Infant Crying Detection in Real-World Environments

Xuewen Yao¹, Megan Micheletti², Mckensey Johnson², Edison Thomaz¹, Kaya de Barbaro²

¹Department of Electrical and Computer Engineering, University of Texas at Austin, USA
²Department of Psychology, University of Texas at Austin, USA

Datasets

- We collected 780 hours of raw audio data using LENA in real-world home environments.
- Real-world: Filtered Dataset (RW-Filter)
  - Filtered using algorithms from LENA software
- Real-world: Unfiltered 24h Dataset (RW-24h)
  - Unfiltered, randomly sampled audio data for testing only
- In-lab (IL-CRIED)
  - CRIED database (microphones over awake infants in a cot in a quiet room)
  - 5587 individual vocalisations of 20 healthy infants
  - Vocalizations: infant neutral/positive, fussing, crying, and overlapping adult vocalizations
- In summary, we have three audio datasets:

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Cys Hrs</th>
<th>Total Hrs</th>
<th>N</th>
<th>Ages (months)</th>
</tr>
</thead>
<tbody>
<tr>
<td>RW-Filter</td>
<td>7.9</td>
<td>66</td>
<td>24</td>
<td>1.53 - 10.8</td>
</tr>
<tr>
<td>RW-24h</td>
<td>14.7</td>
<td>408</td>
<td>17</td>
<td>0.78 - 7.03</td>
</tr>
<tr>
<td>IL-CRIED</td>
<td>1.26</td>
<td>12</td>
<td>20</td>
<td>1.2 - 4</td>
</tr>
</tbody>
</table>

Preprocessing

- Training:
  - Windowing: 5 second windows (w/ 4-second overlap)
  - Augmentation using time masking and deformation technique
- Testing:
  - Removed all audio segments silent above a 350 Hz threshold
  - Windowing: 5 second windows (w/ 4-second overlap)

Crying Detection Models and Results

<table>
<thead>
<tr>
<th>SVM with acoustic features (AF)</th>
</tr>
</thead>
<tbody>
<tr>
<td>34 acoustic features</td>
</tr>
<tr>
<td>SVM classifier with RBF kernel</td>
</tr>
<tr>
<td>End-to-end CNN model (CNN)</td>
</tr>
<tr>
<td>Modified AlexNet with mel-scaled spectrograms as input</td>
</tr>
<tr>
<td>SVM with deep spectrum and acoustic features (DSF + AF)</td>
</tr>
<tr>
<td>Combination of AF and CNN</td>
</tr>
<tr>
<td>Last hidden layer of CNN (size 1000) used as deep spectrum features</td>
</tr>
</tbody>
</table>

- DSF + AF is the best performing model for real-world datasets.
- DSF + AF reaches F1 score 0.613 when trained and tested on real-world datasets.
- End-to-end CNN training contributed most substantially to the DSF + AF model's performance.

Discussion

- Real-world vs. In-lab training data
  - Datasets collected in controlled environments do not represent the full complexity of real-world environments
  - Models trained on in-lab data are of limited use in the context of the real-world crying detection task

We found DSF + AF performed substantially better than LENA's cry classifier in assessment scenarios important to developmental researchers [5].

Acknowledgements

We thank Katrin D. Bartl-Pokorny, Christa Einspieler, Peter B. Marschik, Florian B. Pokorny, and Dajie Zhang of Medical University of Graz, Austria for providing us with CRIED database.

References