Towards Automatic Understanding of Parent-Child Interaction Patterns from Family Audio to Monitor Child Mental Health

Presenter: Jialu Li
ECE 590 Seminar, Spring 2023
4/5/2023
Problem Statement

• Mental, behavioral, and developmental disorders begin in early childhood, and such disorders often go undiagnosed

• Daily interactions with family members that are repeated and reinforced over time are critical to children’s emotional well-being

https://www.cdc.gov/childrensmentalhealth/data.html
Patterns of Parent-Infant Interactions

- **Attachment Theory**
  - The primary caregivers who are available and respond quickly and consistently to an infant's needs allow the child to develop a sense of security.
  - Children may develop **insecure attachment** styles if parents are often unavailable, intrusive, or respond inconsistently to the child’s cues, particularly signs of distress.

No response  \(\rightarrow\) Contingent vocalization  \(\rightarrow\) Co-vocalization  \(\rightarrow\) Turn-taking
Benefits of Turn-taking Interactions

- Parent/infant engaged in more positive interactions, including turn-taking in vocalizations, lead to better infant stress recovery at 6 and 9 months

Our Goals

- Develop robust **machine learning** models to automatically predict and analyze **parent-infant** interactions in infants’ **home environments**
  - **Speaker diarization** (who speaks when)
  - **Parent/Infant vocalization classifications** (type of vocalization for a given speaker)
LittleBeats™

- LittleBeats™ (LB) is a multimodal infant-wearable device that simultaneously records audio, movement of the child (IMU), and heart-rate variability of the child (ECG).

https://littlebeats.hdfs.illinois.edu/
Data Collection

- **Unlabeled data**
  - Child wears the LB or LENA during the day for 2-3 days

- **Labeled data**
  - A few of **10-mins** most vocalized segments are selected for each family to perform manual annotations based on a voice activity detector
  - 10% of collected audio data is double-coded to verify inter-rater reliability scores (all annotated categories have kappa scores > 0.8)

LENA (the Language Environment Analysis device): Another infant-wearable audio-only recording device
https://www.lena.org/
Annotations

- **Speakers**
  - CHN: target infant
  - FAN: adult female (mother)
  - MAN: adult male (father)
  - CXN: another child (sibling), if any

- **Infant vocalizations**
  - CRY: crying
  - FUS: fuss
  - BAB: babble

- **Parent (FAN/MAN) vocalizations**
  - CDS: child-directed speech, motherese
  - ADS: adult-directed speech
  - LAU: laugh
  - SNG: rhythmic, singing

- **Annotating 10 mins of family audio requires 2-3 hours of labor!**

A Praat window of sample labeled segments
# Data Distribution

<table>
<thead>
<tr>
<th>Device</th>
<th>Type</th>
<th>Domain</th>
<th>Context</th>
<th>Age</th>
<th># of families</th>
<th>Total dur (h)</th>
<th>Vocal dur (h)</th>
</tr>
</thead>
<tbody>
<tr>
<td>LB</td>
<td>Labeled</td>
<td>In</td>
<td>Home</td>
<td>&lt;14m</td>
<td>22</td>
<td>10.61</td>
<td>4.78h</td>
</tr>
<tr>
<td>LENA</td>
<td>Labeled</td>
<td>Out</td>
<td>Home</td>
<td>&lt;24m</td>
<td>30</td>
<td>14.59</td>
<td>9.05h</td>
</tr>
<tr>
<td>LB</td>
<td>Labeled</td>
<td>Out</td>
<td>Virtual Visit</td>
<td>&lt;10m</td>
<td>11</td>
<td>1.35</td>
<td>0.8h</td>
</tr>
<tr>
<td>Camera</td>
<td>Labeled</td>
<td>Out</td>
<td>Lab Visit</td>
<td>&lt;14m</td>
<td>105</td>
<td>9.94</td>
<td>3.4h</td>
</tr>
<tr>
<td>LB</td>
<td>Unlabeled</td>
<td>In</td>
<td>Home</td>
<td>&lt;5y</td>
<td>110</td>
<td>1100h</td>
<td>-</td>
</tr>
<tr>
<td>LENA</td>
<td>Unlabeled</td>
<td>Out</td>
<td>Home</td>
<td>&lt;5y</td>
<td>275</td>
<td>3200h</td>
<td>-</td>
</tr>
</tbody>
</table>
Wav2vec 2.0

- The latest cutting-edge self-supervised speech processing model developed by Facebook/Meta AI research.

Unsupervised pretraining

52k hour unlabeled adult speech

Supervised fine-tuning

Limited labeled adult speech

Automatic speech recognition testing

Wav2vec 2.0 on LittleBeats

- **Experimental Setup**
  - **W2V2-based** model is used
    - 12 transformer layers
    - 768 hidden feature dimension
  - Data sample is 2s intervals each for every 0.2s; Label is determined from majority of vocalization labels in centered 1s
  - **In-domain LB data partition**
    - **Training**: 8.67h from 15 families
    - **Development**: 1h from 3 families
    - **Testing**: 2h from 4 families

---

**Unsupervised pretraining**

**Supervised fine-tuning**

**Testing**

- **Unlabeled** family audio
- **Labeled** family audio
- **Labeled in-domain LB home audio**

**Speaker Diarization (SD)**

**Infant Vocalization Classifications (CHN)**

**Parent Vocalization Classifications (ADU)**
Unsupervised Pretraining Comparisons

Baseline architecture
- **MP**: mean pooling
- **WA**: weighted average
- **FFN**: feed-forward networks

<table>
<thead>
<tr>
<th>Pretrained unlabeled data</th>
<th>Duration (h)</th>
<th>SD</th>
<th>CHN</th>
<th>ADU</th>
<th>Avg</th>
</tr>
</thead>
<tbody>
<tr>
<td>Oracle</td>
<td>52k</td>
<td>.628</td>
<td>.549</td>
<td>.592</td>
<td>.590</td>
</tr>
<tr>
<td>LibriSpeech</td>
<td>960</td>
<td>.625</td>
<td>.465</td>
<td>.535</td>
<td>.542</td>
</tr>
<tr>
<td>LB</td>
<td>1100</td>
<td>.682</td>
<td>.663</td>
<td>.630</td>
<td>.658</td>
</tr>
<tr>
<td>LB+LENA</td>
<td>4300</td>
<td>.715</td>
<td>.689</td>
<td>.669</td>
<td>.691</td>
</tr>
</tbody>
</table>

F1-scores (%) for SD, CHN, ADU, and the average of three tiers fine-tuned on in-domain labeled LB data only
Effects of Adding Out-of-domain Data

### Binary-domain learning
- **Domain emb**: One-hot learnable embeddings
- **Domain multi-task learning**

### F1-scores (%) trained on both in- and out-of-domain data

<table>
<thead>
<tr>
<th>Features Used</th>
<th>Domain Learning</th>
<th>SD</th>
<th>CHN</th>
<th>ADU</th>
<th>Avg</th>
</tr>
</thead>
<tbody>
<tr>
<td>Layer 12</td>
<td>-</td>
<td>68.7</td>
<td>68.1</td>
<td>67.6</td>
<td>68.1</td>
</tr>
<tr>
<td>All layers</td>
<td>-</td>
<td>68.9</td>
<td>69.6</td>
<td>70.6</td>
<td>69.7</td>
</tr>
<tr>
<td>Layer 12 One-hot</td>
<td></td>
<td>70.6</td>
<td>70.4</td>
<td>71.9</td>
<td>71.0</td>
</tr>
<tr>
<td>All layers One-hot</td>
<td></td>
<td>69.6</td>
<td>70.6</td>
<td>68.7</td>
<td>69.6</td>
</tr>
<tr>
<td>Layer 12 Multi-task</td>
<td></td>
<td>66.1</td>
<td>68.3</td>
<td>69.7</td>
<td>68.0</td>
</tr>
<tr>
<td>All layers Multi-task</td>
<td></td>
<td>68.4</td>
<td>68.6</td>
<td>67.3</td>
<td>68.1</td>
</tr>
</tbody>
</table>
Effects of Adding Speaker Embeddings

ECAPA-TDNN Speaker Embedding
- **ET emb**: ECAPA-TDNN speaker embeddings
- **ECAPA-TDNN F1-scores (%)** on LB testing data using training data from
  - In-domain: 66.6
  - In- + Out-of-domain: 67.7

<table>
<thead>
<tr>
<th>Data</th>
<th>model</th>
<th>ET emb</th>
<th>SD</th>
<th>CHN</th>
<th>ADU</th>
<th>Avg</th>
</tr>
</thead>
<tbody>
<tr>
<td>In</td>
<td>LB1100h</td>
<td>-</td>
<td>70.8</td>
<td>66.6</td>
<td>60.0</td>
<td>65.8</td>
</tr>
<tr>
<td>In</td>
<td>LB1100h</td>
<td>In</td>
<td>73.2</td>
<td>68.4</td>
<td>62.9</td>
<td>68.2</td>
</tr>
<tr>
<td>In+Out</td>
<td>LL4300h</td>
<td>-</td>
<td>68.9</td>
<td>69.6</td>
<td>70.6</td>
<td>69.7</td>
</tr>
<tr>
<td>In+Out</td>
<td>LL4300h</td>
<td>In+Out</td>
<td>70.2</td>
<td>69.6</td>
<td>70.8</td>
<td>70.2</td>
</tr>
</tbody>
</table>

F1-scores (%) with and w/o ECAPA-TDNN embeddings

## Effects of Data Augmentation

<table>
<thead>
<tr>
<th>Data Augmentation</th>
<th>noise</th>
<th>domain</th>
<th>ET emb</th>
<th>SD</th>
<th>CHN</th>
<th>ADU</th>
<th>Avg</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>SpecAug</strong></td>
<td></td>
<td></td>
<td></td>
<td>68.9</td>
<td>69.6</td>
<td>70.6</td>
<td>69.7</td>
</tr>
<tr>
<td>Random audio chunks dropping</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>speed perturbation</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Environmental Speech Corruption</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reverberation</td>
<td></td>
<td></td>
<td></td>
<td>76.5</td>
<td>69.2</td>
<td>67.7</td>
<td>71.1</td>
</tr>
<tr>
<td>Room impulse responses</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Additive Noise</strong></td>
<td></td>
<td></td>
<td></td>
<td>77.2</td>
<td>70.8</td>
<td>71.5</td>
<td>73.2</td>
</tr>
<tr>
<td>MUSAN (a variety of noises)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CHiME-Home (domestic noises)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Reverberation + Noise</strong></td>
<td></td>
<td></td>
<td></td>
<td>75.2</td>
<td>71.8</td>
<td>67.8</td>
<td>71.6</td>
</tr>
<tr>
<td>M(SD)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>M(CHN, ADU)</td>
<td></td>
<td></td>
<td></td>
<td>76.2</td>
<td>72.0</td>
<td>68.7</td>
<td>72.3</td>
</tr>
<tr>
<td>M(All)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>M(SD)</td>
<td></td>
<td></td>
<td></td>
<td>77.4</td>
<td>68.5</td>
<td>64.3</td>
<td>70.1</td>
</tr>
<tr>
<td>M(SD)</td>
<td></td>
<td>One-hot</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>M(SD)</td>
<td></td>
<td>One-hot</td>
<td>In+Out</td>
<td>78.2</td>
<td>70.1</td>
<td>68.8</td>
<td>72.4</td>
</tr>
</tbody>
</table>


Future Work

- Investigate **active learning** to improve W2V2

  - Select the most uncertain samples from unlabeled data
  - W2V2
  - Human coders annotate selected data

- Automatically analyze important parent-child interaction patterns
  - **Coordinated**: positive affect, turn-taking
  - **Uncoordinated**: parent unresponsive, parent or child interrupt
Acknowledgement

- This work is supported by Beckman Graduate Fellowship.
- Thank YOU to my supervisors, collaborators, project manager, graduate RAs, undergraduate RAs, and anyone who have helped me conducting research on this amazing project!

Professor Mark Hasegawa-Johnson, Electrical and Computer Engineering
Professor Nancy McElwain, Human Development and Family Studies (HDFS)
Professor Bashima Islam, Worcester Polytechnic Institute
https://littlebeats.hdfs.illinois.edu/team/
Yannan Hu, Graduate student at HDFS
Jordan Bodway, Project Manager
Lily Gaskin, LittleBeats Coordinator
Thank you for listening. Questions?