Infant Crying Detection in Real-World Environments

Xuewen Yao[1], Megan Micheletti[2], Mckensey Johnson[2], Edison Thomaz[1], Kaya de Barbaro[2]
[1]Department of Electrical and Computer Engineering, University of Texas at Austin, USA
[2]Department of Psychology, University of Texas at Austin, USA

Introduction

- Infant crying is a critical signal for communication and a known parental stressor.
- Many researchers have tried to detect crying, and it appears the models do well [1].
    - Previous crying models either were developed and evaluated using data in controlled settings or trained and evaluated on short, preparsed segments containing non-overlapping individual sound.
- Detection and classification in real-world settings is much harder than clean-lab conditions, such as in cough [2] and laughter [3] detections

Contribution

- We collected and annotated a real-world infant crying dataset
    - https://homebank.talkbank.org/access/Password/deBarbaroCry.html
- We developed a robust crying detection model in real-world
    - F1 score: 0.613 (Precision: 0.672, Recall: 0.552)
    - https://github.com/AgnesMayYao/Infant-Crying-Detection
- We concluded that In-lab crying dataset does not generalize to real-world situations
    - Trained on in-lab, tested on In-lab F1 score: 0.656
    - Trained on in-lab, tested on real-world F1 score: 0.236

Datasets

- We collected 780 hours of raw audio data using LENA in real-world home environments.
- Real world: Filtered Dataset (RW-Filt)
    - 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 1. Crying Dataset Statistics

Dataset	Cry Hrs	Total Hrs	N	Ages (months)
RW-Filt	7.9	66	24	1.53 - 10.8
RW-24h	14.7	408	17	0.78 - 7.03
IL-CRIED	1.26	14	20	1 - 4

- Annotation
    - At level of crying episodes according to the best practice in behavioral science
    - Include both fussing and crying vocalizations
    - Inter-rater reliability kappa score: 0.85 (strong agreement)

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

Crying Detection Models and Results

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

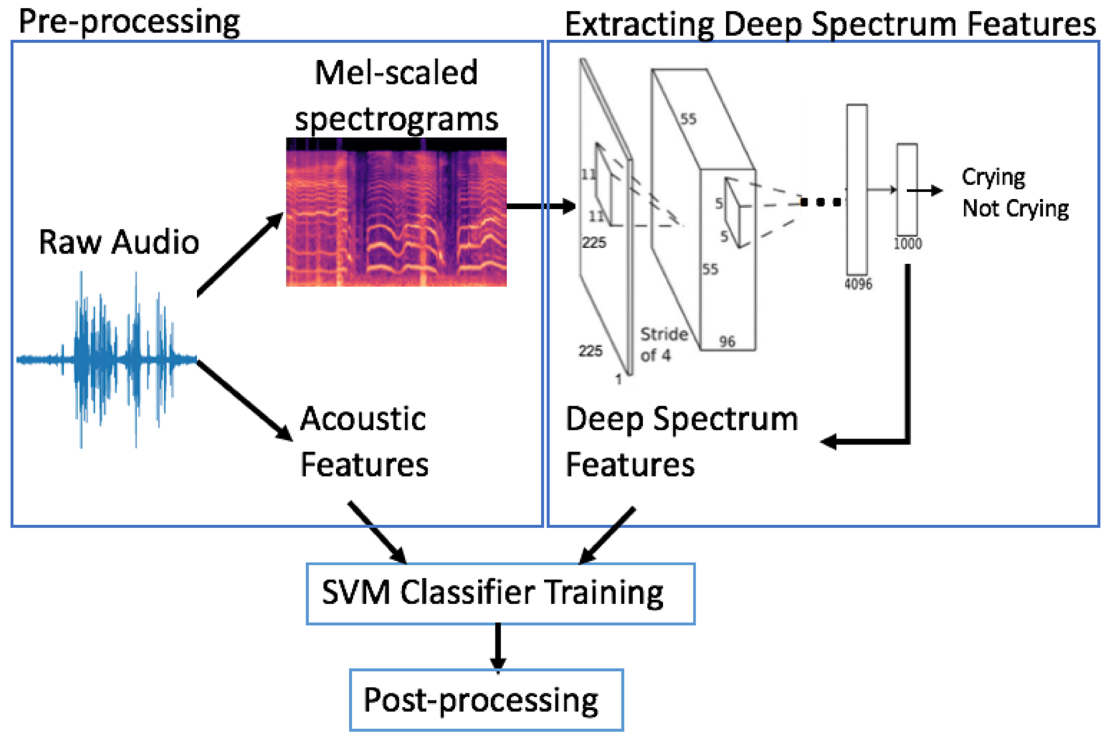

Table 2. Infant cry detection performance on both real-world and in-lab dataset, with second-by-second accuracy averaged across participants.

	Results on RW-Filt (LOPO)			Results on RW-24h		
Train on RW-Filt	F1	Precision	Recall	F1	Precision	Recall
AF	0.515(±0.185)	0.42(±0.225)	0.847(±0.140)	0.502(±0.204)	0.481(±0.239)	0.586(±0.191)
CNN	0.620(±0.182)	0.505(±0.206)	0.873(±0.110)	0.589(±0.194)	0.642(±0.217)	0.580(±0.178)
DSF + AF	0.615(±0.170)	0.521(±0.191)	0.820(±0.147)	0.613(±0.184)	0.672(±0.219)	0.552(±0.178)
VGGish	0.574(±0.204)	0.445(±0.216)	0.936(±0.062)	0.543(±0.204)	0.489(±0.228)	0.652(±0.182)
Train on IL-CRIED	**Results on IL-CRIED (LOPO)**			**Results on RW-24h**		
DSF + AF	0.656(±0.191)	0.578(±0.255)	0.808(±0.128)	0.236(±0.122)	0.143(±0.084)	0.851(±0.162)

- 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

[1]. C. Ji, T. B. Mudiyanselage, Y. Gao, and Y. Pan, "A review of infant cry analysis and classification," *EURASIP Journal on Audio, Speech, and Music Processing*, no. 8, 2021.
[2]. D. Liaqat, S. Liaqat, J. L. Chen, T. Sedaghat, M. Gabel, F. Rudzicz, and E. de Lara, "Coughwatch: Real-world cough detection using smartwatches," in *ICASSP 2021*, 2021, pp.8333–8337.
[3]. J. Gillick, W. Deng, K. Ryokai, and D. Bamman, "Robust Laughter Detection in Noisy Environments," in *Proc. Interspeech 2021*, 2021, pp. 2481–2485.
[4]. P. Marschik, F. Pokorny, R. Peharz, D. Zhang, J. O'Muircheartaigh, H. Roeyers, S. B ̈olte, A. Spittle, B. Urlesberger, B. Schuller, L. Poustka, S. Ozonoff, F. Pernkopf, T. Pock, K. Tammimies, C. Enzinger, M. Krieber, I. Tomantschger, K. Bartl-Pokorny, J. Sigafoos, L. Roche, G. Esposito, M. Gugatschka, K. Nielsen-Saines, C. Einspieler, W. Kaufmann, and The BEE-PRI Study Group, "A novel way to measure and predict development: A heuristic approach to facilitate the early detection of neurodevelopmental disorders," *Current Neurology and Neuroscience Reports*, vol. 17, no. 5, pp. 43, Apr 2017.
[5]. M. Micheletti, X. Yao, M. Johnson, and K. de Barbaro, "Validating a Model to Detect Infant Crying from Naturalistic Audio," *Behavior Research Methods* (In Review).