UVA DATA SCIENCE

ML Audio Classification of Parent-Infant Interactions

Gwen Powers, Will Sivolella, Elisabeth Waldron

May 1, 2024

Table of Contents

Purpose and Background

Purpose

- Distinguish between distinct vocal patterns of parents, infants, and other sounds through audio classification
- Provide a foundational step towards UVA STAR's mission to better inform early autism interventions in Latinx communities through vocal interaction analysis

Background

UVA STAR Initiative

- Research program at UVA School of Education and Human Development
- Focus on interdisciplinary research to improve autism interventions

LENA Technology

- Analyzes child vocalizations and conversational patterns
- High cost limits accessibility; our project aims to replicate its functions affordably

Specific to Latinx Communities

- Cultural nuances, language barriers, and limited awareness impact autism diagnosis
- Project focuses on culturally sensitive approaches and interventions

Scope

Data Overview

- Dataset consists of ~200 hours of audio recordings of infants (6 months to 2 years) from 32 different Lantinx families in central Virginia
- Recordings consist of back-and-forth conversations between infants and parents/caretakers as they go about daily activities at home
- Also provided with 10 hours of labeled audio data
- Target variable: infant or caretaker speaking
- Relevant predictors: frequencies present, amplitudes of audio

Assumptions

- Accurate pre-labeled data
- Minimal background noise
- Wide enough variety of voices

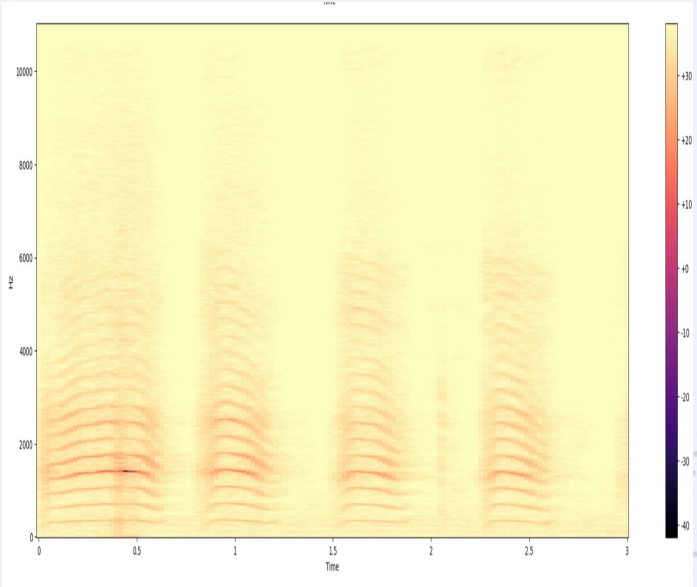
Limitations

- Storage + Computational Resources
- Ethical certification and handling of human data

Methodology

Data Preprocessing

- **Converted** mp3 audio files to csv and image formats to support varied analysis approaches.
- **Segmented** audio with metadata to ensure precise data alignment for training.
- **Extracted** key features (frequency and time) from audio using STFT and MFCC
- **Standardized** data shapes via padding of CSV and image files to ensure consistent processing



Frequency Plot Example

CNN Model Architecture

Chosen for its ability to effectively process and learn from image-based data

- **Base model:** Resnet50 pretrained model (frozen)
- **Top model:** additional pooling, dense and output layers (trainable)
- **Optimizer:** Adam (Adaptive Moment Estimation)
- **Criterion:** Categorical Cross Entropy
- **Pooling:** Global Average Pooling
- **Activation:** Relu
- **Data used: 3024** train files, **759** test files, all evenly split between 3 classes

LSTM Model Architecture

Chosen for its ability to bridge long-time lags in certain problems (also handles noise, distributed representations, continuous values, and overall bias)

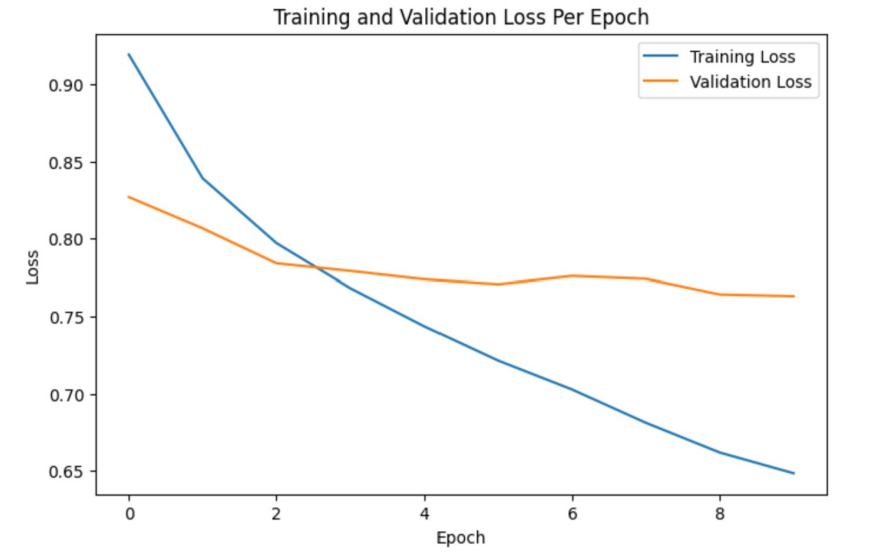
- **Optimizer:** Adam (Adaptive Moment Estimation)
- **Criterion:** Categorical Cross Entropy
- **Pooling:** Global Average Pooling
- **Activation:** Relu
- **Layers:** BatchNormalization, Dense (FC – vanish gradient), Dropout (bias)
- **Train files:** Adult - 120, Infant - 120, Background - 120

Model Enhancements

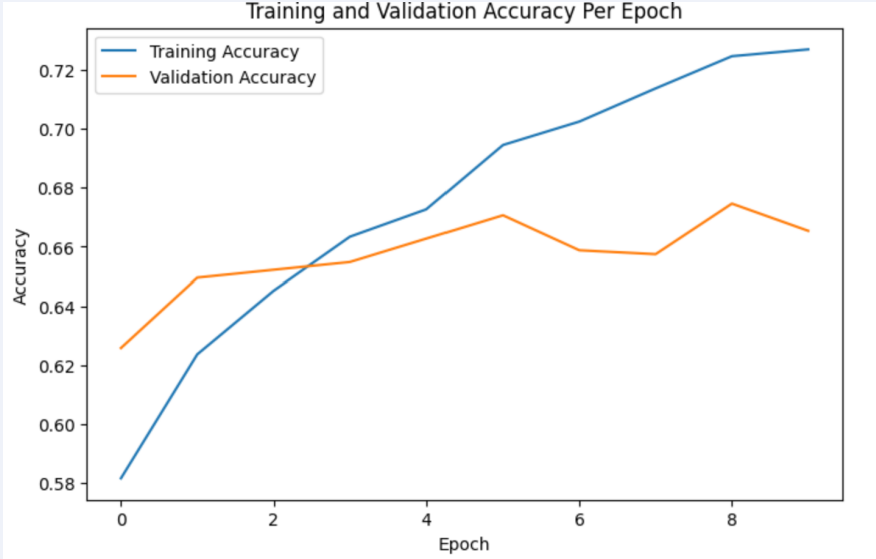
- **Train & Test Segmentation** using metadata
- **Hyperparameter Tuning:** LR, Optimizer, Criterion, Early Stopping (just for LSTM)
- **Statistical Analysis:** F1-Score, Confusion Matrix (under-sampling specific classifications)

Results

CNN Findings



Train and Validation Loss vs Epoch

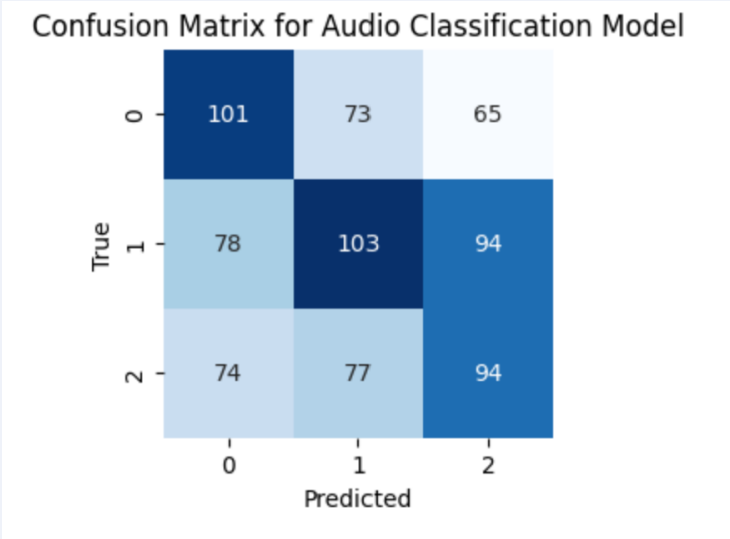


Train and Validation Accuracy vs Epoch

CNN Findings

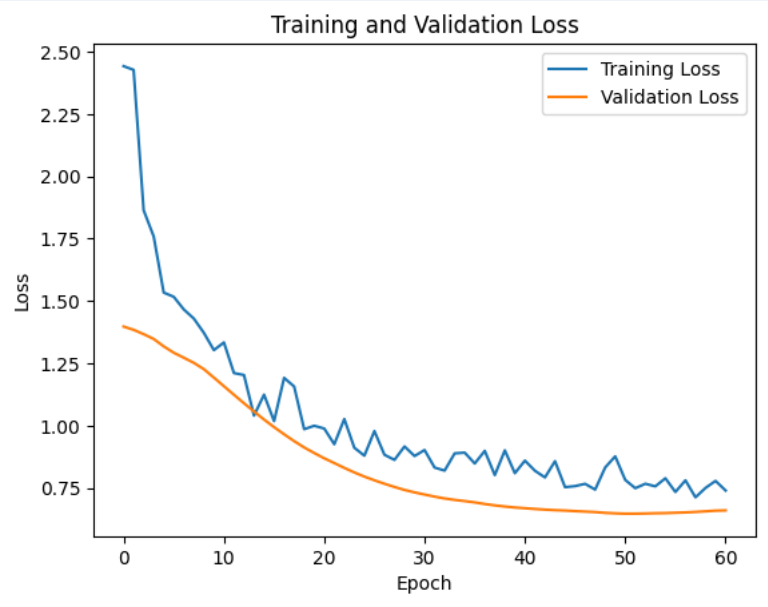
- **Accuracy: 67 %**
- **Loss: 0.76**
- **Weighted Average**

  F1-Score: 39 %

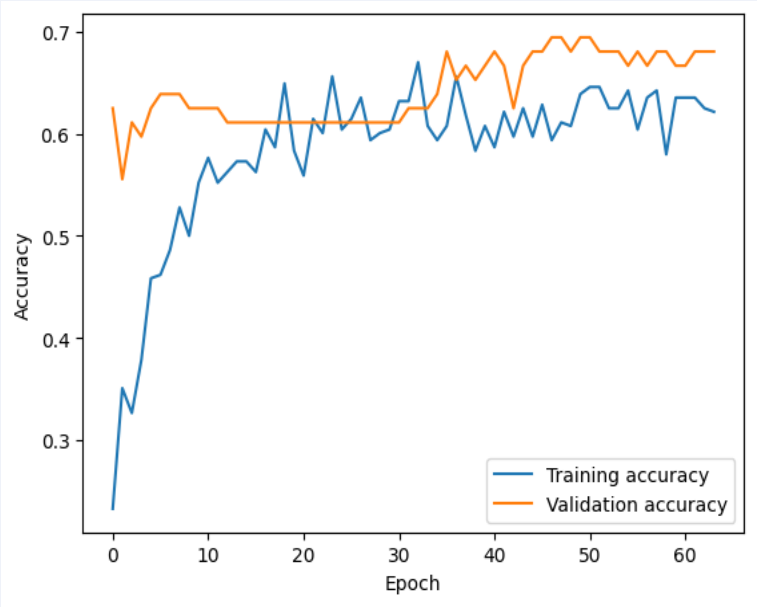


Confusion Matrix of Predictions

LSTM Findings



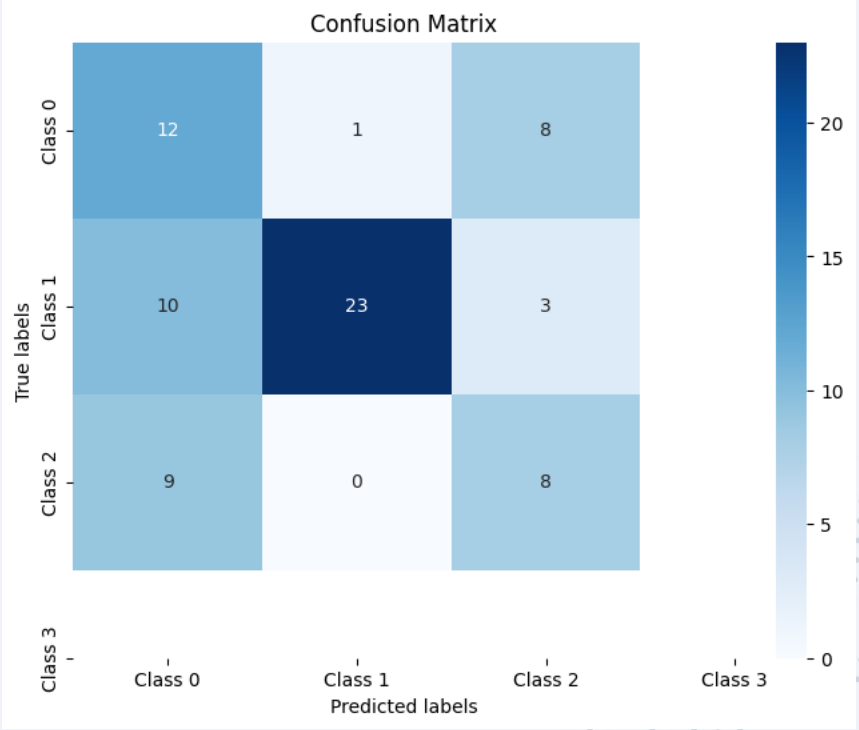
Train and Validation Loss vs Epoch



Train and Validation Accuracy vs Epoch

LSTM Findings

- **Accuracy: 69 %**
- **Loss: 0.65**
- **Weighted Average F1-Score: 60 %**



Final Thoughts

Takeaways

- Achieved similar accuracy to LENA and similar trained models (50 - 70 %)
- Importance of data recording quality (environment trade-offs – (un)controlled)
- Balance of computational resources with ethical considerations
- Using statistical analysis earlier in training

Applications and Future Work

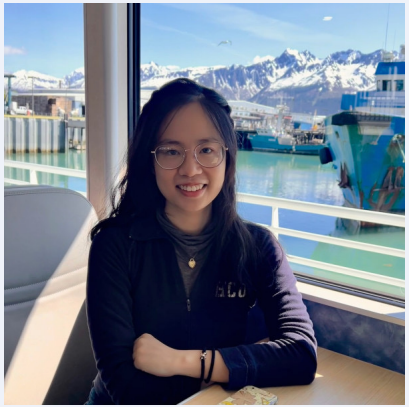
- Apply models for preliminary turn-taking analysis of vocal interactions
- Build a user-friendly interface that also provides special privileges for security
- Moving forward, our sponsor Professor DuBay will pass along our work to other students who will use it to aid in conversational analysis methods and research applications

Acknowledgements

MICHAELA DUBAY

SPONSOR

Assistant Professor of
Education and Human Development



AIYING ZHANG

FACULTY MENTOR

Assistant Professor
of Data Science

ADAM TASHMAN

PROFESSOR

Associate Professor
of Data Science

UVA DATA SCIENCE

Capstone Team

GWEN POWERS

WILL SIVOLELLA

ELISABETH WALDRON

UVA DATA SCIENCE

References

https://www.lena.org/

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7855224/

https://education.virginia.edu/research-initiatives/research-centers-labs/supporting-transformative-autism-research

https://www.geeksforgeeks.org/understanding-of-lstm-networks/