Human-Speech Detection in Naturalistic Child-Centered Audio using an Interpretable Open-Source Ensemble Voice Activity Detection Approach

Jun Ho Chai[1] & Eon-Suk Ko[1,2]
[1]Center for Data Science in Humanities, Chosun University
[2]Department of English Language and Literature, Chosun University
junho.chai@chosun.ac.kr; eonsukko@chosun.ac.kr

Naturalistic child-centered audio recordings are increasingly used to study children's everyday language environments, but their scientific value depends on accurately identifying when human speech occurs (VanDam et al., 2016; Gilkerson et al., 2018). Errors at this initial stage propagate to downstream measures of speech exposure, conversational turns, diarization, transcription, and interactional timing. Although proprietary systems have enabled large-scale analysis of developmental recordings, their closed pipelines limit transparency, adaptation, and independent validation (McDonald et al., 2021; Lavechin et al., 2020). The present study evaluated whether lightweight open-source voice activity detection (VAD) systems can provide a reliable, interpretable, and auditable front end for developmental audio research.

We benchmarked four complementary VAD systems—Kaldi, RNNoise, Silero, and WebRTC—against expert human annotation on 60 five-minute caregiver–child audio clips drawn from a validated child-centered corpus. To examine whether detector complementarity improves performance, we evaluated transparent ensemble strategies using majority voting and temporal regularization. Performance was assessed using duration accuracy, segmentation quality, temporal boundary agreement, and frame-level speech detection.

Results showed that no single detector performed best across all targets. Ensemble procedures consistently improved performance over individual detectors. Majority voting followed by temporal regularization produced the strongest interpretable profile, improving segment structure and temporal coherence while maintaining robust frame-level detection (F1 = .872). Supervised expert-fusion models further improved broad frame-level detection and segment overlap with human annotation, particularly XGBoost (micro-F1 = .895), but did not improve precise boundary localization relative to transparent voting-based methods.

Beyond benchmarking frame-level speech detection, VAD-derived segments preserved developmentally meaningful structure in children's everyday language environments. The derived measures captured speech quantity, temporal organization, and interactional dynamics, with systematic associations to child vocalization rate, conversational turns, vocalization length, and chronological age (Figure 2).

Figure 1. Off-the-shelf ensemble workflow for human-speech detection

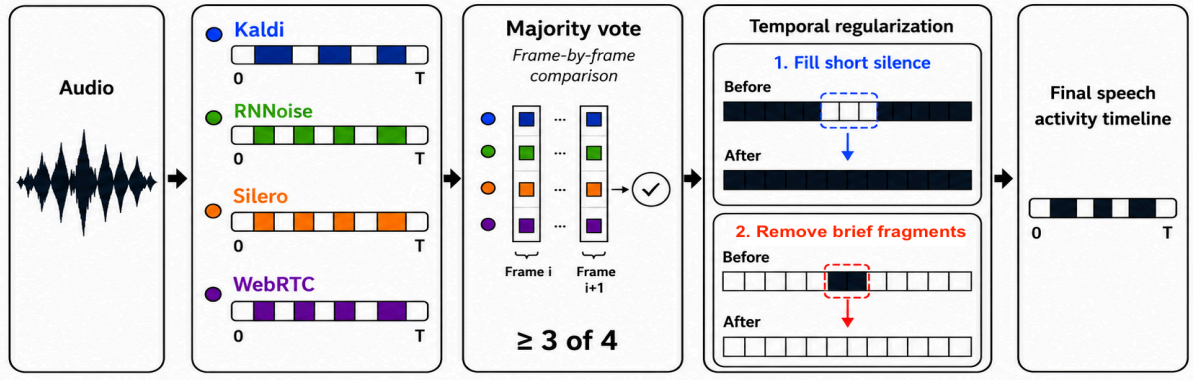
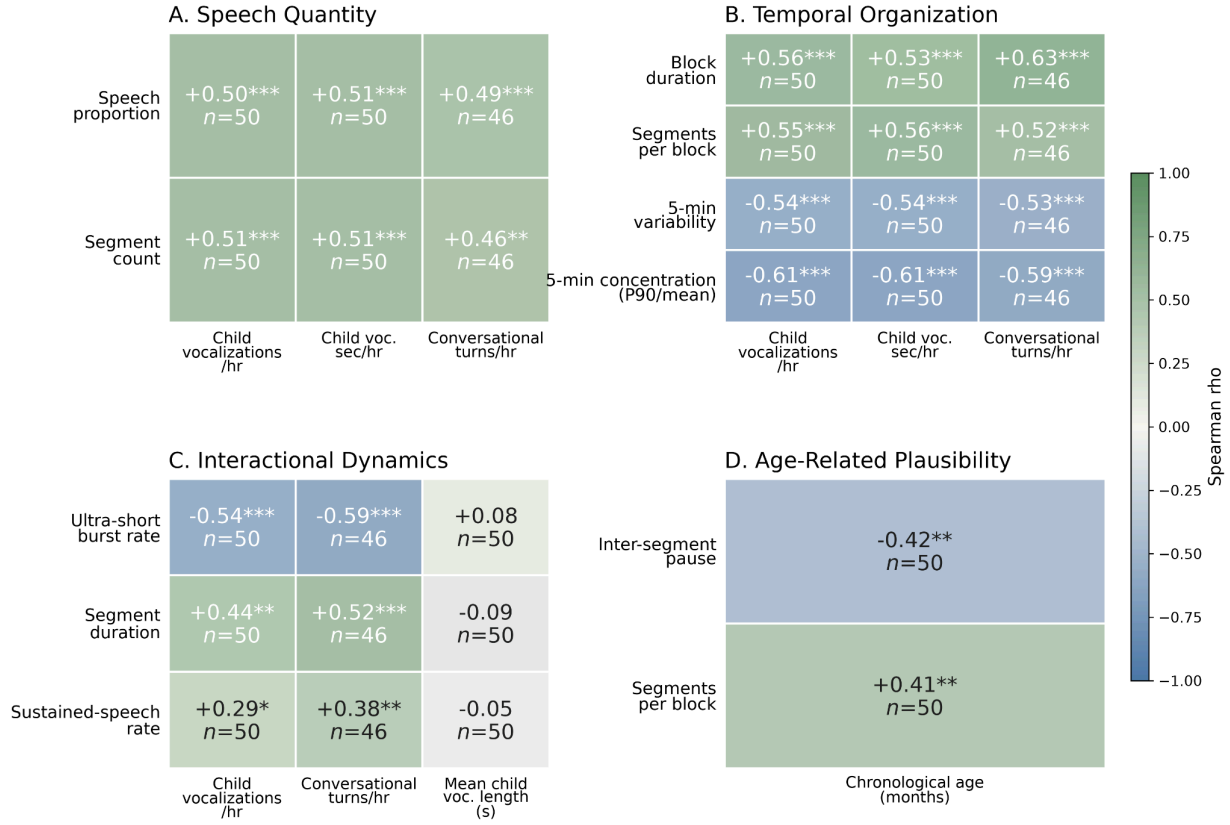


Figure 2. Downstream Utility of First-Pass Human Speech Detection with Expert Fusion (XGBoost)



References

Gilkerson, J., Richards, J. A., Warren, S. F., Oller, D. K., Russo, R., & Vohr, B. (2018). Language experience in the second year of life and language outcomes in late childhood. *Pediatrics*, *142*(4).

Lavechin, M., Bousbib, R., Bredin, H., Dupoux, E., & Cristia, A. (2020). An open-source voice type classifier for child-centered daylong recordings. *arXiv preprint arXiv:2005.12656*.

McDonald, M., Kwon, T., Kim, H., Lee, Y., & Ko, E. S. (2021). Evaluating the language environment analysis system for Korean. *Journal of Speech, Language, and Hearing Research*, *64*(3), 792-808.

VanDam, M., Warlaumont, A. S., Bergelson, E., Cristia, A., Soderstrom, M., De Palma, P., & MacWhinney, B. (2016, May). HomeBank: An online repository of daylong child-centered audio recordings. In *Seminars in speech and language* (Vol. 37, No. 02, pp. 128-142). Thieme Medical Publishers.