

ChildVox: A Speech, Audio, and Large Audio-Language Model Benchmark in Understanding and Characterizing Sound across Childhood

Tiantian Feng¹ Anfeng Xu¹ Xuan Shi¹ Aditya Kommineni¹ Shakhrul Iman Siam³
Megan Micheletti⁴ Zhonghao Shi⁵ Helen Tager-Flusberg⁶ Mi Zhang³
Lynn K. Perry² Catherine Lord⁴ Daniel Messinger² Shrikanth Narayanan¹

¹University of Southern California ²University of Miami ³The Ohio State University
⁴University of California, Los Angeles ⁵Harvard University ⁶Boston University

Correspondence: tiantiaf@usc.edu

Abstract

We present ChildVox, a novel benchmark for characterizing the diverse acoustic signals through which children communicate. Specifically, ChildVox follows the full developmental trajectory from birth through school age, covering physiological sounds, non-linguistic vocalizations, canonical syllables, and spoken language. ChildVox integrates more than 20 sub-tasks across 17 child-centered audio and speech datasets, enabling systematic cross-corpus and cross-domain comparison. We evaluate a representative range of audio and speech foundation models, including self-supervised, ASR-oriented, and large audio-language models, on tasks including physiological sound classification, vocalization and canonical syllables modeling, and speech quality assessment and recognition. Benchmark results show that ChildVox provides a suite of high-performance models in recognizing a wide range of acoustic signals from children, supporting downstream applications such as characterizing children’s language levels and tracking speech production with age. Our benchmark models on public datasets are released at: <https://github.com/tiantiaf0627/childvox-release>.

1 Introduction

Deep learning and Artificial Intelligence for child-centered speech or voice processing and modeling have long focused on automatic speech recognition (ASR) (Fan et al., 2024; Ying et al., 2025; Singh et al., 2026), a general-purpose task for populations with established spoken language skills. However, children “express” themselves through a wide range of verbal and non-verbal skills at different stages of childhood (Li et al., 2026a), shaped by their physical (Namasivayam et al., 2020) and neuro-developmental status (Lord et al., 2018) and educational backgrounds (Lesaux and Siegel, 2003). Therefore, tasks like ASR cover only a subset of children’s communicative skills and life

experiences, and do not fully capture signals critical for understanding development, language level, health, and behaviors across childhood.

This limitation is amplified for children who have not yet fully acquired socially recognized spoken language, or who experience speech sound disorders (SSD) and other conditions that impact speech production. In early childhood, for example, communication is often dominated by non-spoken forms, including vocalized events (e.g., cries). Moreover, physiological sounds (e.g., cardiac sounds (Oliveira et al., 2021)) can carry critical information when verbal or non-verbal cues are insufficient for understanding a child’s condition.

In this work, we therefore rethink what constitutes speech, voice, or language in the context of child development. We introduce ChildVox, an audio benchmark that extends conventional ASR modeling to broader forms of children’s voices”, enabling the characterization of sounds across childhood experiences for child-centered applications. We broadly view “voice” in children as embodied communication that includes not only spoken language, but also vocalization, canonical syllables, and physiological sounds, all of which work together to convey meaning and express internal states. This definition calibrates communication skills to the developmental stage, recognizing that how children interact with the world evolves as their linguistic, motor, and cognitive skills develop.

Our proposed ChildVox benchmark offers unique contributions compared to prior works: (1) Unlike previous studies that primarily focus on children’s ASR, ChildVox considers the full developmental trajectory from birth through school age, covering physiological sounds, vocalizations, canonical syllables, and speech within a single unified benchmark. Here, we specifically refer to “unified” as the integration and evaluation of heterogeneous child-centered acoustic signals drawn from different datasets and developmental stages.

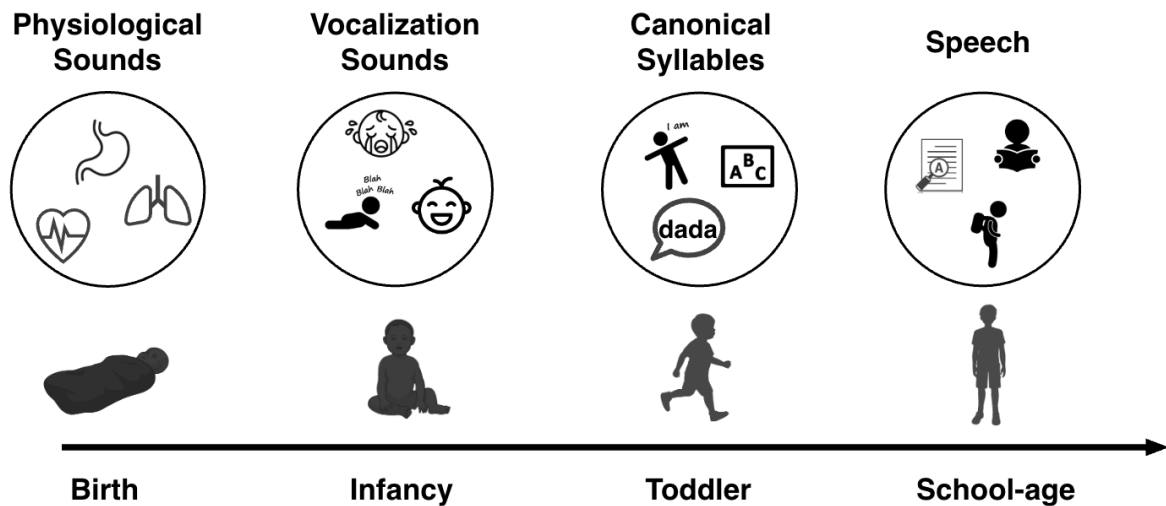


Figure 1: Overview of the wide range of ways that children “express” themselves. ChildVox benchmarks audio, speech, and large audio-language models on physiological sounds, vocalizations, canonical syllables, and speech.

(2) ChildVox integrates more than 20 sub-tasks across 17 child-centered audio and speech datasets into a consistent evaluation protocol. (3) We evaluate a representative range of self-supervised, ASR-oriented, and large audio-language models, showing that different pre-training setups offer complementary measures of child-centered acoustic signals. Moreover, models provided in ChildVox can outperform the frontier proprietary models like Gemini 3.5 Flash in zero-shot settings on the five selected tasks. Extensive experiments show that ChildVox provides a robust platform for supporting downstream applications in understanding children’s audio and spoken language across developmental contexts.

2 Related Works

2.1 Child ASR Benchmarks

Several recent studies have proposed benchmark-style evaluations for children’s ASR, where many of these works consistently show substantial degradation when adult-trained ASR systems are applied to children’s speech, due to developmental acoustic variability, pronunciation differences, and conversational dynamics. For example, [Fan et al. \(2024\)](#); [Ying et al. \(2025\)](#) evaluate supervised and self-supervised speech foundation models for child ASR, analyzing the effects of model scaling, training paradigms, and datasets. Other studies examine ASR robustness of speech from children with developmental delays, highlighting challenges from atypical vocal behaviors ([McGonigle et al., 2024](#)). Recently, causal analyses of children’s ASR errors further quantify the contributions of physiological,

cognitive, and environmental factors to recognition failures ([Singh et al., 2026](#)).

2.2 Child Speech Benchmarks beyond ASR

Apart from ASR, recent work has applied speech foundation models to a range of child-centered tasks, showing that self-supervised representations capture meaningful developmental information from children’s speech and infant vocalizations ([Li et al., 2024](#)). Other studies have explored speech foundation models for child-adult speaker diarization in naturalistic dyadic interactions ([Xu et al., 2024](#)), spoken language development analysis in children with ASD ([Xu et al., 2023a](#)), and phoneme recognition for language learning applications ([Medin et al., 2024](#)). Additionally, layer-wise analyses from a recent work further reveal that different model layers encode distinct developmental information ([Sinha et al., 2026](#)). Collectively, these studies show the growing potential of speech foundation models beyond ASR in children’s speech.

3 ChildVox Benchmark

Figure 1 presents the overview of benchmark design to characterize the range of ways that children “express” themselves from birth through school age.

3.1 Benchmark Overview

Unlike conventional ASR-centric benchmarks on children’s speech, ChildVox captures the diverse spectrum of child-produced and child-related acoustic signals that emerge throughout development: physiological sounds, such as cardiac sounds that provide markers of health conditions; vocaliza-

Table 1: Summary of the audio and speech data used in the ChildVox benchmark. The proposed benchmark includes more than 20 sub-tasks related to physiological sound classification, vocalization classification, canonical syllables classification, and speech recognition and classification.

Dataset	Evaluation Tasks	Labels	Data License
CirCor (Oliveira et al., 2021)	Murmur Detection	Absent, Present, Unknown	Open Database
ICBHI (Rocha et al., 2019)	Crackle Detection	Crackles, No crackles	Open Database
	Wheeze Detection	Wheezes, No wheezes	
	Respiratory Condition	Healthy, COPD, Other condition	
SPRSound (Zhang et al., 2022)	Respiratory Sound	Normal, CAS, DAS, CAS & DAS, Poor Quality	Not Specified
Donate-a-cry	Cry Classification	Hunger, Other	Open Database
What’s in a cry? (Lockhart-Bouron et al., 2023)	Cry Classification	Hunger, Loneliness, Discomfort	Not Specified
AudioSet (Gemmeke et al., 2017)	Child Sound Classification	Child speech, Babbling, Giggle, Children shouting, Baby laughter, Baby cry, Whimper, Child singing, Children playing, Music for children	CC-BY-4.0
ReCANVo (Johnson et al., 2023) (include adult audio data)	Affective Status	Delighted, Dysregulated, Frustrated, Request, Selftalk, Social	Not Specified
BabbleCor (Cychosz et al., 2021)	Vocal Development Classification	Crying, Laughing, Canonical, Non-Canonical, Junk	Customized License
SpeechMaturity (Hitczenko et al., 2025)	Vocal Development Classification	Crying, Laughing, Canonical, Non-Canonical, Junk	Customized License
C-BESD (Rao et al., 2024)	Emotion Recognition	Anger, Happy, Neutral, Sad	Not Specified
PERCEPT-R (Benway et al., 2022)	Pronunciation	Rhotic, Derhotic	PhonBank License
SpeechOcean762 (Zhang et al., 2021) (includes adult audio data)	Prosody	Poor, Nearly correct, Correct intonation	CC-BY-4.0
	Fluency	Little fluent, Fluent in general, Fluent	
UltraSuite (Eshky et al., 2018)	Speech Articulator	Poor or Understandable, Good, Excellent	CC-BY-NC-4.0
		Bilabial, Labiodental, Dental, Glottal Labiovelar, Alveolar, Palatal, Velar	
Natural Language Sampling (NLS) (Butler et al., 2022)	Diarization	Speaker labels and timestamps	Private
In-house ADOS2-Mod3	Speech Intelligibility	Intelligible, Unintelligible, Vocalization	
	Diarization	Speaker labels and timestamps	Private
MyST (Pradhan et al., 2024)	ASR	Speech Transcript	
	ASR	Speech Transcript	Customized License
TinyVox (Lavechin et al., 2026)	ASR	Phoneme Transcript	Not Specified

tions, such as crying in infancy; canonical syllables and emerging language behaviors in toddlers, such as proto-speech that reflect the progressive development of communicative skills; and the more complex social communicative behaviors of school-age children associated with spoken language interaction in everyday interactions. Collectively, this developmental framing across childhood positions the ChildVox as an ecologically grounded platform for advancing child-centered audio intelligence.

3.2 Benchmark Tasks and Datasets

ChildVox evaluates four broad categories of child-centered audio datasets: physiological sounds, vocalizations, canonical syllables, and speech. The information regarding evaluation tasks, labels, and data license is provided in Table 1. The details about each dataset are described in the Appendix A.

Physiological Sound Classification ChildVox includes three benchmark datasets for pediatric physiological sound classification covering both cardiac and respiratory assessment. The benchmark targets four tasks: murmur detection, crackle detection, wheeze detection, and respiratory condition classification. Murmur detection is evaluated on CirCor (Oliveira et al., 2021), a collection of 5,272 pediatric cardiac auscultation recordings, while crackle detection, wheeze detection, and respiratory condition classification are evaluated on ICBHI (Rocha et al., 2019) and SPRSound (Zhang et al., 2022), two respiratory sound datasets consisting of pediatric and general patient populations.

Vocalization Event Classification ChildVox targets two vocalization-related tasks: generic child sound and infant cry classification. Child sound classification covers ten child-related audios:

speech, babble, giggle, shout, laughter, cry, whimper, sing, play, and music for children. This is drawn from AudioSet (Gemmeke et al., 2017), a large-scale ontology-driven collection of YouTube data. Cry cause classification is evaluated on two infant cry corpora: Donate-a-Cry¹, where we classify hunger versus other, given its imbalanced distribution, and What’s in a cry? (Lockhart-Bouron et al., 2023) contains labels: hunger, loneliness, and discomfort.

Canonical Syllables Classification We experimented with ReCANVo (Johnson et al., 2023), a database of affective nonverbal sounds recorded in everyday interactions from minimally speaking children and adults with neurodevelopmental conditions. This corpus supports affective status classification (e.g., dysregulated). Although ReCANVo includes recordings from both children and adults, it remains a unique resource for studying affective nonverbal vocalizations in children. BabbleCor (Cychosz et al., 2021) and SpeechMaturity (Hitzenko et al., 2025; Zhang et al., 2025) target early vocal development in children, supporting vocal development classification across cry, laugh, canonical, non-canonical, and junk.

Speech Quality Assessment We include three datasets evaluating speech production, pronunciation, and articulation. PERCEPT-R (Benway et al., 2022) (with PhonBank (Rose and MacWhinney, 2014)) contains recordings from children with typical speech development and speech sound disorders, emphasizing rhotic production in American English. SpeechOcean762 (Zhang et al., 2021) provides 5,000 English utterances from 250 non-native speakers (half children), each annotated by five expert raters for pronunciation accuracy, fluency, and prosody. Similar to the ReCANVo dataset, although SpeechOcean762 includes half adult speech data, it remains one of the largest and well-annotated corpora to evaluate children’s speech quality. UltraSuite (Eshky et al., 2018) pairs synchronized ultrasound tongue imaging with acoustic recordings for research on child speech disorders. Inspired by Ribeiro et al. (2021), we use it for articulator classification across eight places of articulation (e.g., velar) on word-level audios.

Adult-child Speaker Diarization A core component of analyzing child speech in interaction is speaker diarization, which detects speech segments

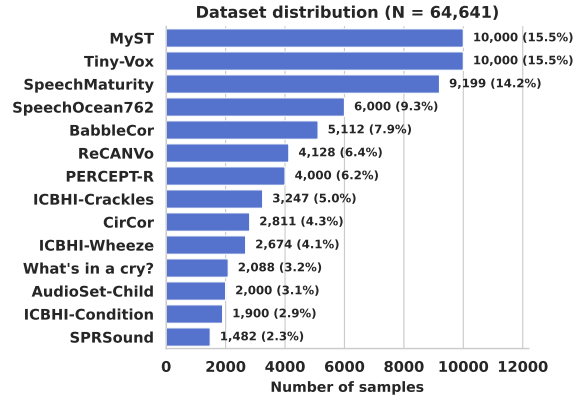


Figure 2: Training data distribution of each sub-dataset in ChildVox-Balanced dataset for training LALMs.

and labels “who spoke when”. Here, we include two in-house datasets to benchmark the speaker diarization. Specifically, the Natural Language Sampling (NLS) (Butler et al., 2022) dataset includes child-parent conversations collected from remote dyadic interactions involving minimally verbal children with ASD. The corpus includes 73 English-speaking child-parent sessions with detailed child/ adult speaker and speech intelligibility labels. Moreover, the ADOS2-Mod3 dataset (Xu et al., 2026) consists of recordings from more than 100 children participating in Autism Diagnostic Observation Schedule-Module3 (Lord et al., 2008) assessments. Within this dataset, interactions related to emotion and social-difficulty segments are annotated with speaker labels, which are kept for the fine-tuning experiment.

Automatic Speech Recognition We target both word-level and phoneme-level transcription of children’s speech. Word-level recognition is evaluated on MyST (Pradhan et al., 2024), a corpus of elementary school students interacting with a virtual science tutor, and on ADOS2-Mod3, whose transcripts are annotated with emotion and social-difficulty segments. Phoneme-level recognition is evaluated on TinyVox (Lavechin et al., 2026), which provides over half a million phoneme-level transcriptions across multiple languages. Here, we focus on the English subset with one item.

3.3 ChildVox-Balanced Dataset

Given the substantial aggregate scale of the combined datasets, we construct ChildVox-Balanced, a curated subset with labels balanced within each classification task, to support practical and reproducible supervised fine-tuning of Large Audio Language Models (LALMs) mentioned later. It draws

¹<https://github.com/gveres/donateacry-corpus>

Pre-trained Model	Pre-Training Objectives	Child-centered Audio Pre-Training
SSAST-Base	Self-supervised	×
voc2vec-HuBERT		×
wav2vec-LB		✓
BabyHuBERT		✓
WavLM Large		×
Whisper Base	VAD, ASR, LID	×
Whisper Small		×
Whisper Large V3		×
Qwen2-Audio Instruct	Multi-stage Uni-modal and Multi-modal Learning	×
AudioFlamingo 3		×

Table 2: Summary of audio and speech foundation models used in the ChildVox benchmark. We mark whether the pre-trained models are pre-trained with child-centered audio.

only on publicly accessible resources and limits training and testing samples per label to 2,000 and 50, respectively. For ASR evaluation, we include 10,000/500 utterances from MyST and TinyVox for training/testing. Figure 2 shows the resulting per-dataset distribution, yielding a total of 64,641 audio and speech recordings across 14 subtasks.

4 Benchmark Models

In ChildVox, we evaluate several representative audio, speech, and large audio-language models for benchmark experiments. The details of the experimented pre-trained models are listed in Table 2.

4.1 Encoder-based Models

Self-supervised Models We evaluate five representative self-supervised audio and speech foundation models: SSAST (Gong et al., 2022), voc2vec (Koudounas et al., 2025), WavLM (Chen et al., 2022), wav2vec-LB (Li et al., 2023), and BabyHuBERT (Charlot et al., 2025). SSAST, voc2vec, and WavLM apply self-supervised learning to learn representations from generic audio events, non-verbal vocalizations, and adult-centered speech data, respectively. These pre-trained models have shown strong performance on audio event detection, speech recognition, and speaker diarization tasks. Moreover, BabyHuBERT and wav2vec-LB are two pretrained speech models trained primarily on child-centered audio and speech data using HuBERT and wav2vec pre-training, respectively.

ASR Models We primarily evaluate models from the Whisper family, an encoder-decoder architecture trained on large-scale multilingual speech primarily for ASR. Apart from ASR, Whisper encoders have shown strong transferability across

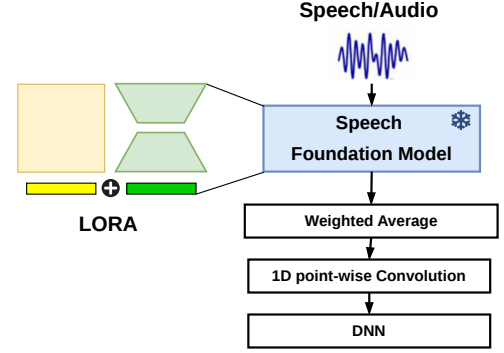


Figure 3: LoRA fine-tune in most encoder-based pre-trained models except SSAST, BabyHuBERT and wav2vec-LB.

speech processing problems, such as child-adult speaker diarization (Xu et al., 2024). To explore the effect of model scale, we evaluate encoder embeddings from three Whisper variants: Whisper-Base, Whisper-Small, and Whisper-Large-v3. Moreover, we include Parakeet-TDT (Rekesh et al., 2023; Xu et al., 2023b) for fine-tuned ASR tasks.

4.2 Large Audio-Language Model

We include two widely used Large Audio-Language Models (LALMs) in our benchmark: Qwen2-Audio-Instruct (Chu et al., 2024) and Audio Flamingo 3 (AF3) (Ghosh et al., 2025). Both models unify diverse audio understanding across speech, environmental sounds, and music. Specifically, Qwen2-Audio combines a Whisper-large-v3 audio encoder with the Qwen-7B language model backbone. Similar to Qwen2-Audio, the audio encoder of AF3 is initialized from Whisper-Large-v3.

5 Experiments

5.1 Data Preprocessing

For all datasets included in ChildVox, audio samples are resampled to 16 kHz to match the sampling rate of all pre-trained models. Across most experiments, we restrict a minimum input duration of 200 ms to maintain sufficient acoustic information for downstream evaluation. For most datasets without predefined training, development, and test partitions, we adopt a 5-fold cross-validation setup in the experiment. More details, such as the prompt design for the SFT experiments with LALM and data filtering, are in the Appendix A and C.

5.2 Modeling Details

The details of training settings, like learning rate and training epochs, are listed in the Appendix B.

Table 3: Benchmark performance across physiological sounds, vocalization, canonical syllables, and speech from children on each complete dataset. voc2vec-H, WavLM-L, Whisper-L are voc2vec-HuBERT, WavLM-Large, Whisper-Large, respectively. SO represents the SpeechOcean dataset. Additional results on Whisper-base and Whisper-small are reported in the Appendix D. We report the average Macro-F1 score (higher the better), and the best and second-best results per row are in **bold** and underlined, respectively.

Task Category	Dataset	SSAST	voc2vec-H	WavLM-L	wav2vec-LB	BabyHuBERT	Whisper-L
Physiological Sounds	CirCor	<u>0.636</u>	0.563	0.643	0.575	0.632	0.552
	ICBHI-crackles	<u>0.644</u>	0.626	0.612	0.625	0.654	0.586
	ICBHI-wheezes	0.638	0.592	0.614	0.613	<u>0.636</u>	0.609
	ICBHI-Condition	0.466	0.445	0.392	<u>0.472</u>	0.401	0.508
	SPRSound	0.448	0.368	<u>0.439</u>	0.189	0.359	0.414
Vocalization	AudioSet-Child	0.657	0.560	0.539	0.606	<u>0.632</u>	0.619
	What’s in a cry?	0.300	0.303	0.321	0.308	0.326	<u>0.323</u>
	donate-a-cry	0.473	<u>0.486</u>	0.468	0.424	0.430	<u>0.502</u>
Canonical Syllables	ReCANVo	<u>0.444</u>	0.416	0.423	0.425	0.446	0.359
	Speech Maturity	0.686	0.640	0.620	<u>0.696</u>	0.722	0.640
	BabbleCor	0.569	0.514	0.485	<u>0.605</u>	0.667	0.594
Speech	NLS-Intelligibility	0.588	0.548	0.603	0.602	<u>0.614</u>	0.661
	Percept-R	0.789	0.788	<u>0.813</u>	0.805	0.823	0.811
	Ultrasuite	0.850	0.635	<u>0.882</u>	0.768	0.818	0.888
	SO-Fluency	0.601	0.554	<u>0.613</u>	0.516	0.627	0.627
	SO-Accuracy	0.564	0.557	<u>0.625</u>	0.575	0.598	0.649
	SO-Prosody	0.613	0.613	<u>0.705</u>	0.585	0.669	0.715
	C-BESD	<u>0.753</u>	0.566	0.892	0.658	0.665	0.532

Model Fine-tuning Because the experimental pretrained models differ in architecture and implementation, we adopt model-specific fine-tuning strategies. For SSAST, we choose to fine-tune the model using full-model optimization, whereas the pretrained backbone weights of all other models remain frozen. For wav2vec-LB, we use representations from the final hidden layer for downstream fine-tuning. For BabyHuBERT, we learn a weighted combination of representations from all hidden layers for fine-tuning. For voc2vec-HuBERT, WavLM-Large, and the Whisper-family models, we adopt parameter-efficient fine-tuning through LoRA. Specifically, we apply LoRA with a rank of 64 to the feedforward layers for voc2vec, WavLM, and Whisper as shown in Figure 3. For LALMs, LoRA modules with rank 64 are inserted into the query, key, value, down-projection, and up-projection layers of the language model, and the feedforward layers of the audio encoder.

Evaluation We report utterance-level Macro-F1 on the test set for most classification tasks in the benchmark. For speaker diarization, we report diarization error metrics, while ASR and phoneme recognition are evaluated using word error rate (WER) and phoneme error rate (PER), respectively. Results for encoder-based models are reported as averages across multi-folds in Appendix A. Specifically, we use the default test set from datasets like

SpeechOcean762 and MyST.

6 Benchmark Results

6.1 Encoder-based Model Results

Table 3 presents the results of the encoder-based model on the complete ChildVox benchmark data.

Physiological Sound Classification Performance on physiological sound classification varies across models and datasets. For example, WavLM-Large achieves the best result on CirCor with a Macro F1 score of 0.643, while SSAST performs best on ICBHI-wheezes and SPRSound. BabyHuBERT achieves the highest score on ICBHI-crackles with a Macro F1 score of 0.654, whereas Whisper-Large performs best on ICBHI-Condition. Overall, models pretrained on general audio or self-supervised speech objectives perform strongly across most tasks in physiological sound classification. These results suggest that models learned with generic audio representation may better capture fine-grained spectral and temporal characteristics present in physiological signals. In contrast, Whisper-based models show comparatively lower performance on most physiological sound tasks except on ICBHI-Condition, indicating that speech-oriented pretraining may not provide benefits for physiological sound classification, particularly in children.

Table 4: Diarization and ASR performance. We report Diarization Error Rate (DER) and Word Error Rate (WER), where the lower is better. Best results per column in **bold**, second-best underlined.

Model	DER ($\downarrow$)		WER ($\downarrow$)	
	NLS	ADOS	MyST	ADOS
WavLM-Large	22.10	45.60	16.56	59.34
Whisper-Base	24.20	48.20	16.99	55.85
Whisper-Small	<u>18.70</u>	<u>44.00</u>	15.93	59.91
Whisper-Large	17.70	42.50	14.80	40.20
Parakeet-TDT	–	–	<u>15.82</u>	<u>45.60</u>

Vocalization and Canonical Syllables Similar to the physiological sound classification, for vocalization and canonical-syllable classification, no single model consistently dominates across datasets. SSAST achieves the best performance on AudioSet-Child and competitive performance in ReCANVo and Speech Maturity, while BabyHuBERT performs best on What’s in a cry?, ReCANVo, Speech Maturity, and BabbleCor. Whisper-Large achieves the highest score on donate-a-cry and remains competitive on several other tasks. These results suggest that general audio pretraining is particularly effective for broad child-related acoustic events, whereas pre-trained representations learned from child-relevant speech or sound may better capture early vocal and syllabic patterns. Nevertheless, cry-cause classification remains particularly challenging under the current benchmark setting, with relatively low performance across all evaluated models.

Speech Quality Assessment For children’s speech quality assessment, Whisper-Large achieves the best performance on NLS-Intelligibility, UltraSuite, and all three SpeechOcean762 sub-tasks, including fluency, accuracy, and prosody. On the other hand, BabyHuBERT performs best on Percept-R, while WavLM-Large substantially outperforms the other models on C-BESD for emotion classification. These findings suggest the benefits of large-scale multilingual ASR pretraining for predicting children’s speech intelligibility, articulation, and overall speech quality. At the same time, the strong results from BabyHuBERT and WavLM-Large indicate that self-supervised representations pre-trained from child-centered speech recordings remain valuable for tasks involving pronunciation and speech emotion. Overall, these results indicate that speech-based pre-training benefits the classification of children’s speech, despite such models being trained predominantly on adult data.

Speaker Diarization and Speech Recognition

Table 4 reports diarization and ASR performance across the NLS, ADOS2-Mod3, and MyST datasets. For speaker diarization, Whisper-Large achieves the lowest error rates on both NLS (17.70 DER) and ADOS2-Mod3 (42.50 DER). For ASR, Whisper-Large also obtains the best results, reaching 14.80 WER on MyST and 40.20 WER on ADOS, with Parakeet-TDT ranking second on both ASR benchmarks. Overall, model scale is the dominant factor, with the largest Whisper model leading on the speech recognition and diarization tasks.

Findings Overall, performance varies substantially across child-centered audio categories, highlighting the diverse acoustic properties of physiological sounds, vocalizations, canonical syllables, and speech. Meanwhile, no single model dominates across all tasks, whereas different encoders capture complementary aspects of these signals.

6.2 Large Audio Language Models Results

Table 5 compares the two LALMs against the encoder-based models on the ChildVox-Balanced subset, where two LALMs behave substantially differently in their performance. Qwen2-Audio is consistently competitive with the strongest encoder-based models, achieving the best scores on AudioSet and ReCANVo. It remains close to the best encoder-based models on SpeechMaturity and Percept-R, and obtains a low WER on MyST. In contrast, AudioFlamingo3 performs substantially worse on nearly every task, with particularly low scores on vocalization and canonical syllables classification, such as AudioSet and SpeechMaturity. It also returns an extremely large error rate on TinyVox phoneme recognition (0.958 PER).

Error Inspection Manual inspection indicates that this performance difference is driven by inconsistent instruction-following from the response, where AudioFlamingo3 often returns free-form descriptions (e.g., the audio clip is of a baby laughing”) instead of the requested label. It also tends to summarize or hallucinate on transcription tasks. For example, given the reference transcript “Well, without electricity, an electromagnet would never be possible, because it gives it the magnetism,” AF3 hallucinates a summary rather than a transcription, producing “The electromagnet is not working.” Moreover, we show the confusion matrices of Gemini 3.5 Flash and ChildVox Qwen2-Audio on PERCEPT-R and SpeechOcean in Appendix E.

Table 5: Benchmark performance on ChildVox-Balanced test set using encoder-based models and LALMs. We report PER and WER (lower the better) for TinyVox and MyST, respectively. We report the Macro-F1 score (higher the better) on the remaining sets. SO refers to the SpeechOcean762 dataset.

Model	CirCor (↑)	SPRSound (↑)	AudioSet (↑)	ReCANVo (↑)	Speech Maturity (↑)	PERCEPT-R (↑)	SO-Prosody (↑)	TinyVox (↓)	MyST (↓)
SSAST	0.543	0.508	0.638	0.500	0.735	0.800	0.644	-	-
BabyHuBERT	0.663	0.445	0.635	0.433	0.786	0.809	0.745	-	-
WavLM-Large	0.571	0.518	0.538	0.448	0.612	0.843	0.726	-	0.147
Whisper-Base	0.349	0.362	0.605	0.373	0.663	0.769	0.745	-	0.147
Whisper-Large	0.402	0.452	0.636	0.370	0.689	0.815	0.759	-	0.128
AudioFlamingo3	0.452	0.162	0.089	0.157	0.069	0.323	0.400	0.958	0.274
Qwen2-Audio	0.397	0.541	0.699	0.514	0.726	0.779	0.671	0.602	0.133

Table 6: Macro-F1 performance of Gemini 3.5 Flash across sampling temperatures. SO denotes the SpeechOcean dataset. The best results for each dataset are shown in **bold**.

Dataset	Temperature		
	0.1	0.5	1.0
CirCor	0.449	0.364	0.342
SPRSound	0.267	0.269	0.240
ReCANVo	0.224	0.192	0.221
PERCEPT-R	0.703	0.668	0.693
SO-Prosody	0.427	0.456	0.394

6.3 Comparison with Proprietary Baselines

We compare zero-shot Gemini 2.5 Flash and Gemini 3.5 Flash against our ChildVox-trained encoder (best one) and Qwen2-Audio on five public datasets in Figure 4. We limit this comparison to these five datasets to ensure that no restricted or license-protected data is transmitted to third-party APIs.

Sensitivity to Temperature We evaluate both proprietary models across multiple sampling temperatures and compare performance sensitivity to temperature of Gemini 3.5 Flash in Table 6. Specifically, Gemini 3.5 Flash shows sensitivity to the sampling temperature. A temperature of 0.1 provides the strongest overall performance, achieving the best results on CirCor, ReCANVo, and PERCEPT-R, with an average Macro-F1 of 0.414. A temperature of 0.5 performs best on SPRSound and SO-Prosody. At a temperature of 1.0, the average score decreases to 0.378. Overall, lower-temperature decoding is generally more reliable for Gemini 3.5 Flash, although the optimal temperature remains task-dependent. Therefore, we apply a temperature of 0.1 in the remaining analysis.

Comparing with ChildVox Overall, specialized models from ChildVox outperform both proprietary baselines in zero-shot settings on five selected task. Specifically, both Gemini models perform particularly poorly on CirCor, SPRSound, and Re-

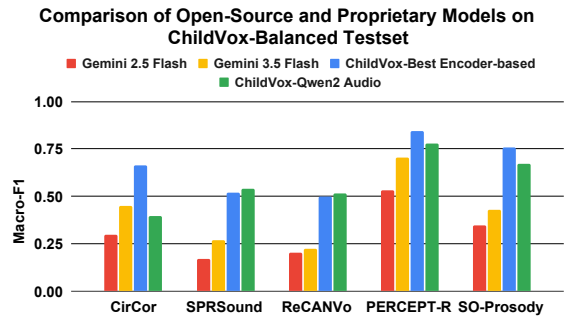


Figure 4: Macro-F1 on ChildVox-Balanced test set. Zero-shot proprietary models (Gemini 2.5/3.5 Flash) underperform the ChildVox -trained encoder and Qwen2-Audio baselines across all five datasets.

CANVo, suggesting that general-purpose LALMs struggle with fine-grained pediatric vocal and physiological audio. In summary, these results highlight the effectiveness of models trained on the ChildVox benchmark. Zero-shot Gemini settings are detailed in the Appendix E.

7 Exemplar Applications with ChildVox

Inspired by the previous work (Lavechin et al., 2026), we design two experiments using the benchmark models developed from the ChildVox to measure the child’s speech production and language development in two different applications.

7.1 Characterizing the Language Level

Application Setup Our first application is on the in-house NLS dataset described in Section 3. In this study, domain experts categorize each child’s spoken language level into pre-verbal communication, first words, and word combinations, following the definitions by Tager-Flusberg et al. (2009). Specifically, we refer to these three levels as LL-1, LL-2, and LL-3, respectively. We then apply our speaker diarization estimation models to compare language level differences across three groups.

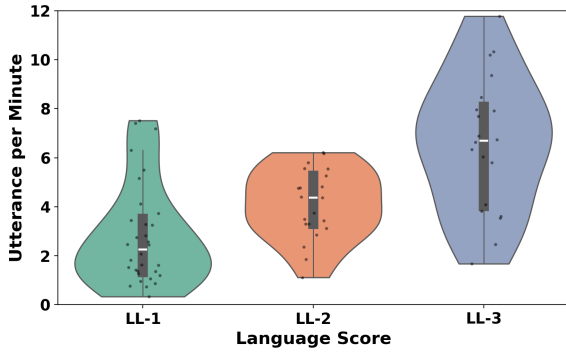


Figure 5: Utterance rate (utterances per minute) derived from the speaker diarization model by language level (LL-1 to LL-3) in the NLS study. The violin plot shows that the utterance rate increases with the language levels.

Observations Figure 5 reveals a clear monotonic relationship between language level and utterance rate: the median number of utterances per minute increases steadily from LL-1 (pre-verbal communication) through LL-2 (first words) to LL-3 (word combinations). This trend indicates that children at more advanced spoken language levels are more vocally active, producing speech at a higher rate. Moreover, the variability in utterance rate is increased with language level, with LL-3 showing the most varied distribution. Overall, these results show that diarization-derived utterance rate captures a meaningful developmental signal consistent with expert-assigned language levels.

7.2 Speech Production Development

Application Setup Our second application uses the PERCEPT-R corpus (Section 3), which contains single-word rhotic-targeted speech samples from children with typical development and SSDs, each manually annotated for correct rhotic production. We apply our Whisper-Large and BabyHuBERT Rhotic/Derhotic classification model to investigate how the predicted probability of rhotic production varies with chronological age. Specifically, we restrict the analysis only to typically developing children, using recordings labeled “PRE” in the CRESULSMOSAIC subset (Ayala et al., 2023), and average the predictions from Whisper-Large and BabyHuBERT as the final output.

Observations Figure 6 shows a moderate positive correlation between chronological age and the model-predicted probability of rhotic production in the CRESULSMOSAIC subset of the PERCEPT-R ($r = 0.573$). Specifically, younger children (~ 115 -135 months) show larger variability,

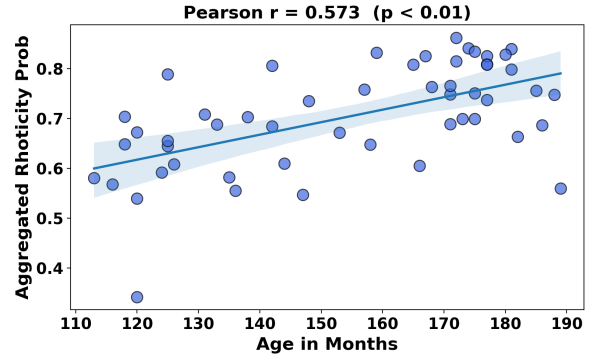


Figure 6: Correlation between model-predicted probability of correct rhotic production and chronic age in the CRESULSMOSAIC subset of the PERCEPT-R dataset. Each dot represents a single participant, where the model-predicted probability is aggregated from all reading samples of a participant.

ity, while predicted probabilities from older children centered more consistently around 0.75. Several lower-probability outliers remain present at older ages, highlighting continued inter-speaker variability. These results indicate the ChildVox provides models that can capture developmental changes in rhotic articulation across late childhood

8 Conclusion and Future Work

We introduce ChildVox, a benchmark characterizing the diverse acoustic signals through which children communicate from birth through school age, integrating more than 20 sub-tasks across 17 child-centered audio and speech datasets. We would highlight that the scope of the benchmark design is rarely presented in the literature. Evaluated on a representative range of speech and audio foundation models, ChildVox supports downstream applications such as characterizing children’s language levels and tracking speech production with age. In the future, we plan to expand it toward broader phoneme recognition (Bharadwaj et al., 2026; Guo et al., 2026), additional recording environments (Li et al., 2021; Feng et al., 2025b; Long et al., 2024), and to apply our benchmark models to enrich existing child audio corpora to support early screening (Islam et al., 2024), longitudinal monitoring (Kalenkovich et al., 2025), and education (Li et al., 2026b; Wagner et al., 2025) for children.

9 Acknowledgement

This work was supported by the Simons Foundation under Grant SFI-ARHUMAN-00004115-03 and Grant 655054.

10 Limitations

While ChildVox facilitates the systematic evaluation of audio, speech, and LALMs on child-centered audios, several limitations remain.

Language, Demographics, and Task Coverage

The majority of datasets in ChildVox include most English-language recordings in speech categories, with limited representation of other languages or dialects. Specifically, our ASR evaluation is restricted to the English subset. Therefore, conclusions about model performance may not generalize to children speaking other languages, such as Mandarin (Chua et al., 2023) or Spanish (Gildersleeve-Neumann et al., 2008). Similarly, the demographic factors (e.g., educational background or developmental status) across the experimental datasets are, in many cases, not fully documented, which may introduce sampling biases that propagate into model evaluation and result interpretation. Moreover, we would consider integrating more spoken language understanding tasks as described in Dynamic-superb (Huang et al., 2025b, 2024).

Annotation variability Many child-centered audio tasks in ChildVox, particularly affective vocalization classification (ReCANVo), cry-cause classification (Donate-a-Cry), and canonical syllables labeling (SpeechMaturity), involve inherently subjective categories with inter-rater disagreement. Therefore, reported classification scores may only reflect the ceiling imposed by annotation reliability.

Restricted set of foundation models Although ChildVox includes several representative self-supervised, ASR-oriented, and LALMs, the development of audio foundation models is expanding rapidly. Specifically, we did not evaluate every recent open-source LALMs, such as GAMA (Ghosh et al., 2024), SALMONN (Tang et al., 2024), Step-Audio (Huang et al., 2025a), and Kimi-Audio (Ding et al., 2025). Moreover, comparisons with proprietary baselines are limited to two Gemini Flash models in a zero-shot setting, and our zero-shot prompt approach did not cover different prompt-engineering techniques.

11 Ethical Considerations

Overall, all models released in this work are derived from publicly available sources, and ChildVox adheres to the scope of each dataset’s license. Moreover, modeling children’s speech can expose sen-

sitive information about children, and to mitigate misuse, we plan to release code and checkpoints under the Responsible AI License (RAIL). Users are expected to respect the privacy of data subjects and comply with applicable laws in their jurisdictions. We encourage the use of these models for research aimed at developing robust speech processing technology for and about children. Specifically, use of the models for clinical or diagnostic applications, surveillance, privacy-invasive applications, or commercial purposes is strictly prohibited.

References

- Samantha A Ayala, Amanda Eads, Heather Kabakoff, Michelle T Swartz, Douglas M Shiller, Jennifer Hill, Elaine R Hitchcock, Jonathan L Preston, and Tara McAllister. 2023. Auditory and somatosensory development for speech in later childhood. *Journal of Speech, Language, and Hearing Research*, 66(4):1252–1273.
- Nina Benway, Jonathan L Preston, Elaine Hitchcock, Asif Salekin, Harshit Sharma, and Tara McAllister Byun. 2022. Percept-r: An open-access american english child/clinical speech corpus specialized for the audio classification of /r/. In *INTERSPEECH*, pages 3648–3652.
- Shikhar Bharadwaj, Chin-Jou Li, Yoonjae Kim, Kwanghee Choi, Eunjung Yeo, Ryan Soh-Eun Shim, Hanyu Zhou, Brendon Boldt, Karen Rosero Jacome, Calvin Chang, et al. 2026. Prism: Benchmarking phone realization in speech models. *arXiv preprint arXiv:2601.14046*.
- Lindsay K Butler, Chelsea La Valle, Sophie Schwartz, Joseph B Palana, Cerelia Liu, Natalie Peterman, Lue Shen, and Helen Tager-Flusberg. 2022. Remote natural language sampling of parents and children with autism spectrum disorder: Role of activity and language level. *Frontiers in Communication*, 7:820564.
- Théo Charlot, Tarek Kunze, Maxime Poli, Alejandra Cristia, Emmanuel Dupoux, and Marvin Lavechin. 2025. Babyhubert: Multilingual self-supervised learning for segmenting speakers in child-centered long-form recordings. *arXiv preprint arXiv:2509.15001*.
- Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, et al. 2022. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518.
- Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, et al. 2024. Qwen2-audio technical report. *arXiv preprint arXiv:2407.10759*.

- Victoria YH Chua, Hexin Liu, Leibny Paola Garcia, Fei Ting Woon, Jinyi Wong, Xiangyu Zhang, Sanjeev Khudanpur, Andy WH Khong, Justin Dauwels, and Suzy J Styles. 2023. Merlion ccs challenge: A english-mandarin code-switching child-directed speech corpus for language identification and diarization. In *Proc. Interspeech 2023*, pages 4109–4113.
- Margaret Cychosz, Alejandrina Cristia, Erika Bergelson, Marisa Casillas, Gladys Baudet, Anne S Warlaumont, Camila Scaff, Lisa Yankowitz, and Amanda Seidl. 2021. Vocal development in a large-scale crosslinguistic corpus. *Developmental science*, 24(5):e13090.
- Ding Ding, Zeqian Ju, Yichong Leng, Songxiang Liu, Tong Liu, Zeyu Shang, Kai Shen, Wei Song, Xu Tan, Heyi Tang, et al. 2025. Kimi-audio technical report. *arXiv preprint arXiv:2504.18425*.
- Aciei Eshky, Manuel Sam Ribeiro, Joanne Cleland, Korin Richmond, Zoe Roxburgh, James M Scobbie, and Alan Wrench. 2018. Ultrasuite: A repository of ultrasound and acoustic data from child speech therapy sessions. In *Proc. Interspeech 2018*, pages 1888–1892.
- Ruchao Fan, Natarajan Balaji Shankar, and Abeer Alwan. 2024. Benchmarking children’s asr with supervised and self-supervised speech foundation models. *Interspeech*.
- Tiantian Feng, Jihwan Lee, Anfeng Xu, Yoonjeong Lee, Thanathai Lertpetchpun, Xuan Shi, Helin Wang, Thomas Thebaud, Laureano Moro-Velazquez, Dani Byrd, et al. 2025a. Vox-profile: A speech foundation model benchmark for characterizing diverse speaker and speech traits. *arXiv preprint arXiv:2505.14648*.
- Tiantian Feng and Shrikanth Narayanan. 2024. Foundation model assisted automatic speech emotion recognition: Transcribing, annotating, and augmenting. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 12116–12120. IEEE.
- Tiantian Feng, Anfeng Xu, Xuan Shi, Somer Bishop, and Shrikanth Narayanan. 2025b. Egocentric speaker classification in child-adult dyadic interactions: From sensing to computational modeling. In *Proc. Interspeech 2025*, pages 2835–2839.
- Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. 2017. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 776–780. IEEE.
- Sreyan Ghosh, Arushi Goel, Jaehyeon Kim, Sonal Kumar, Zhifeng Kong, Sang-gil Lee, Chao-Han Yang, Ramani Duraiswami, Dinesh Manocha, Rafael Valle, and Bryan Catanzaro. 2025. [Audio flamingo 3: Advancing audio intelligence with fully open large audio language models](#). In *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pages 41819–41886.
- Sreyan Ghosh, Sonal Kumar, Ashish Seth, Chandra Kiran Reddy Evuru, Utkarsh Tyagi, S Sakshi, Oriol Nieto, Ramani Duraiswami, and Dinesh Manocha. 2024. Gama: A large audio-language model with advanced audio understanding and complex reasoning abilities. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6288–6313.
- Christina E Gildersleeve-Neumann, Ellen S Kester, Barbara L Davis, and Elizabeth D Peña. 2008. English speech sound development in preschool-aged children from bilingual english–spanish environments. *Language, Speech, and Hearing Services in Schools*, 39(3):314–328.
- Yuan Gong, Cheng-I Lai, Yu-An Chung, and James Glass. 2022. Ssast: Self-supervised audio spectrogram transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 10699–10709.
- Chenxu Guo, Jiachen Lian, Yisi Liu, Baihe Huang, Shriyaa Narayanan, Cheol Jun Cho, and Gopala Anumanchipalli. 2026. Huper: A human-inspired framework for phonetic perception. *arXiv preprint arXiv:2602.01634*.
- Kasia Hitczenko, Loann Peurey, WN Harvard, Kai Jia Tey, Amanda Seidl, Chiara Semenzin, Camila Scaff, Marvin Lavechin, Bridgette Kelleher, Lisa Hamrick, et al. 2025. Speech maturity dataset: A cross-cultural corpus of naturalistic child and adult vocalizations. *OSF Preprints*.
- Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2022. Lora: Low-rank adaptation of large language models. *International Conference on Learning Representations*.
- Ailin Huang, Boyong Wu, Bruce Wang, Chao Yan, Chen Hu, Chengli Feng, Fei Tian, Feiyu Shen, Jingbei Li, Mingrui Chen, et al. 2025a. Step-audio: Unified understanding and generation in intelligent speech interaction. *arXiv preprint arXiv:2502.11946*.
- Chien-yu Huang, Wei-Chih Chen, Shu-wen Yang, Andy T Liu, Chen-An Li, Yu-Xiang Lin, Wei-Cheng Tseng, Anuj Diwan, Yi-Jen Shih, Jiatong Shi, et al. 2025b. Dynamic-superb phase-2: A collaboratively expanding benchmark for measuring the capabilities of spoken language models with 180 tasks. In *International Conference on Learning Representations*.
- Chien-yu Huang, Ke-Han Lu, Shih-Heng Wang, Chi-Yuan Hsiao, Chun-Yi Kuan, Haibin Wu, Siddhant Arora, Kai-Wei Chang, Jiatong Shi, Yifan Peng, et al. 2024. Dynamic-superb: Towards a dynamic, collaborative, and comprehensive instruction-tuning benchmark for speech. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 12136–12140. IEEE.

- Bashima Islam, Nancy L McElwain, Jialu Li, Maria I Davila, Yannan Hu, Kexin Hu, Jordan M Bodway, Ashutosh Dhekne, Romit Roy Choudhury, and Mark Hasegawa-Johnson. 2024. Preliminary technical validation of littlebeats™: A multimodal sensing platform to capture cardiac physiology, motion, and vocalizations. *Sensors*, 24(3):901.
- Kristina T Johnson, Jaya Narain, Thomas Quatieri, Pattie Maes, and Rosalind W Picard. 2023. Recanvo: A database of real-world communicative and affective nonverbal vocalizations. *Scientific Data*, 10(1):523.
- Evgenii Kalenkovich, Sharath Koorathota, Shaelise Tor, Andrei Amatuni, Shannon Egan-Dailey, Charlotte Moore, Catherine Laing, Hallie Garrison, Gladys Baudet, Federica Bulgarelli, et al. 2025. A year of nouns from english-learning infants’ daily lives: The seedlings-nouns dataset: E. kalenkovich et al. *Behavior Research Methods*, 57(11):298.
- Alkis Koudounas, Moreno La Quatra, Sabato Marco Siniscalchi, and Elena Baralis. 2025. voc2vec: A foundation model for non-verbal vocalization. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.
- Marvin Lavechin, Erika Bergelson, and Roger Levy. 2026. Babar: from phoneme recognition to developmental measures of young children’s speech production. *arXiv preprint arXiv:2603.05213*.
- Nonie K Lesaux and Linda S Siegel. 2003. The development of reading in children who speak english as a second language. *Developmental psychology*, 39(6):1005.
- Jialu Li, Mark Hasegawa-Johnson, and Nancy L McElwain. 2021. Analysis of acoustic and voice quality features for the classification of infant and mother vocalizations. *Speech communication*, 133:41–61.
- Jialu Li, Mark Hasegawa-Johnson, and Nancy L McElwain. 2023. Towards robust family-infant audio analysis based on unsupervised pretraining of wav2vec 2.0 on large-scale unlabeled family audio. In *Interspeech*, volume 2023, page 1035.
- Jialu Li, Mark Hasegawa-Johnson, and Nancy L McElwain. 2024. Analysis of self-supervised speech models on children’s speech and infant vocalizations. In *2024 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*, pages 550–554. IEEE.
- Jialu Li, Marvin Lavechin, Xulin Fan, Nancy L McElwain, Alejandrina Cristia, Paola Garcia-Perera, and Mark A Hasegawa-Johnson. 2026a. Automated analysis of naturalistic recordings in early childhood: Applications, challenges, and opportunities. *IEEE Signal Processing Magazine*, 42(6):16–34.
- Shuhe Li, Chenxu Guo, Jiachen Lian, Cheol Jun Cho, Wenshuo Zhao, Xiner Xu, Ruiyu Jin, Xiaoyu Shi, Xuanru Zhou, Dingkun Zhou, et al. 2026b. K-function: Joint pronunciation transcription and feedback for evaluating kids language function. In *ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3231–3235. IEEE.
- Marguerite Lockhart-Bouron, Andrey Anikin, Katarzyna Pisanski, Siloé Corvin, Clément Cornec, Léo Papet, Florence Levréro, Camille Fauchon, Hugues Patural, David Reby, et al. 2023. Infant cries convey both stable and dynamic information about age and identity. *Communications Psychology*, 1(1):26.
- Bria Long, Sarah Goodin, George Kachergis, Virginia A Marchman, Samaher F Radwan, Robert Z Sparks, Violet Xiang, Chengxu Zhuang, Oliver Hsu, Brett Newman, et al. 2024. The babyview camera: designing a new head-mounted camera to capture children’s early social and visual environments. *Behavior Research Methods*, 56(4):3523–3534.
- Catherine Lord, Mayada Elsabbagh, Gillian Baird, and Jeremy Veenstra-Vanderweele. 2018. Autism spectrum disorder. *The lancet*, 392(10146):508–520.
- Catherine Lord, Michael Rutter, Pamela C DiLavore, Susan Risi, K Gotham, SL Bishop, et al. 2008. *Autism diagnostic observation schedule (ADOS): Manual*. Western Psychological Services Los Angeles.
- Emma McGonigle, Mark VanDam, Carol Wilkinson, and Kristina T Johnson. 2024. Benchmarking automatic speech recognition technology for natural language samples of children with and without developmental delays. In *2024 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 1–5. IEEE.
- Lucas Block Medin, Thomas Pellegrini, and Lucile Gelin. 2024. Self-supervised models for phoneme recognition: Applications in children’s speech for reading learning. *Interspeech*.
- Aravind Kumar Namasivayam, Deirdre Coleman, Aisling O’Dwyer, and Pascal Van Lieshout. 2020. Speech sound disorders in children: An articulatory phonology perspective. *Frontiers in psychology*, 10:2998.
- Jorge Oliveira, Francesco Renna, Paulo Dias Costa, Marcelo Nogueira, Cristina Oliveira, Carlos Ferreira, Alípio Jorge, Sandra Mattos, Thamine Hatem, Thiago Tavares, et al. 2021. The circor digiscope dataset: from murmur detection to murmur classification. *IEEE journal of biomedical and health informatics*, 26(6):2524–2535.
- Leonardo Pepino, Pablo Riera, and Luciana Ferrer. 2021. Emotion recognition from speech using wav2vec 2.0 embeddings. In *Proc. Interspeech 2021*, pages 3400–3404.
- 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.
- Dhulipalla Venkata Rao, Kodali Radha, Sai Rekha Gudivaka, and Pranuthi Polipogu. 2024. Children’s speech emotion recognition using feature fusion-based deep neural networks. In *2024 IEEE International Conference on Intelligent Signal Processing and Effective Communication Technologies (IN-SPECT)*, pages 1–6. IEEE.
- Dima Rekesh, Nithin Rao Koluguri, Samuel Kriman, Somshubra Majumdar, Vahid Noroozi, He Huang, Oleksii Hrinchuk, Krishna Puvvada, Ankur Kumar, Jagadeesh Balam, et al. 2023. Fast conformer with linearly scalable attention for efficient speech recognition. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1–8. IEEE.
- Manuel Sam Ribeiro, Joanne Cleland, Aciel Eshky, Korin Richmond, and Steve Renals. 2021. Exploiting ultrasound tongue imaging for the automatic detection of speech articulation errors. *Speech Communication*, 128:24–34.
- Bruno M Rocha, Dimitris Filos, Luís Mendes, Gorkem Serbes, Sezer Ulukaya, Yasemin P Kahya, Nikša Jakovljevic, Tatjana L Turukalo, Ioannis M Voziatzis, Eleni Perantoni, et al. 2019. An open access database for the evaluation of respiratory sound classification algorithms. *Physiological measurement*, 40(3):035001.
- Yvan Rose and Brian MacWhinney. 2014. The phonbank project. *The Oxford Handbook of Corpus Phonology*, pages 380–401.
- Vishwanath Pratap Singh, Md Sahidullah, and Tomi H Kinnunen. 2026. Causal analysis of asr errors for children: Quantifying the impact of physiological, cognitive, and extrinsic factors. *Computer Speech & Language*, 95:101859.
- Abhijit Sinha, Hemant Kumar Kathania, and Mikko Kurimo. 2026. A study on the layer-wise transferability of self-supervised learning features for children’s speech processing tasks. *Speech Communication*, page 103392.
- Helen Tager-Flusberg, Sally Rogers, Judith Cooper, Rebecca Landa, Catherine Lord, Rhea Paul, Mabel Rice, Carol Stoel-Gammon, Amy Wetherby, and Paul Yoder. 2009. Defining spoken language benchmarks and selecting measures of expressive language development for young children with autism spectrum disorders. *Journal of Speech, Language, and Hearing Research*, 52(3):643–652.
- Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, and Chao Zhang. 2024. Salmonn: Towards generic hearing abilities for large language models. In *International Conference on Learning Representations*.
- Laura Wagner, Sharifa Alghowinhem, Abeer Alwan, Kristina Bowdrie, Cynthia Breazeal, Cynthia G Closser, Eric Fosler-Lussier, Izabela A Jamsek, Devan Lander, Rajiv Ramnath, et al. 2025. The ohio child speech corpus. *Speech Communication*, 170:103206.
- Anfeng Xu, Tiantian Feng, Somer Bishop, Catherine Lord, and Shrikanth Narayanan. 2026. End-to-end joint asr and speaker role diarization with child-adult interactions. *IEEE Transactions on Audio, Speech and Language Processing*.
- Anfeng Xu, Rajat Hebbar, Rimita Lahiri, Tiantian Feng, Lindsay Butler, Lue Shen, Helen Tager-Flusberg, and Shrikanth Narayanan. 2023a. Understanding spoken language development of children with asd using pre-trained speech embeddings. *Interspeech*.
- Anfeng Xu, Kevin Huang, Tiantian Feng, Lue Shen, Helen Tager-Flusberg, and Shrikanth Narayanan. 2024. Exploring speech foundation models for speaker diarization in child-adult dyadic interactions. In *Proc. Interspeech 2024*, pages 5193–5197.
- Hainan Xu, Fei Jia, Somshubra Majumdar, He Huang, Shinji Watanabe, and Boris Ginsburg. 2023b. Efficient sequence transduction by jointly predicting tokens and durations. In *International Conference on Machine Learning*, pages 38462–38484. PMLR.
- Anyu Ying, Natarajan Balaji Shankar, Chyi-Jiunn Lin, Mohan Shi, Pu Wang, Hye-jin Shim, Siddhant Arora, Abeer Alwan, Shinji Watanabe, et al. 2025. Benchmarking training paradigms, dataset composition, and model scaling for child asr in espnet. *Workshop on Child Computer Interaction (WOCCI)*.
- 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 *Proc. Interspeech 2021*, pages 3710–3714.
- Qing Zhang, Jing Zhang, Jiajun Yuan, Huajie Huang, Yuhang Zhang, Baoqin Zhang, Gaomei Lv, Shuzhu Lin, Na Wang, Xin Liu, et al. 2022. Sprsound: Open-source sjtu paediatric respiratory sound database. *IEEE Transactions on Biomedical Circuits and Systems*, 16(5):867–881.
- Theo Zhang, Madurya Suresh, Anne Warluamont, Kasia Hitczenko, Alejandrina Cristia, and Margaret Cychoz. 2025. Employing self-supervised learning models for cross-linguistic child speech maturity classification. In *Interspeech 2025*, pages 2825–2829.

A Dataset Details

CirCor The CirCor DigiScope Phonocardiogram Dataset (Oliveira et al., 2021) is one publicly available pediatric heart-sound corpus, which includes 5,272 phonocardiogram recordings from 1,568 subjects aged 0–21 years. Each subject is annotated

by an expert cardiac physiologist with a murmur label drawn from present, absent, or unknown. We use the publicly released training/testing partition to classify these three labels. We segment each recording into non-overlapping 10-second windows. We further perform between-subject 5-fold cross-validation, partitioning the train data with 80/20 into train and development sets.

ICBHI ICBHI (Rocha et al., 2019) is a publicly available benchmark for respiratory sound analysis, targeting the recognition of crackles, wheezes, and respiratory conditions. It comprises 920 recordings collected from 126 subjects, accompanied by two complementary annotation schemes. The dataset provides cycle-level labels for 6,898 respiratory cycles, each categorized as normal, containing crackles, containing wheezes, or both. Similar to the CirCor dataset, we use the official ICBHI train/test split and apply 5-fold cross-validation over the training subjects to construct train/dev folds. For respiratory condition classification, recordings are segmented into 10-second windows with a 5-second hop, with diagnoses mapped to healthy, COPD, and other. For crackle and wheeze detection, we instead segment each recording according to the provided respiratory-cycle annotations. Segments shorter than 3 seconds (recording-level) or 0.5 seconds (cycle-level) are discarded.

SPRSound SPRSound (Zhang et al., 2022) is a paediatric respiratory sound database, designed for automatic diagnosis. It contains 2683 records and 9089 respiratory sound events collected from 292 participants and annotated by 11 annotators. This dataset categorizes respiratory sound into two major labels: normal and adventitious. Specifically, there are continuous adventitious (CAS, such as Rhonchi, Wheeze, and Stridor) and discrete adventitious (DAS, such as Coarse Crackle and Fine Crackle) based on duration. We use the official SPRSound train/test split and apply 5-fold cross-validation over the training subjects to construct training and development folds. For respiratory sound classification, we use the record-level annotations provided with each recording, which are normal, CAS, DAS, CAS & DAS, and poor quality.

Donate-a-cry Donate-a-cry is a public infant cry audio corpus with labels for crying reasons. The audio samples in the corpus are collected with mobile devices and cleaned in the post-processing. There are 465 samples in the dataset, from children

aged from 0 weeks to 2 years. The dataset is distributed with extreme imbalance crying reasons in belly_pain (16), burping (8), discomfort (27), tired (24), and hungry (382). Therefore, we simplify the label as hungry and other.

What’s in a cry? What’s in a cry? (Lockhart-Bouron et al., 2023) is collected for a longitudinal infant crying study. The samples are recorded from 24 families with 10 girls and 14 boys. In our study, we focus on three primary cry causes, which are hunger, loneliness, and discomfort, discarding segments labeled as pain, other, or unknown due to limited sample availability. We merge consecutive cry segments within each recording when the inter-segment gap is less than 1 second to form continuous cry bouts, and retain only merged segments with a duration of at least 3 seconds. The resulting dataset is partitioned at the subject level using 5-fold cross-validation, where within each fold 80% of the training subjects are used for training and the remaining 20% for validation.

AudioSet AudioSet (Gemmeke et al., 2017) is a large-scale ontology-driven collection of over one million YouTube video clips. In this benchmark, we focus on child sound event classification and filter data belonging to the following ten child-related audio categories: speaking, babbling, giggling, shouting, laughing, crying, whimpering, singing, playing, and music for children. We select the AudioSet-balanced dataset as the training set.

ReCANVo ReCANVo (Johnson et al., 2023) is designed to investigate the informative expressions in nonverbal vocalizations from real-world interaction. The participants are 6 males and 2 females, aged from 6 to 23, with different developmental disorders. The speech samples from each speaker have at least 10 pieces with fewer than 10 words. Overall, the data has 7,077 audio samples, recorded in a home environment. In our benchmark, we use six expression categories: delighted, dysregulated, frustrated, request, selftalk, and social, while discarding samples that fall outside this label set. The dataset is partitioned at the utterance level using 5-fold cross-validation, where within each fold 80% of the training and validation samples are used for training and the remaining 20% for testing.

BabbleCor BabbleCor (Cychosz et al., 2021) is a crosslingual corpus of infants and children to investigate vocal development in typical speech.

52 participants are exposed to five different languages, including English, Spanish, Yêlî-Dnye, Tseltal Mayan, and bilingual Quechua-Spanish. Audio samples are clipped (400ms) from a day-long real-life recording for each child and annotated into five categories: canonical, non-canonical, crying, laughing, and junk. In our data processing, we use all five vocalization categories: crying, laughing, canonical, non-canonical, and junk, while discarding data marked as NO-LABEL. The dataset is partitioned at the subject level using 5-fold cross-validation, where within each fold 80% of the training subjects are used for training and validation, and the remaining 20% for testing.

SpeechMaturity SpeechMaturity (Hitezenko et al., 2025; Zhang et al., 2025) is also designed to investigate children’s vocal development. Following BabbleCor, SpeechMaturity acquires clips from day-long children’s speech recordings and gets clips annotated into the five categories. But SpeechMaturity is collected from a larger population size, containing 242,004 labeled vocalizations from more than 25 languages. Our data processing approach follows the same processing steps as stated in the BabbleCor dataset.

C-BESD C-BESD (Rao et al., 2024) is a bilingual dataset developed for children’s emotion development research. The dataset contains 4200 utterances from 70 child speakers, aged from 6 to 12, and gender-balanced. Each speaker has 5 utterances for each of the 6 emotions (Happy, Neutral, Angry, Disgust, Fear, and Sad) in both Telugu and English. In our study, we focus on four primary emotion categories: angry, happy, neutral, and sad. This choice is similar to common emotion recognition benchmarks (Feng and Narayanan, 2024). The dataset is partitioned at the speaker level using 5-fold cross-validation, where within each fold 80% of the training subjects are used for training and validation, and the remaining 20% for testing.

PERCEPT-R PERCEPT-R (Benway et al., 2022) focuses on the speech of children from fully-rhotic American English with Rhotacism Speech Sound Disorder (RSSD). PERCEPT-R contains 32.47 hours of citation speech recordings, comprising 105,232 word-level utterances. The 281 children (121 girls and 160 boys) are involved in speech recording, ranging from 72 months to 288 months. In data processing, we discard any utterances without a valid perceptual label. We restrict

the data from the recordings including PRE-1 and DP-1 due to computational constraints. The dataset is partitioned at the subject level using 5-fold cross-validation, where within each fold 80% of the training subjects are used for training and validation, and the remaining 20% for testing.

SpeechOcean762 SpeechOcean762 (Zhang et al., 2021) is an open-source non-native English speech corpus for multi-granularity pronunciation assessment. It contains 5,000 utterances from 250 Mandarin-speaking learners (half children, half adults) with balanced gender and proficiency distributions. Each utterance is annotated by five experts to assess its accuracy, fluency, and prosody. Specifically, for accuracy and prosody, scores of 6 or below are mapped to “Poor”, 7–8 to “Nearly correct”, and 9–10 to “Correct”. For fluency, scores below 5 are mapped to “Intermittent”, 5–6 to “Fluent in general”, and 7 or above to “Fluent”. We adopt the official train and test splits, and further split the training set at the speaker level using 5-fold cross-validation.

UltraSuite UltraSuite (Eshky et al., 2018) is a multimodal repository of synchronized ultrasound and acoustic data from child speech therapy sessions, containing 86 children (58 typically developing, 28 with SSDs) across three sub-datasets totaling 37.28 hours of audio and 14,456 utterances. Data spans six prompt types and includes longitudinal recordings across therapy stages for children with disorders such as phonological delay and childhood apraxia of speech. In our benchmark, we focus on the typically developing children subset (core-uxtd) and restrict the data to Type A prompts, which consist of single-word utterances for word-level articulation analysis. We formulate a multi-label classification task targeting eight places of articulation: alveolar, bilabial, palatal, glottal, labiovelar, velar, dental, and labiodental, by mapping each target word to its constituent consonant phonemes via the CMU Pronouncing Dictionary. We adopt the official train, dev, and test partitions provided with the dataset.

Natural Language Sampling (NLS) The dataset comprises 73 English-speaking child-parent dyadic interaction sessions that occurred in naturalistic or home settings. No child appears more than twice in the dataset. Most children are diagnosed with ASD and are minimally verbal. Recordings were made remotely by research personnel over Zoom.

Domain experts assigned each child’s spoken language levels to one of the three levels: pre-verbal (LL-1), first words (LL-2), or word combinations (LL-3). This definition was based on the prior study by Tager-Flusberg et al. (2009).

In-house ADOS2-Mod3 The ADOS (Lord et al., 2008) is a semi-structured assessment protocol lasting 40–60 minutes, where ADOS-Module 3 (Mod3) targets verbally fluent children across 14 pre-defined activities. The in-house ADOS-Mod3 dataset (Xu et al., 2026) focuses on two interview-style components, Social Difficulties and Annoyance and Emotions. This produces 352 segments from 180 children (mean age 8.53 years, range 3–13). Roughly half received an ASD diagnosis, with most others diagnosed with ADHD or other neurodevelopmental or psychiatric conditions. We used the raw annotated data, whereas other works (Xu et al., 2026) applied additional preprocessing steps to clean the data.

MyST MyST (Pradhan et al., 2024) is one of the largest children’s conversational speech corpora, containing 393 hours and 228,874 utterances from 1,371 third-to-fifth-grade students engaging in one-on-one spoken dialog with a virtual science tutor. The corpus includes a stratified train/dev/test split for the ASR experiment.

TinyVox TinyVox (Lavechin et al., 2026) is a large-scale cross-linguistic corpus of over 500,000 IPA-transcribed child vocalizations (387.7 hours) from 560 children aged 5–96 months, compiled and standardized from 31 PhonBank corpora spanning English, French, Portuguese, German, and Spanish, with a unified 57-phoneme target inventory mapped via phonological feature edit distance. The dataset provides a speaker-independent 80/10/10 train/validation/test split. In this study, we only perform the phoneme experiments with LALM, since the token spaces of the decoders for the evaluated ASR models are not directly compatible with phoneme recognition targets.

B Training Details

B.1 Training Parameters

For non-ASR tasks, during fine-tuning, SSAST models are trained with learning rates from 1×10^{-4} , 5×10^{-4} , while other pre-trained speech models use learning rates from 2×10^{-4} , 5×10^{-4} , 1×10^{-3} . We use a learning

rate of 1×10^{-4} in fine-tuning Qwen2Audio and AudioFlamingo3. All experiments are trained for 10 epochs. We set the maximum audio duration of 10 seconds for most audio and speech classification tasks. We apply a batch size of 256 with a mini-batch size of 8 when fine-tuning LALMs. During training, we applied several augmentations to the input waveforms: Gaussian noise was added with a probability of 1.0, using an SNR range of 3–30 dB; time stretching was used with a probability of 1.0, with stretch rates ranging from 0.9 to 1.1; and polarity inversion was applied with a probability of 0.5.

For ASR fine-tuning, all models are trained with a learning rate of 1×10^{-4} and a batch size of 32. We train for 5 epochs on the MyST dataset and 20 epochs on the ADOS dataset. Prior to fine-tuning, all transcripts are normalized using Whisper’s BasicTextNormalizer, and the same normalization is applied to the model outputs during inference. Unlike the work in (Fan et al., 2024), we do not filter samples based on transcription length or the zero-shot Whisper-large-v3 WER, allowing for a more realistic and comprehensive evaluation of model performance.

B.2 Training Architecture

Encoder-based Fine-tuning Prior works (Pepino et al., 2021; Feng et al., 2025a) have shown that fine-tuning hidden representations from pre-trained encoder layers of speech foundation models can effectively adapt to a broad range of downstream speech tasks. Based on this, we adopt the architecture in Figure 3, where a learnable weighted average over all encoder layers (both convolutional and transformer) is passed through a 1D point-wise convolution, followed by temporal averaging, and fed to a fully connected classifier. We further apply LoRA (Hu et al., 2022) to all fine-tuning experiments for parameter-efficient adaptation of the foundation model backbones, including for the ASR fine-tunings. However, due to implementation differences with wav2vec-LB and BabyHuBERT, we choose to fine-tune these two models only using hidden representations without LoRA.

LALMs As stated in the main paragraph, we directly apply LoRA modules with rank 64 into the query, key, value, down-projection, and up-projection layers of the language model, and feed-forward layers of the audio encoder. We use 4-bit

quantization during training and normal 16-bit precision at inference.

B.3 Training Hardware

All experiments with public datasets are conducted on a High-Performance Computing (HPC) cluster. Most experiments are conducted with a single A40 or V100 GPU. Experiments with private data are conducted on a GPU server with 4 A6000 GPUs. Most experiments on encoder-based models require a fine-tuning time of less than 2 days, and experiments on LALMs require 4-5 days on an A40 GPU.

C Prompt Design

While LALMs are trained on generalized speech tasks, they may not have the domain knowledge about vocalizations/speech patterns specific to children. Hence, in our SFT pipeline, we design system prompts for each dataset so as to include descriptions for each of the classes in the datasets. System prompt was first designed by the authors and then filtered using a frontier model (Gemini 3.1-pro). System prompts corresponding to SFT of each dataset are as follows:

C.1 SpeechMaturity

Listen to the child audio recording and classify the vocalization type.

Output only the label - no explanation or punctuation.

Labels:

- Crying: distress vocalizations, whimpers, or wails
- Laughing: laughter or giggling sounds
- Canonical: well-formed syllables with clear consonant-vowel structure (e.g., ``ba'', ``da'')
- Non-Canonical: vocalizations lacking clear consonant-vowel structure (e.g., squeals, growls, vowel-only sounds)
- Junk: non-speech noise, silence, or unintelligible audio

C.2 BabbleCor

Listen to the child audio recording and classify the vocalization type.

Output only the label - no explanation or punctuation.

Labels:

- Crying: distress vocalizations, whimpers, or wails
- Laughing: laughter or giggling sounds
- Canonical: well-formed syllables with clear consonant-vowel structure (e.g., ``ba'', ``da'')
- Non-Canonical: vocalizations lacking clear consonant-vowel structure (e.g., squeals, growls, vowel-only sounds)
- Junk: non-speech noise, silence, or unintelligible audio

C.3 PERCEPT-R

Listen to the child reading a word. Evaluate the pronunciation of the 'r' sound specifically.

Output only the label - no explanation or punctuation.

Labels:

- Rhotic: the 'r' sound is produced correctly with appropriate rhoticity
- Derhotic: the 'r' sound is missing, distorted, or replaced (e.g., substituted with 'w')

C.4 SpeechOcean762-Prosody

Listen to the child reading a sentence.

Evaluate the prosody - focus on intonation, stress, and rhythm.

Output only the label - no explanation or punctuation.

Labels:

- Poor intonation: flat, monotone, or clearly unnatural intonation patterns
- Nearly correct intonation: mostly natural prosody with minor errors
- Correct intonation: natural, native-like intonation and rhythm throughout

C.5 SpeechOcean762-Fluency

Listen to the child reading a sentence.

Evaluate the fluency - focus on flow, pausing, and hesitations.

Output only the label - no explanation or punctuation.

Labels:

- Intermittent or little fluent: frequent pauses, hesitations, or disruptions in speech flow
- Fluent in general: mostly smooth delivery with occasional minor disfluencies
- Fluent: smooth, continuous speech with no notable disfluencies

C.6 SpeechOcean762-Accuracy

Listen to the child reading a sentence.

Evaluate the pronunciation accuracy - focus on correct phoneme production.

Output only the label - no explanation or punctuation.

Labels:

- Poor or Understandable: many mispronunciations; difficult to understand
- Good: mostly accurate pronunciation with some errors
- Excellent: highly accurate, native-like pronunciation throughout

C.7 CirCor

Listen to the heart sound recording.

Determine whether a cardiac murmur is present.

Output only the label - no explanation or punctuation.

Labels:

- Absent: no murmur detected; normal heart sounds
- Present: a murmur is clearly audible
- Unknown: audio quality or ambiguity prevents a confident determination

C.8 ICBHI-Crackles

Listen to the lung sound recording. Determine whether crackles are present. Output only the label - no explanation or punctuation.

Labels:

- Crackles: discontinuous, explosive sounds (fine or coarse crackles) are audible
- No crackles: no crackle sounds detected

C.9 ICBHI-Wheezes

Listen to the lung sound recording. Determine whether wheezes are present. Output only the label - no explanation or punctuation.

Labels:

- Wheezes: continuous, high-pitched whistling sounds are audible
- No wheezes: no wheeze sounds detected

C.10 SPRSound

Listen to the lung sound recording. Classify the respiratory sound pattern. Output only the label - no explanation or punctuation.

Labels:

- Normal: clear breath sounds with no adventitious sounds
- CAS: continuous adventitious sounds only (e.g., wheezes, rhonchi)
- DAS: discontinuous adventitious sounds only (e.g., crackles)
- CAS & DAS: both continuous and discontinuous adventitious sounds present
- Poor Quality: recording is too noisy or unclear for reliable classification

C.11 ReCANVo

Listen to the child audio recording and classify the communicative intent or emotional state of the vocalization. Output only the label - no explanation or punctuation.

Labels:

- Delighted: joyful, excited, or positively expressive sounds
- Dysregulated: distressed, overwhelmed, or emotionally dysregulated vocalizations
- Frustrated: sounds expressing frustration or dissatisfaction
- Request: vocalizations used to request an object, action, or attention
- Selftalk: vocalizations directed inward, not toward a social partner
- Social: vocalizations directed toward engaging or interacting with others

C.12 What's in a cry?

Listen to the infant cry recording and identify the most likely cause of the cry. Output only the label - no explanation or punctuation.

Labels:

- Hunger: cry pattern associated with feeding need
- Loneliness: cry pattern associated with desire for social contact or comfort. Loneliness was defined as the infant seeking contact or attention
- Discomfort: cry pattern associated with physical discomfort. Examples of discomfort included cold, fever, and full diaper.

C.13 MyST

Listen to the child audio recording and transcribe exactly what is spoken. Output only the transcript - no labels, explanations, punctuation beyond what is spoken, or formatting. If the audio is unintelligible, output: [unintelligible]

C.14 TinyVox

Listen to the child speaking a single word. Transcribe the pronunciation using standard phoneme symbols (IPA). Output only the phoneme sequence - no labels, explanations, or formatting. Example format (IPA): /b ae t/

D Additional Whisper Results

Additional results on Whisper-base and Whisper-small are presented in Table 7.

E Proprietary Baselines

We apply both Gemini 2.5 Flash and Gemini 3.5 Flash APIs on 5 selected public datasets in the main paragraph. We apply a temperature of 0.1, 0.5, and 1.0 in prompting both models. We run the prompt experiment on these two proprietary baselines for a single run due to the budget limitation. Figure 7 shows the confusion matrices of Gemini 3.5 Flash vs. ChildVox Qwen2-Audio on PERCEPT-R and SpeechOcean Intonation predictions. Specifically, on PERCEPT-R, Gemini 3.5 Flash over-predicts derhotic, while our ChildVox Qwen-Audio 2 has better rhotic prediction. On SO-Prosody, Gemini fails toward the middle class, and ChildVox Qwen2-Audio instead over-predicts Correct intonation.

F Generative AI Use Disclaimer

Generative AI tools were used solely for grammar checking and polishing the writing of the manuscript. All research concepts, experimental design, implementations, analyses, and reported

Table 7: Additional benchmark performance across physiological sounds, vocalization, canonical syllables, and speech from children on each complete dataset. Whisper-B and Whisper-S are Whisper-Base and Whisper-Small, respectively. We report the average Macro-F1 score (higher the better).

Task Category	Dataset	Whisper-B	Whisper-S
Physiological Sounds	CirCor	0.499	0.546
	ICBHI-crackles	0.581	0.597
	ICBHI-wheezes	0.614	0.621
	ICBHI-Condition	0.473	0.483
	SPRSound	0.345	0.362
Vocalization	AudioSet-Child	0.598	0.634
	What's in a cry?	0.333	0.326
	donate-a-cry	0.467	0.528
Canonical Syllables	ReCANVo	0.414	0.415
	Speech Maturity	0.625	0.634
	BabbleCor	0.546	0.596
Speech	NLS-Speech Intelligibility	0.635	0.658
	Percept-R	0.794	0.807
	Ultrasuite	0.923	0.926
	SpeechOcean-Fluency	0.625	0.626
	SpeechOcean-Accuracy	0.605	0.628
	SpeechOcean-Prosody	0.701	0.704
	C-BESD	0.596	0.593

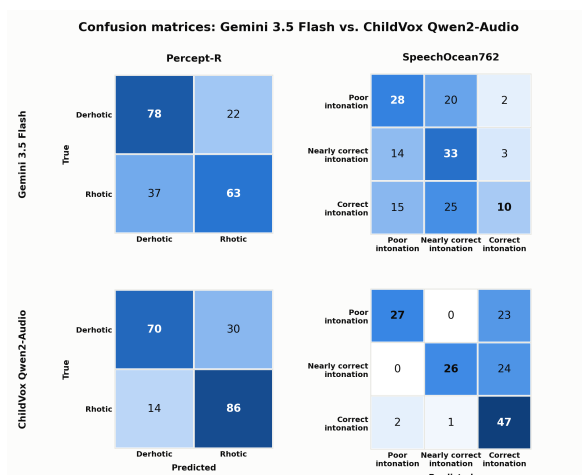


Figure 7: Confusion matrices of Gemini 3.5 Flash vs. ChildVox Qwen2-Audio on PERCEPT-R and SpeechOcean Intonation predictions.

results were developed and conducted by the authors. The conceptual framing, methodology, and scientific contributions of this work are entirely human-generated.