

ChildVoice - Finding the Voice of Children in Audio Recordings

A model to diarize and isolate a four-year-old's speech from the adult test administrator to produce database for automated diagnostic tools

Kanagha Sivakumar · Sophie Liu · Jack Peterson · PedzSTAR Lab · Dr. Marisha Speights · Bharath Yedla

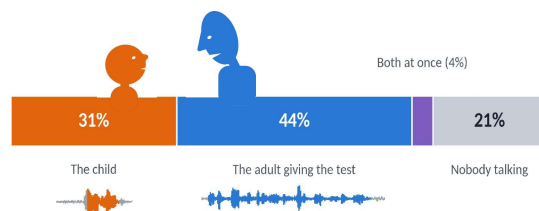


Why build a child-focused speech corpus?

About 1 in 14 children aged 3-17 has a disorder of speech or language, and 4 in 10 of them received no intervention service last year. Finding the children who need help means listening closely to a great many of them, and there are not enough clinicians to do that.

The lab recorded 305 four-year-olds taking standard speech and language tests amounting to 220 hours of audio to help flag the children who need a closer look.

Hand-transcribing child recordings runs at least seven minutes of annotator time per minute of audio amounting to about 1,500 hours for this corpus. ChildVoice annotates, facilitating transcription, in about 30 minutes on one A100 GPU.



The child wears a microphone and so does the adult giving the test and both hear everyone. On hand-marked second by second across ten children, more than half the speech on the child's own microphone is the adult's.

A third of what the child produces is not words. Partial word attempts, breaths, laughs. Widely used speaker-labelling tool only find 24-49% of them.

This corpus enables models to learn from only the children's successful utterances and speech errors to build automated diagnostic tools.

How do we build a corpus that can be used for any future diagnostic tool?

1. Adaptable corpus building tool.

New corpora can be built for diverse applications by changing the parameters of ChildVoice.

2. Never lose the child's speech.

The recording is kept whole and bit-for-bit identical; ChildVoice writes a score beside it rather than cutting it. Contamination can be removed later, but deleted speech cannot be recovered.

3. Label the sound, while not manipulating it.

Filtering the adult out of the recording removed almost none of the adult and cost the accuracy of the detector by about 12%.

4. Never generate speech.

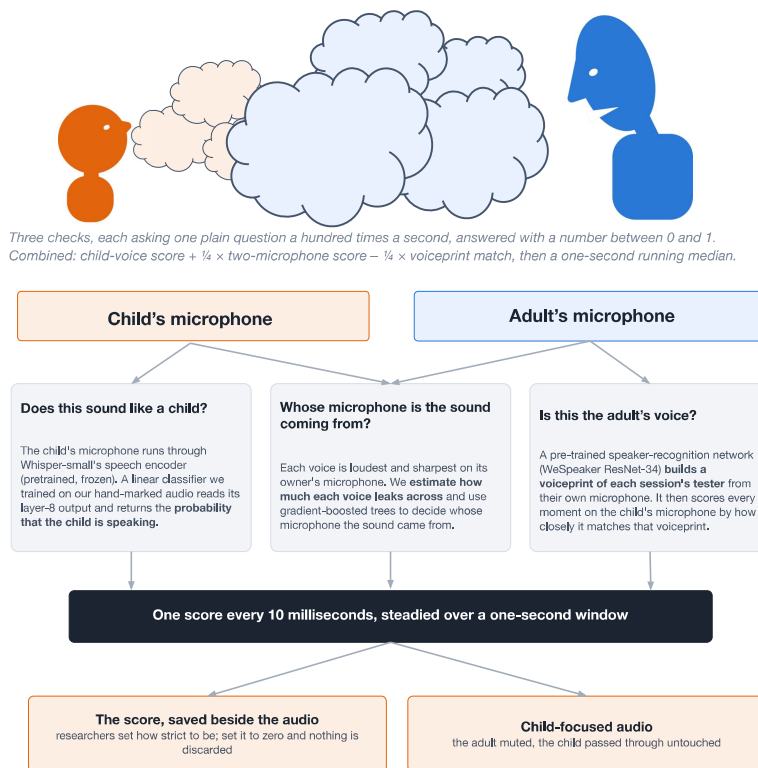
Enhancement models that rebuild a voice have been shown to erase the acoustic marks of disordered speech. ChildVoice only turns sound down; nothing is synthesised.

5. Model the adult and infer the child.

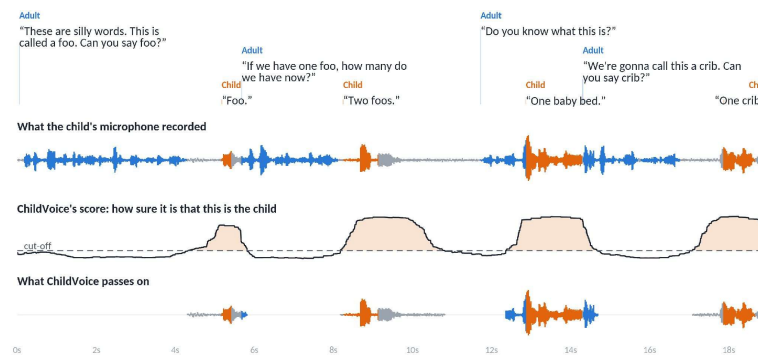
Recognising a four-year-old by voice is unreliable; recognising the adult, who is on tape for hours, is easy and a child is easily identified as "not the adult".

Sources

Prevalence and Intervention: NIDCO, Quick Statistics About Voice, Speech, Language (2025); Black, Vahtrian & Hoffman, NCHS Data Brief 205 (2015). - Annotation time: Casillas & Crista, Collabra: Psychology 5(1):24 (2019). - LENA: Crista et al., Behavior Research Methods 53:467-486 (2021). - Enhancement erases disordered speech cues: Reszka et al., arXiv:2412.13933 (2024). - Child-adult model: Xu et al., Interspeech 2024. - pynote audio: Bredin et al., ICASSP 2020. - Whisper: Radford et al., ICML 2023. - Voiceprints: Wang et al., WeSpeaker, ICASSP 2023. - Audio shown and played here: Ohio Child Speech Corpus, Wagner et al., Speech Communication 170:103206 (2025), TalkBank-CHLDES, CC BY-NC-SA 3.0.



Twenty seconds of a session, start to finish



A four-year-old being taught new words. Orange is the child, blue the adult. The adult's turns are muted; the child's speech passes through unchanged. From OSCS corpus as the PedzSTAR recordings cannot be shown.

ChildVoice retains high fidelity to child speech

97%

of the child's speech kept

70%

of the adult's speech removed at the same time

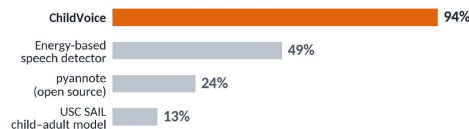
94%

of non-word sounds, such as partial words, clicks, breaths, kept

0.00 dB

change in loudness of the child's audio that is kept

Retention of incomplete or non-word sounds



Non-word sounds kept, measured on the same hand-marked recordings.

Consistent on new corpora: On sessions from OSCS of four-year-olds, one microphone, never seen in training it keeps the same 97% of the child while removing about half of the adult.



...but is also scalable as the corpus grows and stays flexible for different applications

- A new version of the corpus can be infinitely regenerated as no child speech is lost allowing for better detectors in future and accommodating downstream models.
- A tool that loses non-word sounds unevenly across sites and tasks can create group differences that aren't real. With a small blind spot, site and task comparisons reflect the children.
- Researchers in fields such as child speech pathology, developmental psychology, linguistics, pediatric neuroscience, etc., can use ChildVoice to automatically clean their child speech datasets.
- ChildVoice is scalable and thus can accommodate the SPROUT dataset as it incorporates more children from other US cities.
- In addition to successful word attempts, ChildVoice retains partial attempts, errors, breath sounds, etc., enabling a much fuller picture of children's developmental patterns than is not afforded by other systems.

