AUTOMATED ALIGNMENT OF AUDIO DATA FOR DEVELOPMENTAL PSYCHOLOGY RESEARCH USING AUDIO FLAGS AND FORENSICS

Lemuel Mojica Vazquez[5], Elliot Huang[4], Kavya Jain[1], Vanessa Aguiar-Pulido[1], Laura Vitale[2], Daniel S. Messinger[2, 3]

Introduction

Advancements in audio recording technology have opened unprecedented opportunities to observe and analyze developmental interactions in naturalistic settings like classrooms(1). Traditionally, LENA recorders(2) (Fig. 1) with a 16 kHz sample rate were used(3), but the increasing need for higher audio quality and reduced noise has led to the adoption of Sony recorders (Fig. 2) with a 44.1 kHz sample rate(4).

Figure 1. Lena Recorder

Despite their advantages, the absence of an accurate real-time measure in Sony Recorders poses a significant challenge. Precise synchronization of audio data with real-time is crucial for aligning developmental interactions with timestamps, facilitating in-depth analysis, and enhancing the validity of research outcomes.

Figure 2. Sony Recorder

To overcome this challenge, *this poster presents an innovative approach to automating the alignment of audio data using audio flags and signal processing methods*. Audio flags, specific audio signals at known frequencies played for a short duration, serve as reference points for synchronizing the audio data from each recorder with real time. The positions of these flags within the waveform data are accurately identified using Fourier transform techniques.

Methods: Process

1. Audio Flag Generation
To facilitate calibration, carefully designed audio flags are introduced. Currently, they are set at 3000 kHz.

2. Real-Time Measurement
A central timekeeping system measures the timestamp to achieve precise real-time synchronization.

3. Flag Capture by Recorders
Each recorder deployed in the classroom captures the audio flag at the precise moment it is played.

4. Automated Alignment
The synchronization algorithm leverages Short-term Fourier Transformations (STFT) to detect the audio flags. STFT technique (Fig. 3) enables analysis of time-varying frequency components by applying Fourier Transform to short, overlapping windows of the audio signal. The transform was implemented with the assistance of the "librosa" Python package (3).

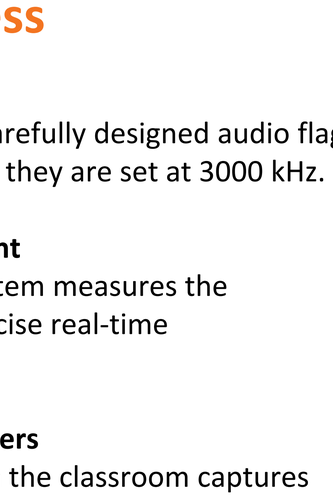

Figure 3. STFT Process

Methods: Algorithm

After performing a short-term Fourier transformation, the waveform data recorded by the Sony recorder is converted to three-axis data: time, frequency, and magnitude. When visualized with a spectrogram, the location of the flag can be quite easily identified visually (Fig. 4). However, reliably locating the flag by computational means proved less straightforward.

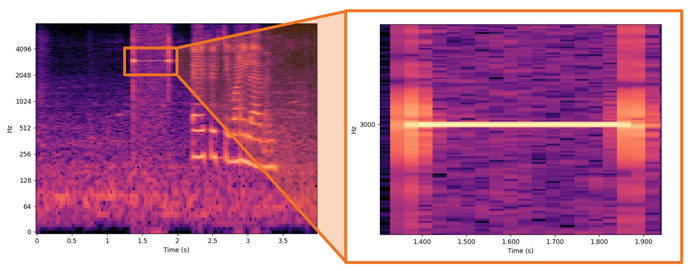

Figure 4. Flag STFT Spectrogram Visualization

The core mechanism by which we detect the presence of the flag is recognizing when the maximum magnitude of a time frame is at the flag frequency (3000 Hz). However, such detection is disrupted by outlier frequency gaps (Fig. 5) due to the fixed frequency nature of short-term Fourier transforms. This is a major obstacle, as it introduces a tradeoff between temporal localization and frequency precision(4).

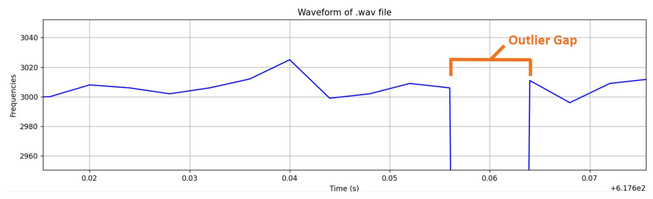

Figure 5. Graph of an outlier frequency gap.

A simple solution is to add a threshold in which detections with small enough gaps would be considered a continuous signal. Even so, this does not remove the possibility of outliers obstructing the beginning of the flag. Therefore, to minimize this error, we also optimized the window size.

Results

Through careful experimentation, a window size of 256 samples (0.006 sec) was identified as the ideal balance between accuracy and stability. Forensic analysis also revealed that gaps in continuity caused by outliers typically range from 0.004 to 0.012 seconds, so the threshold was set at 0.035 seconds. An offset of 100 Hz was also set to account for slight variability in flag frequency.

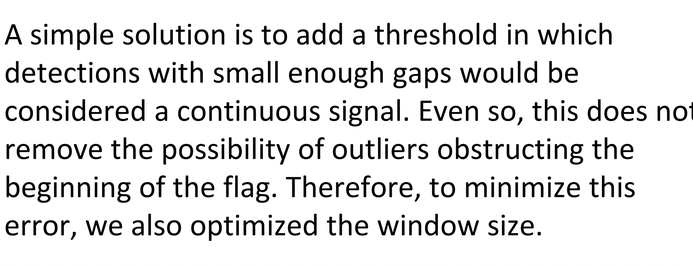

Figure 6. Graph of maximum magnitudes with flag

The algorithm developed can detect the presence and exact onset of flags (Fig. 6) at an accuracy of around 0.05 seconds, far exceeding the target accuracy of a tenth of a second. Current reliability tests have also not discerned any failure points, though further assessment is required.

After testing, the algorithm was packaged and integrated within the IBIS Lab data processing pipeline. Presently, the alignment can be completed with a single command.

Conclusion

The successful calibration of audio data marks a transformative milestone in the transition from Lena to Sony Recorders, and consequently developmental psychology research. The enhanced audio quality will empower the extraction of more nuanced insights from classrooms, enabling:

I. Enhanced analysis of vocalizations and language development
II. Elevated investigation of social interactions
III. Augmented examination of developmental disorders
IV. Comprehensive measure of behavior through integration with location tracking(3)
V. Increased reliability and validity of findings

Future Work

Whilst the development of the alignment system is largely complete, there are still a few areas for improvement:

I. Change of flag frequency to 8000 Hz – At 3000 Hz there is a greater chance of other sounds (children screaming) to interfere with the signal
II. Increased testing – The consistency of the algorithm has not been tested with large datasets

References

(1) Mitsven, S. G. et al. (2022). Objectively measured teacher and preschooler vocalizations: Phonemic diversity is associated with language abilities. Developmental Science, 25, e13177.

(2) LENA SP Digital Language Processor, LENA Foundation.

(3) Gilkerson, J. et al. (2008). Audio Specifications of the DLP-0121 (Technical Report LTR-03-02). LENA Foundation.

(4) Digital Voice Recorder, Model: ICD-UX570, SONY ELECTRONICS INC.

(5) McFee, Brian et al. (2023). librosa/librosa: 0.10.0 (0.10.0). Zenodo.

(6) Nasser Kehtarnavaz, CHAPTER 7 - Frequency Domain Processing, Digital Signal Processing System Design (Second Edition), Academic Press, 2008, Pages 175-196, ISBN 9780123744906

(7) Lynn K. Perry et al., Reciprocal patterns of peer speech in preschoolers with and without hearing loss, Early Childhood Research Quarterly, Volume 60, 2022, Pages 201-213, ISSN 0885-2006.

Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant No. CNS-1949972. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

1 Department of Computer Science, University of Miami, Coral Gables, FL, USA
2 Department of Psychology, University of Miami, Coral Gables, FL, USA
3 Linda Ray Intervention Center, University of Miami Miller School of Medicine, Miami, FL, USA
4 School of Computer Science, College of Computing, Georgia Institute of Technology, Atlanta, GA, USA
5 Department of Computer and Electrical Engineering, Ana G. Méndez University Gurabo, PR, USA