IEEE ICASSP 2022

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 their models do well [1].
- However, previous models are typically developed and evaluated with "clean-lab" data
  - Controlled settings
  - Short, preparsed segments containing non-overlapping individual cry sounds.

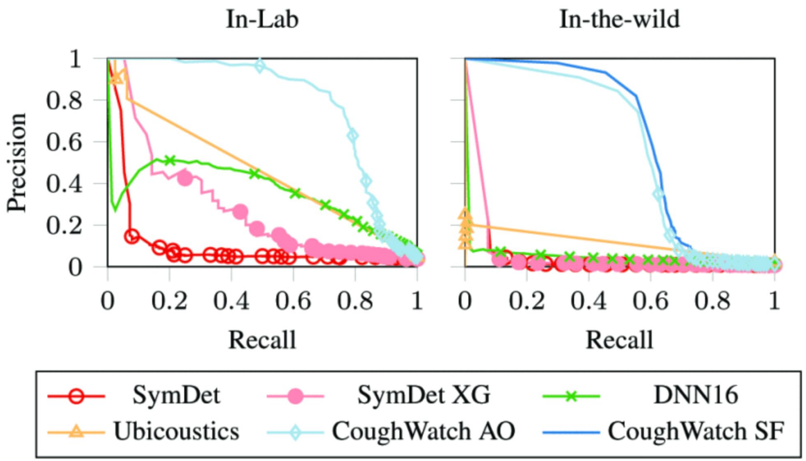
First Author	Dataset	Features	Classifiers	Best Performance
Chang (2019)	Self-recorded (Crying with TV, Speech, etc.)	Spectrogram	CNN	99.83%
Manikanta (2019)	Recorded in homes (Crying with AC, Fan, etc.)	MFCC	1D-CNN FFNN SVM	86%
Dewi (2019)	Self-recorded samples Cry and Not Cry	LFCC	KNN	90%
Gu (2018)	Self-recorded (Crying with laughter, barking, etc.)	LPC	Dynamic time warping algorithm	97.1%
Ferretti (2018)	Real Dataset: recorded in the NICU of a hospital. Synthetic DB: Crying with speech, "beep" sounds, etc.)	Log-Mel Coefficients	DNN	Real dataset 86.58% Synthetic DB 92.92%
Feier (2017)	TUT Rare Sound Events 2017 (Crying with "glass breaking", "gunshot", etc.)	log-amplitude mel-spectrogram	CRNN	85% for baby crying detection

- Thus, their results may not generalize to real-world contexts in which they are most needed

[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.

Introduction

- Detection and classification in real-world settings is much harder than clean-lab conditions
  - E.g. real-world cough [2] and laughter [3] detections



- Training on Switchboard (Precise segmentations, but out-of-domain)

Method	Results on Switchboard Test Data (F1)	Results on New AudioSet Test Data (F1)
Baseline (Featurized MLP)	0.688	0.359
Resnet	0.747	0.573
Resnet + Augmentation	0.752	**0.608**

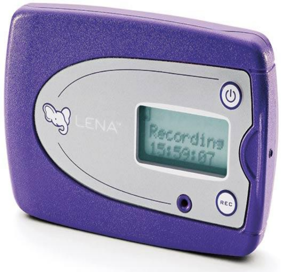
[2]. D. Liaqat, S. Liaqat, J. L. Chen, T. Sedaghat, M. Gabel, F. Rudzicz, and E. de Lara, “Coughwatch: Real-world cough detection using smartwatches,” in *ICASSP 2021*, 2021, pp.8333–8337.
[3]. J. Gillick, W. Deng, K. Ryokai, and D. Bamman, “Robust Laughter Detection in Noisy Environments,” in *Proc. Interspeech 2021*, 2021, pp. 2481–2485.

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

Two novel audio 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

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

One existing audio dataset

- **In-lab (IL-CRIED)**
  - CRIED database published by Marschik et al [4]
  - Microphones over infants in a cot in a quiet room
  - 5587 individual vocalisations from 140 recordings of 20 healthy infants
  - Vocalizations: infant neutral/positive, fussing, crying, and overlapping adult vocalizations
  - Re-annotated to match our real-world datasets

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

[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.

Model development

- Use real-world RW-Filt data to train a set of three models*
  - Test the performance on RW-Filt and RW-24h (raw, unfiltered)
  - Determine the best performing model
- Use lab-clean IL-CRIED data to train the best performing model
  - Test and compare the performance on lab-clean IL-CRIED and real-world RW-24h

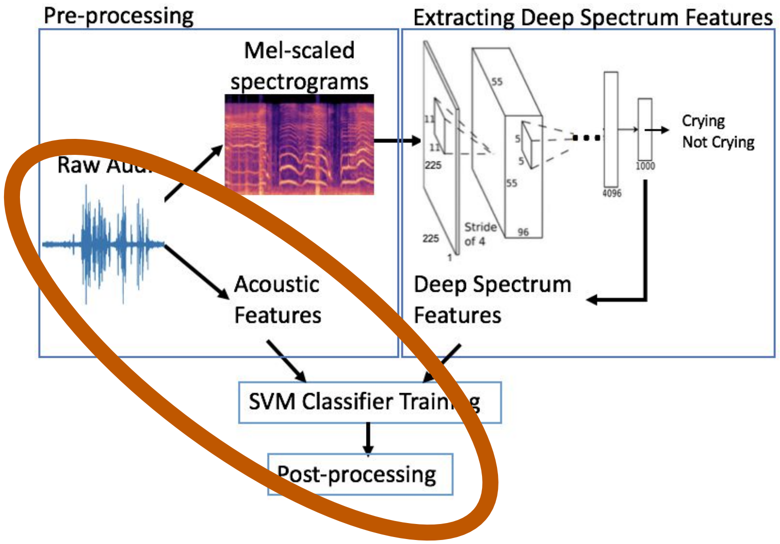
Preprocessing

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

*technically four models -see paper for details

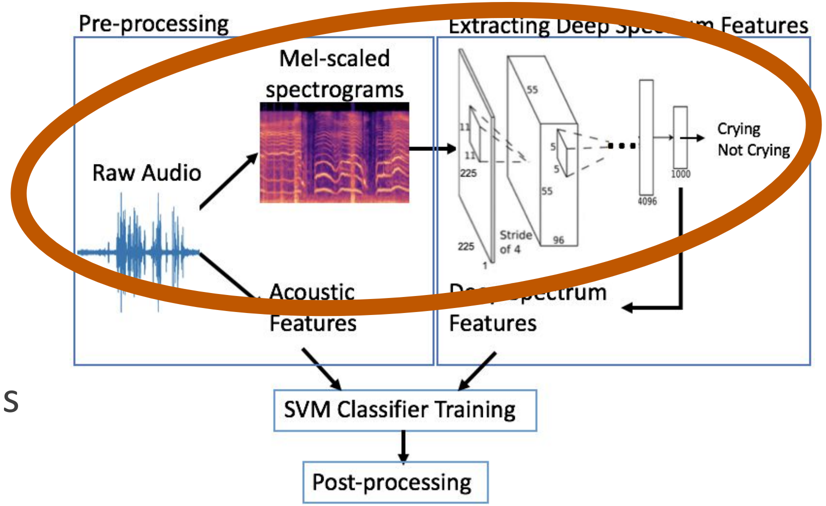
Crying detection models

- **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



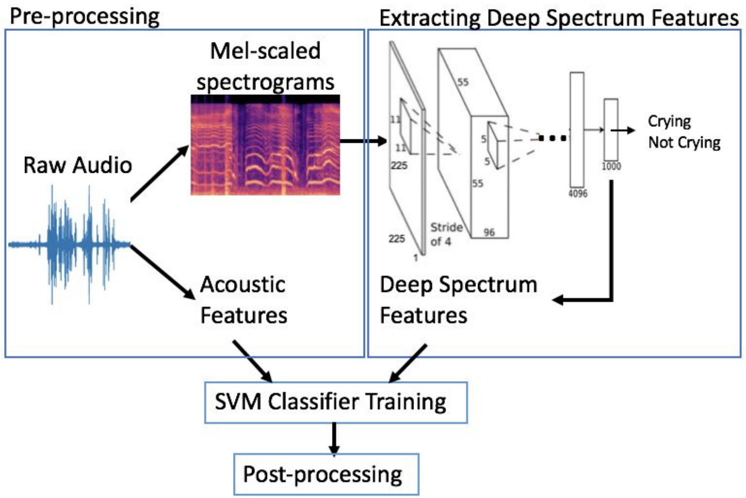
Crying detection models

- 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



Crying detection models

- 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**



Results

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

Discussion: real-world vs. in-lab datasets

- 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

- In other work, we tested DSF + AF in assessment scenarios important to developmental researchers against LENA's cry classifier
  - Our model has substantially higher accuracy metrics (recall, F1, kappa)
  - And stronger correlations with human annotations at all timescales tested (24 hours, 1 hour, and 5 minutes) relative to LENA [5].

[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).