Automated Audio Synchronization Enables Time-Aligned Speech Analysis in Naturalistic Child Development Research

PSYCHOLOGY
EARLY PLAY & DEVELOPMENT LAB
MIAMI

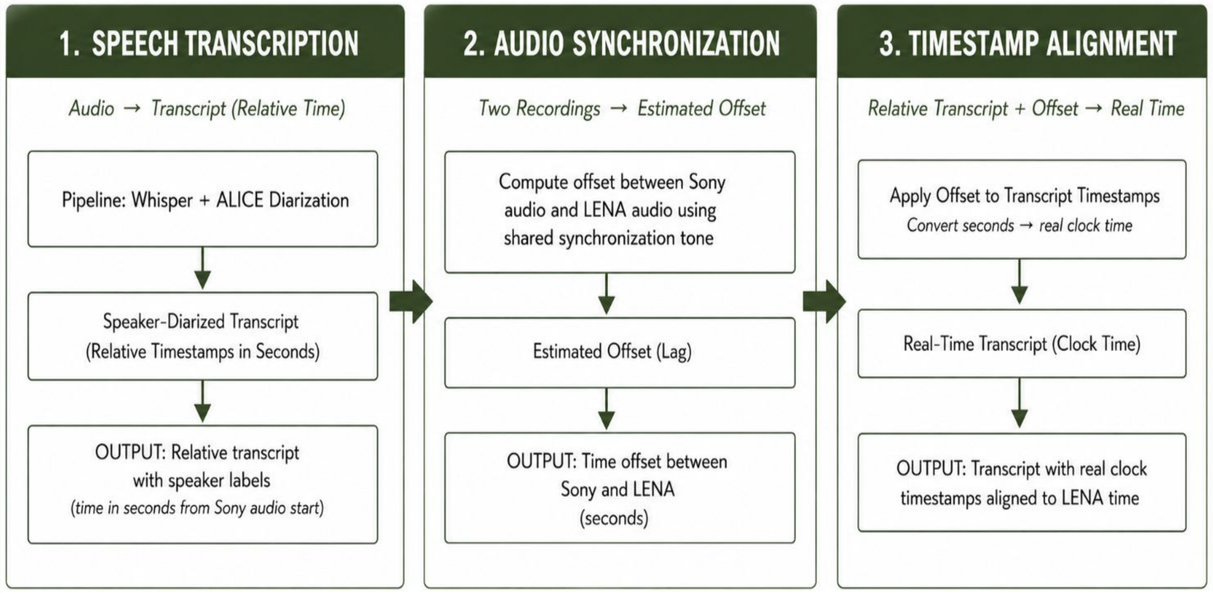
Reese Gmernicki, Laura Vitale, Luis Angeles, Daniel Messinger

Research Experiences for Undergraduates
National Science Foundation

INTRODUCTION

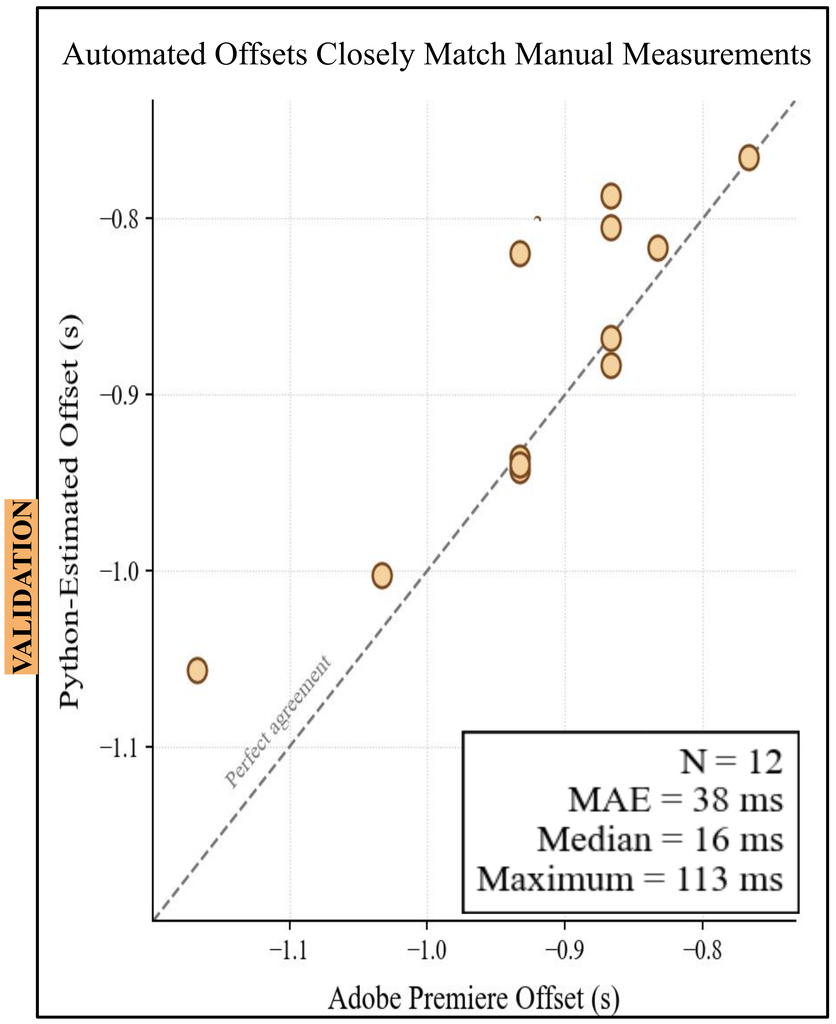
Child development studies often collect audio from multiple devices, but independent clocks cause transcripts and behavioral data to drift apart. Precise timing is essential for linking children's speech to movement, attention, and social context. We developed an automated synchronization pipeline that aligns transcripts to real-world time.

METHODS: AUTOMATED SYNCHRONIZATION PIPELINE

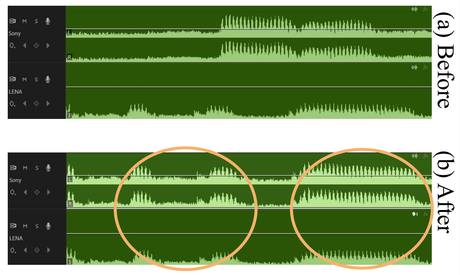


RESULTS

VALIDATION



WAVEFORM ALIGNMENT



Before (a) vs. after (b) of manual wave alignment with Adobe Premiere. Circled segments showcase aligned waves.

TRANSCRIPT CONVERSION

start_sec	end_sec	sentence	speaker
1258	1260	oh you mis	CHI
1270	1272	thank you	MAL
1278	1280	thank you	CHI
1280	1282	teacher	CHI
1284	1286	oh cool	FEM

real_start_	real_end	sentence	speaker
9:13:31	9:13:33	oh you missed i	CHI
9:13:43	9:13:45	thank you	MAL
9:13:51	9:13:53	thank you	CHI
9:13:53	9:13:55	teacher	CHI
9:13:57	9:13:59	oh cool	FEM

Pieces of the original Whisper/ALICE-diarized transcripts with local Sony timestamps vs. the alignment with LENA to show clock timestamps.

Key Takeaway: Automated synchronization achieved **38 ms MAE**, closely matching manual Adobe measurements and enabling accurate transcript-to-clock-time conversion.

DISCUSSION/CONCLUSION

- Automated offset estimation synchronized audio streams with **38 ms MAE**, closely matching manual alignment
- Precision supports **accurate timing** in child communication and eliminates labor-intensive waveform matching
- Real-time conversion lets transcripts sync with other modalities (e.g., Ubisense), allowing richer multimodal analyses

Importance: Enables scalable multimodal developmental research by aligning audio with location, movement, and contextual sensor data

Limitations: Reliance on a detectable synchronization tone, approximate manual identification of tone location

Future work: Fully automated tone detection, tone-free alignment using cross-correlation or spectral similarity, and extension to multi-device synchronization

ACKNOWLEDGEMENTS

The author gratefully acknowledges the Early Play and Development Lab at the University of Miami for providing the research environment and resources that supported this project. Special thanks to Dr. Daniel Messinger for his guidance and mentorship, and to Laura Vitale and Luis Angeles for their technical expertise and continued support. This material is based upon work supported by the National Science Foundation under Grant No. CNS-1949972. Any opinions, findings, and conclusions expressed are those of the author and do not necessarily reflect the views of the National Science Foundation.

REFERENCES

Gilkerson, J., Richards, J. A., Warren, S. F., et al. (2017). Mapping the early language environment using all-day recordings and automated analysis. American Journal of Speech-Language Pathology, 26(2), 248–265.

Radford, A., Kim, J. W., Xu, T., et al. (2023). Robust speech recognition via large-scale weak supervision. Proceedings of the 40th International Conference on Machine Learning.

Teranishi, M., Fujii, K., & Takeda, K. (2021). Quantitative analysis of spontaneous sociality in children's group behavior during nursery activity. PLOS ONE, 16(2), e0246041.

Echeverria-Huarte, I., Feliciani, C., Shi, Z., et al. (2026). Individual locomotor bias drives counterclockwise motion in pedestrian crowds. Nature Communications, 17, 4869.