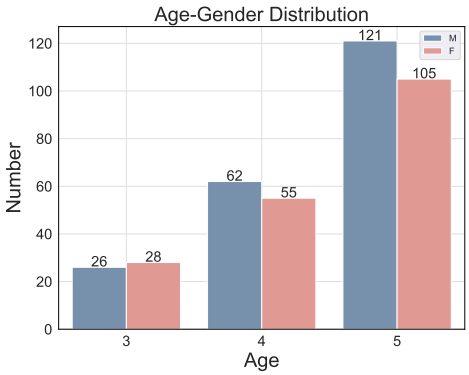

Figure 1: Distribution of speakers by age and gender in our dataset

3 Dataset Description

3.1 Dataset Collection

The dataset consists of 41.25 hours of speech data with carefully crafted character-level manual transcriptions, collected from Mandarin-speaking children aged 3 to 5 years. The gender distribution is balanced across all age groups. To ensure geographic coverage, speakers were selected from different regions of China. A total of 397 speakers participated, representing 22 out of 34 provincial-level administrative divisions. Accents were classified into three categories: heavy (H), moderate (M), and light (L).

Prior to data collection, informed consent was obtained from the parents or legal guardians of all participants. The consent process included detailed explanations of the study's purpose, procedures, and the intended use of the data for academic research. Parents were explicitly informed about their right to withdraw consent at any time without any repercussions.

Our data collection occurred in a conversational context to promote natural interaction, with parents present throughout the sessions to provide emotional comfort and support for the children. The recording content was unrestricted, focusing on age-appropriate daily communication, ensuring that children engaged in familiar and non-stressful activities. Common themes included hobbies, favorite activities, sports, cartoons, recent daily events and so on. Conversations were designed to elicit spontaneous and natural speech patterns from the children. Parents were encouraged to interact with their children using open-ended questions and prompts to sustain dialogue, while avoiding sensitive topics such as violence, politics, or private information. Importantly, only the children's speech was retained during segmentation, and any parental speech was excluded from the dataset. This ensured that the dataset solely captured the linguistic and acoustic characteristics of the target age group.

All recordings followed standardized collection and annotation protocols. Speech samples were captured using smartphones, with a nearly even split between Android (216) and iPhone (181) devices. The recordings were in WAV PCM format, with a 16kHz sampling rate and 16-bit precision, ensuring high-quality audio without clipping or volume inconsistencies. Silence segments of approximately 0.3 seconds were preserved at the beginning and end of each valid speech segment. Segmentation was performed manually based on semantics and pauses. Segments with fewer than 3 characters were excluded due to their lack of sufficient phonetic and linguistic context for effective acoustic modeling. This ensures a dataset with richer, more informative speech samples, beneficial for developing robust ASR models.

3.2 Annotation

To protect participant privacy, all recordings were manually reviewed and any containing identifiable information were removed. The released dataset is fully anonymized, with all metadata and content that could enable re-identification excluded.

Character-level manual annotations were performed by professional transcribers, who meticulously adhered to the audio content, including stutters, disfluencies, and developmental speech patterns. Regional pronunciation variations were transcribed faithfully. Additionally, numbers were transcribed as pronounced, maintaining consistency with the intended meaning of the speech. Further details about the annotators and annotation guidelines are provided in Appendix A.2 and Appendix A.3, respectively.

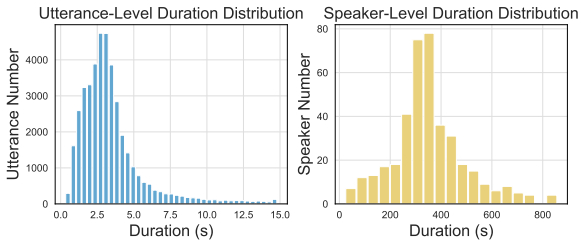

Figure 2: Utterance-level and speaker-level duration distribution in our dataset

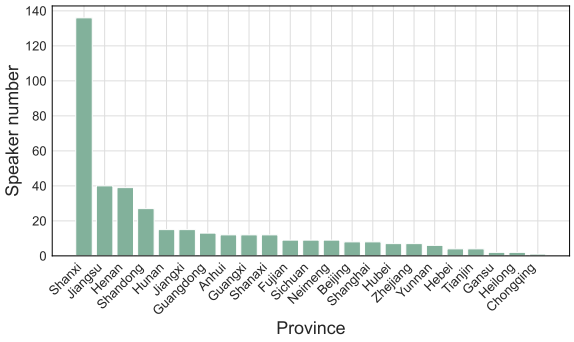

Figure 3: Geographic distribution of speakers in our dataset

3.3 Statistics

As shown in Table 3, our dataset consists of three subsets: training (317 speakers), validation (39 speakers), and test (41 speakers), with no overlap between speakers across the subsets. We further analyze the distribution of speakers based on age, gender, birthplace, accent and recording device.

The age and gender distribution in the dataset, depicted in Figure 1, highlights a decrease in the number of speakers as age decreases, which reflects the challenges in recruiting younger participants. Despite this, the gender distribution remains balanced across all age groups. More details about speaker distribution are provided in Appendix A.1

The distribution of utterance lengths and total speaking duration per speaker is presented in Figure 2. Most utterances are between 1 and 5 seconds long, with very few exceeding 10 seconds. Additionally, the majority of speakers have a total speaking duration between 200 and 600 seconds, which is essential for developing ASR systems tailored to young children.

The geographic distribution of speakers, spanning 22 of China's 34 provincial-level administrative divisions, is summarized in Figure 3. Despite recruitment challenges, broad regional representation was achieved, with Shanxi contributing

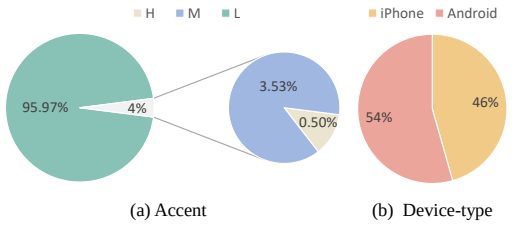

Figure 4: Proportions of accents and recording devices in our dataset

the highest number of participants (136), followed by Jiangsu (40) and Henan (39). Provinces such as Shaanxi, Shandong, and Hunan also contribute significantly. Although some regions, including Gansu, Heilongjiang, and Chongqing, have fewer participants, their inclusion enhances the dataset's comprehensive geographic coverage.

Speaker accents and recording devices are analyzed in Figure 4. Accents are categorized into three levels: heavy (H), moderate (M), and light (L), with the majority of speakers exhibiting light accent variation. Only around 4% of speakers are categorized as having moderate or heavy accents. Furthermore, a balanced representation of iPhone and Android devices was achieved to support diverse ASR system requirements.

4 Tasks and Baselines

In this section, we evaluate our dataset on both ASR and SV tasks.

4.1 Speech Recognition

For child speech recognition, we trained several baseline models from scratch and fine-tuned pre-trained models to assess performance on our dataset. We use the Character Error Rate (CER, %) as the evaluation metric. Refer to the Appendix B for the complete hyperparameter configurations.

4.1.1 Baselines Trained from Scratch

We utilize the open-source Wenet toolkit (Yao et al., 2021) to train ASR models from scratch. Three architectures are chosen: Transformer (Vaswani, 2017), Conformer (Gulati et al., 2020), and Paraformer (Gao et al., 2022). These models incorporate different approaches, including Connectionist Temporal Classification (CTC) (Graves et al., 2006), RNN-Transducer (RNN-T) (Graves, 2012), and Attention-based encoder-decoder (AED) (Chorowski et al., 2014; Chan et al., 2015).

Encoder	Loss	# Params	Decoding method			
			Greedy	Beam	Attention	Attention rescoring
Transformer	CTC AED	29M	34.55	34.4	40.61	32.15
Conformer	CTC AED	31M	**28.73**	**28.72**	**31.60**	**27.38**
Conformer	RNN-T AED	45M	37.11	37.14	33.84	37.14
Paraformer	Paraformer	30M	31.86	28.94	-	-

Table 4: Decoding performance (CER, %) of Transformer, Conformer, and Paraformer models trained from scratch

Model	Architecture	Input	# Params	Sup./Self-sup.	Training Data (hours)
Wav2vec 2.0 (B)	Enc	Waveform	97M	Self-sup.	10K
Wav2vec 2.0 (L)	Enc	Waveform	318M	Self-sup.	10K
HuBERT (B)	Enc	Waveform	97M	Self-sup.	10K
HuBERT (L)	Enc	Waveform	319M	Self-sup.	10K
CW	Enc-Dec	Fbank	122M	Sup.	10K
Whisper	Enc-Dec	Waveform	39M-1,550M	Sup.	680K

Table 5: Details of pre-trained baseline models. Enc and Dec stand for encoder and decoder, while Sup. and Self-sup. represent supervised and self-supervised learning. (B) and (L) denote the base and large versions.

The following models are considered:

- **Transformer**: We trained the widely-used Transformer model with joint CTC/AED training. The training process follows the recipe and configuration provided by Wenet.

- **Conformer**: The Conformer (Gulati et al., 2020) integrates convolutions with self-attention for ASR. We trained two models using both CTC and RNN-T loss functions respectively, following the Wenet recipe.

- **Paraformer**: Proposed by Gao et al. (Gao et al., 2022), Paraformer is a fast and accurate parallel transformer model.

4.1.2 Results of Training Models from Scratch

Table 4 presents the results of models trained from scratch on our dataset, evaluated using various decoding methods provided by Wenet (Yao et al., 2021). For Transformer and Conformer models with joint CTC and AED training (Kim et al., 2017), we report CTC greedy and beam search decoding results. For Conformer models with RNN-T and attention loss, we include RNN-T greedy and beam search decoding results. All beam searches use a beam size of 10. Attention decoding and attention rescoring decoding results are also reported for Transformer and Conformer.

Conformer with CTC-AED performs best overall, achieving the lowest CER of 27.38% with attention rescoring. Its CTC greedy and beam search methods yield nearly identical results (28.73% and 28.72%). In contrast, the Transformer model performs worse, with its best result being 32.15% CER from attention rescoring, while Paraformer achieves competitive results, particularly with beam search (28.94%). RNN-T for Conformer performs less effectively, with no significant improvement from attention rescoring. Overall, Conformer with CTC-AED provides the most reliable performance, especially with attention rescoring.

4.1.3 Pre-trained Baselines

We evaluate our dataset using a range of pre-trained baselines, including both supervised and self-supervised models. The details of these baselines are summarized in Table 5. For the self-supervised models, we utilize Wav2vec 2.0 (Baevski et al., 2020) and HuBERT (Hsu et al., 2021), integrating a CTC decoder with the encoder to perform the ASR task. For the supervised baselines, we include Conformer pre-trained on WenetSpeech (Zhang et al., 2022) and Whisper (Radford et al., 2023).

- **Wav2vec 2.0**: Wav2vec 2.0 (Baevski et al., 2020) is a self-supervised model which jointly captures discrete speech units and contextualized features. We select two versions of Wav2vec 2.0 pre-trained using WenetSpeech.[1]

- **HuBERT** : HuBERT (Hsu et al., 2021) is a self-supervised model that uses k-means clustering to generate target labels and applies

[1] https://huggingface.co/TencentGameMate/chinese-wav2vec2-base and https://huggingface.co/TencentGameMate/chinese-wav2vec2-large

Model	Greedy search	Beam search
Wav2vec 2.0 (B)	20.29	20.29
Wav2vec 2.0 (L)	21.12	21.12
HuBERT (B)	18.74	18.74
HuBERT (L)	**14.97**	**14.97**

Table 6: CER (%) of self-supervised pre-trained baselines with greedy and beam search decoding

Model	# Params	Zero-shot	Fine-tuning
CW	122M	**18.05**	**13.66**
Whisper-tiny	39M	67.63	28.78
Whisper-base	74M	51.49	23.33
Whisper-small	244M	37.99	17.45
Whisper-medium	769M	28.55	18.97
Whisper-large-v2	1,550M	29.43	-

Table 7: CER (%) of supervised pre-trained baselines in zero-shot and fine-tuned settings

BERT-like prediction loss over masked audio regions to learn contextualized representations. We select two versions of HuBERT pre-trained using WenetSpeech.[2]

- **Conformer-WenetSpeech (CW)**: CW is a 122M-parameter Conformer CTC-AED model, trained with supervised learning on the WenetSpeech dataset. Checkpoints are available in Wenet's open-source repository.[3]
- **Whisper**: Whisper (Radford et al., 2023) is a Transformer-based multilingual ASR model trained on 680,000 hours of labeled speech data by OpenAI. We include various versions of Whisper, ranging from tiny to large, with model sizes from 39M to 1550M.[4]

4.1.4 Results of Fine-tuning Pre-trained Models

Table 6 shows the CER for fine-tuning various self-supervised pre-trained models, including Wav2vec 2.0 and HuBERT, using both greedy and beam search decoding methods. HuBERT consistently outperforms Wav2vec 2.0, which is consistent with recent research (Yang et al., 2021). Additionally, HuBERT (L) demonstrates better performance compared to its smaller counterpart, HuBERT (B).

[2] https://huggingface.co/TencentGameMate/chinese-wav2vec2-large and https://huggingface.co/TencentGameMate/chinese-hubert-large

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

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

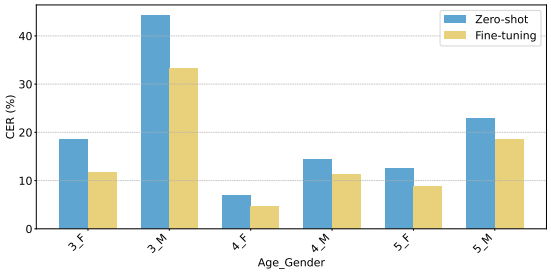

Figure 5: CER (%) comparison of zero-shot and fine-tuning methods using CW model across different age-gender groups

Method	Zero-shot			Fine-tuning		
Age_Gender	S (%)	D (%)	I (%)	S (%)	D (%)	I (%)
3_F	13.87	3.63	1.11	9.03	2.04	0.69
3_M	34.78	5.45	3.97	26.80	4.35	2.11
4_F	6.02	0.57	0.30	3.94	0.53	0.15
4_M	11.55	1.80	1.14	8.92	1.46	0.91
5_F	9.25	2.36	0.95	7.02	1.31	0.45
5_M	16.70	4.36	1.81	14.32	3.04	1.23

Table 8: In-depth comparison of different error types (S: Substitutions, D: Deletions, I: Insertions) between zero-shot and fine-tuning methods using CW model across different age-gender groups

However, Wav2vec 2.0 (L) underperforms relative to Wav2vec 2.0 (B), likely due to overfitting, given the limited data size.

Table 7 presents the CER results for Conformer-WenetSpeech (CW) and Whisper models in zero-shot and fine-tuning settings. Fine-tuning results in substantial CER improvements for all supervised models. Despite Whisper's large parameter size and extensive training data, the limited size of our dataset causes Whisper-medium to perform slightly worse than Whisper-Small after fine-tuning. Overall, CW achieves the best performance in both zero-shot and fine-tuned settings, highlighting its robust ASR capabilities learned from WenetSpeech.

4.1.5 Performance Analysis

Figure 5 shows ASR performance across age and gender groups on the CW model. 3-year-olds exhibits higher error rates than 5-year-olds, reflecting greater variability in younger children's speech. Fine-tuning significantly reduces CER across all age groups, demonstrating its effectiveness in adapting models to children's speech. Male speakers consistently exhibit higher CER than female speakers of the same age. This disparity may arise from greater pitch and articulation variability in young male children.

We further investigate error types in Table 8. Substitutions dominate error types, followed by

Model	# Params	Dim	Dev (%)	PLDA		Cosine similarity	
				EER (%)	minDCF	EER (%)	minDCF
x-vector	4.2M	512	75.4	8.91	0.7198	25.92	0.9780
ECAPA-TDNN	20.8M	192	84.6	13.72	0.8697	27.77	0.9490
ResNet-TDNN	15.5M	256	91.9	9.57	0.6597	22.11	0.9044

Table 9: Results of fine-tuning baselines on the speaker verification task, where Dim indicates the dimension of the extracted embeddings and Dev represents the accuracy on the validation set.

deletions and insertions. Younger children, particularly 3-year-olds, exhibit higher substitution and deletion rates, reflecting challenges in speech recognition.

In summary, age and gender notably influence ASR performance, with younger and male speakers posing greater challenges. Fine-tuning mitigates these issues, highlighting the importance of targeted adaptation strategies. Detailed analysis of fine-tuning performance on specific utterances can be found in Appendix C.

4.2 Speaker Verification

In this section, we evaluate our dataset on the SV task. The evaluation is organized into three parts: dataset repartition, baselines, and results.

4.2.1 Dataset Repartition

For the speaker verification task, the training and validation sets were merged, resulting in a total of 356 speakers. This combined data was then split into new training and validation sets with a 9:1 ratio for each speaker, while the test set remained unchanged. Although the training and validation sets share speakers, their speech samples are distinct. Verification trials were generated entirely from the test set, consisting of 20,000 trials and 41 speakers, with positive and negative trials evenly distributed (50% each). The trials uniformly covered same-speaker pairs (spk_a, spk_a) and different-speaker pairs (spk_a, spk_b).

4.2.2 Speaker Verification Baselines

In this study, three popular speaker embedding extractors, pre-trained on VoxCeleb (Nagrani et al., 2017), were fine-tuned on our dataset: x-vector[5] (Snyder et al., 2018), ECAPA-TDNN[6] (Desplanques et al., 2020), and ResNet-TDNN[7] (Villalba et al., 2020). These models were implemented using the SpeechBrain (Ravanelli et al., 2021) toolkit and fine-tuned for 40 epochs. The embeddings extracted from the verification trials were then used to evaluate the models' performance on the SV task. Refer to the Appendix B for the complete hyperparameter configurations.

4.2.3 Results of Speaker Verification

For evaluation, two scoring methods were applied: Probabilistic Linear Discriminant Analysis (PLDA) (Prince and Elder, 2007) and Cosine Similarity. Performance was measured using two metrics: Equal Error Rate (EER) and Minimum Detection Cost Function (minDCF). EER is computed by finding the verification threshold where the false rejection and false acceptance rates (p_{miss} and p_{fa}) are equal, such that EER $= p_{fa} = p_{miss}$. The DCF is calculated using:

$$C_\delta = c_{miss} \cdot p_{miss} \cdot p_{target} + c_{fa} \cdot p_{fa} \cdot (1 - p_{target})$$

where c_{miss} is the cost of false rejection, c_{fa} is the cost of false acceptance, and p_{target} represents the probability that the target speaker appears in the verification set. In this case, $c_{miss} = c_{fa} = 1$ and $p_{target} = 10^{-2}$.

Table 9 summarizes the performance of the models on the dataset, with both PLDA and Cosine Similarity evaluated using EER and minDCF metrics. Two key insights emerge from the results: First, the dataset proves to be well-suited for speaker-related tasks, as indicated by the strong performance of the three fine-tuned baseline models. However, the underdeveloped vocal characteristics of young children present challenges, potentially masking gender-related features and other distinguishing attributes. Second, due to the relatively small size of the dataset, the larger ECAPA-TDNN model underperformed compared to ResNet and x-vector, likely due to overfitting.

[5] https://huggingface.co/speechbrain/spkrec-xvect-voxceleb

[6] https://huggingface.co/speechbrain/spkrec-ecapa-voxceleb

[7] https://huggingface.co/speechbrain/spkrec-resnet-voxceleb

5 Conclusion

In conclusion, this paper introduces a valuable Mandarin speech dataset specifically designed for young children aged 3 to 5, addressing a crucial gap in ASR resources for this age group. Comprising 41.25 hours of speech data from 397 speakers across diverse provinces in China, the dataset ensures balanced gender representation and board geographic coverage. Our evaluations of ASR models and speaker verification show significant improvements, highlighting the dataset's effectiveness in advancing children's speech technology. This work represents a significant contribution to Mandarin child speech research and holds great promise for applications in educational technology and child-computer interaction.

Limitations

Despite the dataset comprising 41.25 hours of speech data, it remains relatively small compared to adult speech datasets, which typically encompass much larger volumes. Additionally, while the dataset covers 22 provinces across China, the geographic distribution is not fully balanced, and expanding representation from underrepresented regions could improve diversity. Overfitting can occur when fine-tuning pre-trained models with a large number of parameters, particularly on smaller datasets. To address this, parameter-efficient fine-tuning methods like LoRA (Hu et al., 2022) could be explored to enhance model performance.

Ethics Statement

This study was conducted in accordance with strict ethical guidelines to ensure the safety, privacy, and well-being of all participants, particularly given the involvement of young children.

All speech recordings were collected with the informed consent of each participant's parent or legal guardian. The data collection was carried out in child-friendly environments, with a parent present during the entire recording session to provide comfort and supervision. Children participated in age-appropriate conversational tasks without pressure or sensitive prompts. Participants received appropriate compensation of 150 RMB (about $20 USD) for their time and involvement.

To safeguard participants' privacy, all audio recordings were manually reviewed. Any recordings that contained personally identifiable information (e.g., names, contact details) were excluded from the publicly released dataset. The final dataset has been fully anonymized and stripped of metadata or content that could lead to re-identification.

The dataset is distributed exclusively for academic and non-commercial research purposes, and all users must formally agree to a Terms of Access agreement. This agreement explicitly prohibits commercial use, participant re-identification, data redistribution, and unethical applications. In particular:

- If a participant or their guardian requests removal of their data, all dataset recipients are obligated to delete the affected recordings.
- Researchers are required to comply with relevant institutional ethical review protocols (e.g., IRB) and ensure that the dataset is not misused for surveillance, profiling, or other harmful applications.
- Derivative works must not be redistributed beyond the research group without explicit permission.

The dataset is released under a non-commercial research license (CC BY-SA-NC 4.0), and must be cited appropriately in any derived publications or presentations. The dataset is provided "as is" without warranty, and the maintainers reserve the right to revoke access in the event of policy violations.

By taking these measures, we aim to maximize the dataset's value to the research community while upholding the highest ethical standards for child data protection. These safeguards are essential to enabling responsible research on child speech while minimizing potential risks.

Acknowledgement

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

References

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.

Anton Batliner, Mats Blomberg, Shona D'Arcy, Daniel Elenius, Diego Giuliani, Matteo Gerosa, Christian

Hacker, Martin Russell, Stefan Steidl, and Michael Wong. 2005. The pf_star children's speech corpus. pages 2761–2764.

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.

Mohamed Benzeghiba, Renato De Mori, Olivier Deroo, Stephane Dupont, Teodora Erbes, Denis Jouvet, Luciano Fissore, Pietro Laface, Alfred Mertins, Christophe Ris, et al. 2007. Automatic speech recognition and speech variability: A review. *Speech communication*, 49(10-11):763–786.

Vivek Bhardwaj, Mohamed Tahar Ben Othman, Vinay Kukreja, Youcef Belkhier, Mohit Bajaj, B Srikanth Goud, Ateeq Ur Rehman, Muhammad Shafiq, and Habib Hamam. 2022. Automatic speech recognition (asr) systems for children: A systematic literature review. *Applied Sciences*, 12(9):4419.

William Chan, Navdeep Jaitly, Quoc V Le, and Oriol Vinyals. 2015. Listen, attend and spell. *arXiv preprint arXiv:1508.01211*.

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.

Jan Chorowski, Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. End-to-end continuous speech recognition using attention-based recurrent nn: First results. *arXiv preprint arXiv:1412.1602*.

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. 1992. Acquisition of sesotho. In *The Cross-Linguistic Study of Language Acquisition*, pages 557–638. Lawrence Erlbaum Associates.

Katherine Demuth, Jennifer Culbertson, and Jennifer Alter. 2006. Word-minimality, epenthesis and coda licensing in the early acquisition of english. *Language and speech*, 49(2):137–173.

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

Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck. 2020. Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification. *arXiv preprint arXiv:2005.07143*.

Maxine Eskenazi, Jack Mostow, and David Graff. 1997. The cmu kids corpus. *Linguistic Data Consortium*, 11.

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.

Zhifu Gao, ShiLiang Zhang, Ian McLoughlin, and Zhijie Yan. 2022. Paraformer: Fast and accurate parallel transformer for non-autoregressive end-to-end speech recognition. In *Interspeech 2022*, pages 2063–2067.

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.

Thomas Graave, Zhengyang Li, Timo Lohrenz, and Tim Fingscheidt. 2024. Mixed children/adult/childrenized fine-tuning for children's asr: How to reduce age mismatch and speaking style mismatch. In *Interspeech 2024*, pages 5188–5192.

Alex Graves. 2012. Sequence transduction with recurrent neural networks. *arXiv preprint arXiv:1211.3711*.

Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. 2006. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *ICML*, pages 369–376.

Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, and Ruoming Pang. 2020. Conformer: Convolution-augmented transformer for speech recognition. In *Interspeech 2020*, pages 5036–5040.

Andreas Hagen, Bryan Pellom, and Ronald Cole. 2003. Children's speech recognition with application to interactive books and tutors. In *2003 IEEE Workshop on Automatic Speech Recognition and Understanding (IEEE Cat. No. 03EX721)*, pages 186–191. IEEE.

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.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

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.

Suyoun Kim, Takaaki Hori, and Shinji Watanabe. 2017. Joint ctc-attention based end-to-end speech recognition using multi-task learning. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 4835–4839. IEEE.

Joanna Kruyt, Róbert Sabo, Katarína Polónyiová, Daniela Ostatníková, and Štefan Beňuš. 2024. The slovak autistic and non-autistic child speech corpus: Task-oriented child-adult interactions. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 16094–16099.

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*.

Sungbok Lee, Alexandros Potamianos, and Shrikanth Narayanan. 1999. Acoustics of children's speech: Developmental changes of temporal and spectral parameters. *The Journal of the Acoustical Society of America*, 105(3):1455–1468.

R. Gary Leonard and George Doddington. 1993. Tidigits ldc93s10. *Linguistic Data Consortium*.

Arsha Nagrani, Joon Son Chung, and Andrew Zisserman. 2017. Voxceleb: A large-scale speaker identification dataset. In *Interspeech 2017*, pages 2616–2620.

Ronald M Pascual and Rowena Cristina L Guevara. 2012. Developing a children's filipino speech corpus for application in automatic detection of reading miscues and disfluencies. In *TENCON 2012 IEEE Region 10 Conference*, pages 1–6. IEEE.

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.

Sameer Pradhan, Ronald Cole, and Wayne Ward. 2024. 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)*, pages 12040–12045.

Simon JD Prince and James H Elder. 2007. Probabilistic linear discriminant analysis for inferences about identity. In *2007 IEEE 11th international conference on computer vision*, pages 1–8. IEEE.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. 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.

Mirco Ravanelli, Titouan Parcollet, Peter Plantinga, Aku Rouhe, Samuele Cornell, Loren Lugosch, Cem Subakan, Nauman Dawalatabad, Abdelwahab Heba, Jianyuan Zhong, et al. 2021. Speechbrain: A general-purpose speech toolkit. *arXiv preprint arXiv:2106.04624*.

Khaldoun Shobaki, John-Paul Hosom, and Ronald Cole. 2007. Cslu: Kids' speech version 1.1. In *Linguistic Data Consortium*.

David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur. 2018. X-vectors: Robust dnn embeddings for speaker recognition. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5329–5333. IEEE.

A Vaswani. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*.

Jesús Villalba, Nanxin Chen, David Snyder, Daniel Garcia-Romero, Alan McCree, Gregory Sell, Jonas Borgstrom, Leibny Paola García-Perera, Fred Richardson, Réda Dehak, Pedro A. Torres-Carrasquillo, and Najim Dehak. 2020. State-of-the-art speaker recognition with neural network embeddings in nist sre18 and speakers in the wild evaluations. *Computer Speech & Language*, 60:101026.

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

Shuwen Yang, Po-Han Chi, Yung-Sung Chuang, Cheng-I Jeff Lai, Kushal Lakhotia, Yist Y. Lin, Andy T. Liu, Jiatong Shi, Xuankai Chang, Guan-Ting Lin, Tzu-Hsien Huang, Wei-Cheng Tseng, Ko tik Lee, Da-Rong Liu, Zili Huang, Shuyan Dong, Shang-Wen Li, Shinji Watanabe, Abdelrahman Mohamed, and Hung yi Lee. 2021. Superb: Speech processing universal performance benchmark. In *Interspeech 2021*, pages 1194–1198.

Zhuoyuan Yao, Di Wu, 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 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, et al. 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.

Junbo Zhang, Zhiwen Zhang, Yongqing Wang, Zhiyong Yan, Qiong Song, Yukai Huang, Ke Li, Daniel Povey, and Yujun Wang. 2021. speechocean762: An open-source non-native english speech corpus for pronunciation assessment. In *Interspeech 2021*, pages 3710–3714.

Jiaming Zhou, Shiwan Zhao, Ning Jiang, Guoqing Zhao, and Yong Qin. 2023. Madi: Inter-domain matching and intra-domain discrimination for cross-domain speech recognition. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Jiaming Zhou, Shiwan Zhao, Yaqi Liu, Wenjia Zeng, Yong Chen, and Yong Qin. 2024. knn-ctc: Enhancing asr via retrieval of ctc pseudo labels. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 11006–11010. IEEE.

A Dateset Details

A.1 Distribution of Speakers

The table A.1 below provides a detailed breakdown of age, gender, accent severity, and province distribution across the train, development (dev), and test sets.

Feature	Train	Dev	Test
Age			
5 years old	179	22	25
4 years old	96	11	10
3 years old	42	6	6
Gender			
Male (M)	167	19	23
Female (F)	150	20	18
Accent Severity			
Light (L)	305	37	39
Moderate (M)	11	1	2
Heavy (H)	1	1	0
Provinces	23	13	13

Table A.1: Distribution of speakers by age, gender, accent severity, and province in the training, development, and test sets.

A.2 Annotator Information

All annotators are professional transcribers for ASR, and each audio sample was checked at least three times to ensure annotation accuracy. Table A.2 below summarizes the annotators' qualifications.

ID	Age	Gender	Origin	Education
0001	28	Male	Ningxia	Bachelor's
0002	24	Female	Shanxi	Bachelor's
0003	24	Female	Hebei	Bachelor's
0004	28	Female	Hebei	Bachelor's
0005	25	Female	Zhejiang	Bachelor's
0006	22	Male	Jiangxi	Bachelor's
0007	21	Female	Jiangxi	Bachelor's
0008	21	Male	Beijing	Bachelor's
0009	21	Male	Hebei	Bachelor's
0010	21	Female	Shaanxi	Bachelor's
0011	25	Female	Fujian	Master's
0012	24	Female	Shanxi	Master's

Table A.2: Demographic information of selected annotators, including age, gender, origin, and education level.

A.3 Annotation Guidelines

The dataset was annotated at the character level by professional transcribers following the principles below:

- Natural disfluencies (e.g., repetitions, stutters) were preserved if intelligible.
- If the intended meaning was clear, mispronounced words were transcribed with the correct characters; unintelligible speech was discarded.
- Regional pronunciations were transcribed according to the intended standard characters without correction.
- Common articulation errors in young children were interpreted based on intended meaning.
- Transcribed to reflect semantic meaning; letters were uppercased ASCII.
- Full-width Chinese punctuation was used in accordance with writing norms.

B Experimental Configurations

This section provides detailed configurations and hyperparameters used for training and fine-tuning

Encoder	Decoder	Batch size	LR	Warmup	Epochs
Transformer	CTC+ATT	32	1.00E-03	2500	100
Conformer	CTC+ATT	32	1.00E-03	2500	100
Conformer	RNN-T+ATT	32	1.00E-03	2500	100
Paraformer	Paraformer	16	2.00E-03	2500	100

Table B.1: Hyperparameters for training ASR models from scratch.

Model	Batch size	Learning rate	Epochs
CW	16	4.00E-04	100
Wav2vec 2.0 (B)	10	3.00E-05	70
Wav2vec 2.0 (L)	10	1.00E-04	70
HuBERT (B)	10	5.00E-05	70
HuBERT (L)	10	5.00E-05	70
Whisper	16	1.00E-5 $\sim$ 1.00E-6	20

Table B.2: Hyperparameters for fine-tuning pre-trained ASR models.

Model	Batch size	LR Schedule	Init LR	Base LR	Epochs
ECAPA-TDNN	128	Cyclic	5.00E-03	1.00E-08	40
ResNet-TDNN	128	Cyclic	5.00E-03	1.00E-08	40
X-vector	128	Linear	5.00E-03	1.00E-04	40

Table B.3: Hyperparameters for training speaker verification models.

the ASR and speaker verification (SV) models discussed in the paper. All experiments were conducted using four GTX 3090 or GTX 4090 GPUs over several hours.

B.1 ASR Model Training from Scratch

The hyperparameters for training ASR models from scratch are summarized in Table B.1. These models were trained using the Wenet toolkit with the configurations shown below.

B.2 ASR Model Fine-tuning

Table B.2 presents the fine-tuning hyperparameters for pre-trained ASR models, including Wav2vec 2.0, HuBERT, Whisper, and Conformer-WenetSpeech. Fine-tuning was performed using the training subset of our dataset.

B.3 SV Model Training

Table B.3 provides the training configurations for speaker verification models, including ECAPA-TDNN, ResNet-TDNN, and X-vector. These models were trained and evaluated on our dataset for speaker verification tasks.

C Analysis of Fine-Tuning Performance on Specific Utterances

As presented in Figure C.1, the fine-tuning process significantly improved the ASR model's performance across various utterances, with a clear reduction in character error rate (CER). In general, fine-tuning allowed the model to adapt to specific child speech variations, addressing common issues such as phoneme substitutions and mispronunciations. Despite these improvements, some residual errors were still observed, particularly for more complex or longer utterances. Overall, the results demonstrate the effectiveness of fine-tuning for enhancing ASR performance on child speech, though further optimization is necessary to fully address all challenges.

Utterance: 184_5_F_L_NANJING_IPHONE_001_012
Ground truth: 兔 子 会 穿 上 很 很 大 的 毛 衣
Zero-shot: 兔 子 会 穿 上 很 大 的 毛 衣 CER: 9.09 % N=11 C=10 S=0 D=1 I=0
Fine-tuning: 兔 子 会 穿 上 很 很 大 的 毛 衣 CER: 0.00 % N=11 C=11 S=0 D=0 I=0

Utterance: 080_5_F_L_CHENGDU_Android_005
Ground truth: 我 跟 同 学 一 起 玩 橡 皮 泥
Zero-shot: 我 跟 同 学 一 起 玩 橡 皮 离 CER: 10.00 % N=10 C=9 S=1 D=0 I=0
Fine-tuning: 我 跟 同 学 一 起 玩 橡 皮 泥 CER: 0.00 % N=10 C=10 S=0 D=0 I=0

Utterance: 320_5_F_L_SHANXI_iPhone12_001_013
Ground truth: 因 为 怪 兽 都 是 男 生 没 有 女 生
Zero-shot: 一 位 怪 兽 都 是 男 生 没 有 女 神 CER: 25.00 % N=12 C=9 S=3 D=0 I=0
Fine-tuning: 因 为 怪 兽 都 是 男 生 没 有 女 生 CER: 0.00 % N=12 C=12 S=0 D=0 I=0

Utterance: 403_5_M_L_ZHENGZHOU_Android_005
Ground truth: 我 要 唱 好 多
Zero-shot: 我 要 差 好 多 CER: 20.00 % N=5 C=4 S=1 D=0 I=0
Fine-tuning: 我 要 唱 好 多 CER: 0.00 % N=5 C=5 S=0 D=0 I=0

Utterance: 235_3_M_M_CHIFENG_opporeno3pro_001_108
Ground truth: 锄 禾 日 当 午 汗 滴 禾 下 土 谁 知 盘 中 餐 粒 粒 皆 辛 苦
Zero-shot: 锄 禾 日 到 五 太 低 和 下 土 直 他 嘟 他 秘 密 接 心 CER: 75.00 % N=20 C=5 S=13 D=2 I=0
Fine-tuning: 锄 禾 日 当 午 汗 滴 禾 下 土 谁 - 盘 中 餐 粒 粒 皆 辛 - CER: 10.00 % N=20 C=18 S=0 D=2 I=0

Figure C.1: Performance comparison of zero-shot and fine-tuned models on specific utterances