Baby Cry Detection in Domestic Environment using Deep Learning

Yizhar Lavner*, Rami Cohen†, Dima Ruinskiy*‡ and Hans IJzerman§
*Dept. of Computer Science, Tel-Hai College, Upper Galilee, Israel
†Andrew and Erna Viterbi Faculty of Electrical Engineering, Technion – Israel Institute of Technology, Haifa, Israel
‡Intel Corporation, Haifa, Israel
§Dept. of Clinical Psychology, Vrije Universiteit, Amsterdam, the Netherlands
Email: yizhar.lavner@gmail.com, rc@campus.technion.ac.il, dima.ruinskiy@intel.com, h.ijzerman@gmail.com

***Abstract*—Automatic detection of a baby cry in audio signals is an essential step in applications such as remote baby monitoring. It is also important for researchers, who study the relation between baby cry patterns and various health or developmental parameters. In this paper, we propose two machine-learning algorithms for automatic detection of baby cry in audio recordings. The first algorithm is a low-complexity logistic regression classifier, used as a reference. To train this classifier, we extract features such as Mel-frequency cepstrum coefficients, pitch and formants from the recordings. The second algorithm uses a dedicated convolutional neural network (CNN), operating on log Mel-filter bank representation of the recordings. Performance evaluation of the algorithms is carried out using an annotated database containing recordings of babies (0-6 months old) in domestic environments. In addition to baby cry, these recordings contain various types of domestic sounds, such as parents talking and door opening. The CNN classifier is shown to yield considerably better results compared to the logistic regression classifier, demonstrating the power of deep learning when applied to audio processing.**

I. Introduction

Automatic detection and classification of acoustic events in audio signals is a challenging research area in auditory machine perception [1], related to computational auditory scene analysis [2]. Due to the vast amount of acoustic data collected and accumulated in recent years, manual annotation of the data is impractical. This raises the need for developing reliable and efficient algorithms for automatic detection and classification of acoustic events. Such algorithms are a pre-requisite for automatic recognition and labeling of audio content.

In this study, we focus on the detection and classification of baby cry sounds in various domestic environments. Crying is one of the major means of infants to communicate distress and attachment needs (such as being hungry or cold) to their caregivers [3]. One common application of automatic cry detection is a baby remote monitor, where parents are alerted if their baby is crying. Another important application is enablement of non-intrusive psychological research of infants and their caregivers in the earliest days of life. The bond between caregiver and infant is formed through physiological co-regulation processes that take place throughout the day [4]. Thus, monitoring often needs to be conducted over many hours or days to collect meaningful data. The sheer volume of data and the difficulty to decide a priori which variables to target make precise measurements and classification of acoustic events necessary to understand the co-regulation patterns.

A baby cry is elicited from rhythmical transitions between inhalation and exhalation, due to a vibration of the vocal cords that produces periodic air pulses. The period of these pulses is called the fundamental frequency (pitch), and its typical values in healthy babies are $250-600$ Hz. The cry signal is shaped by the vocal tract, leading to resonant frequencies termed as *formants*. The first two formants occur typically around 1100 Hz and 3300 Hz, respectively [3]. The detection of cry signals is usually carried out by extracting distinguishing features from segments of the audio signal. Apart from pitch and formants, these include temporal and spectral features such as short-time energy, Mel-frequency cepstrum (MFC) coefficients and others [5]–[7].

In this work, we implement two methods for the detection of cry signals in audio recordings: a low-complexity classifier based on *logistic regression* and a *convolutional neural network* (CNN) classifier, and compare their performance.

II. Methods

A. Database

The database for this study contains recordings of several tens of hours of audio recordings made by parents of babies in the Netherlands. The babies were in their first 6 months of life, and were recorded 24/7 in a domestic environment. The recordings contain various types of sounds, such as crying, parents talking, door opening etc. The database was collected as a part of a pilot study aimed at investigating ”attachment formation” (forming the bond between caregiver and child) [8]. Three hours of the recordings were fully annotated, down to the millisecond level, with about 50 different event types. The sampling frequency of the recordings is $F_s = 44100$ Hz.

B. Preprocessing and feature extraction

The audio recordings are divided into consecutive overlapping segments of 4096 samples (about 93ms) with an overlap of 50%. These segments are further divided into frames of 16ms with a step size of 8ms. A pitch detector [9] is applied to each frame, using peaks in the cepstral domain for rough pitch

978-1-5090-2152-9/16/$31.00 ©2016 IEEE

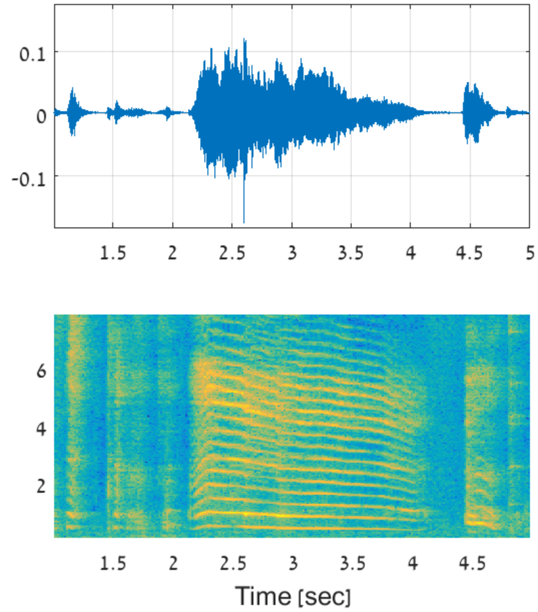

Fig. 1: An example of a baby cry signal. Top: the signal waveform. Bottom: the signal spectrogram, demonstrating the harmonic structure of the cry signal.

value estimation, and cross-correlation in the time-domain for refinement of the initial pitch value. Possible pitch period durations are restricted to the range of $1.6 - 3.3$ms due to the expected baby cry pitch period.

The following features are computed for each audio segment. The reader is referred to [7] and [10] for details.

1) 38 Mel-Frequency Cepstrum coefficients (MFCC).
2) Short-time energy (STE).
3) Zero-crossing rate (ZCR).
4) Pitch median value within a segment.
5) Run-length of pitch, defined as the number of consecutive voiced frames within a segment where pitch was detected.
6) Harmonicity factor (HF).
7) Harmonic-to-average power ratio (HAPR).
8) First formant, based on the line-spectral pair representation.
9) Band energy ratio, defined as the ratio (in dB) between the spectral energy in the frequency bands $[0, 3.5]$kHz and $[3.5, 22.5]$kHz.
10) Spectral rolloff point: the frequency below which 75% of the spectral energy is concentrated.

Figure 2 shows an example of the distribution of the 5th MFC coefficient among baby cry sections (red) vs. all other sound events (blue) in the training set (about 320 seconds). The discriminating potential of this feature is evident, although there is a wide overlapping area.

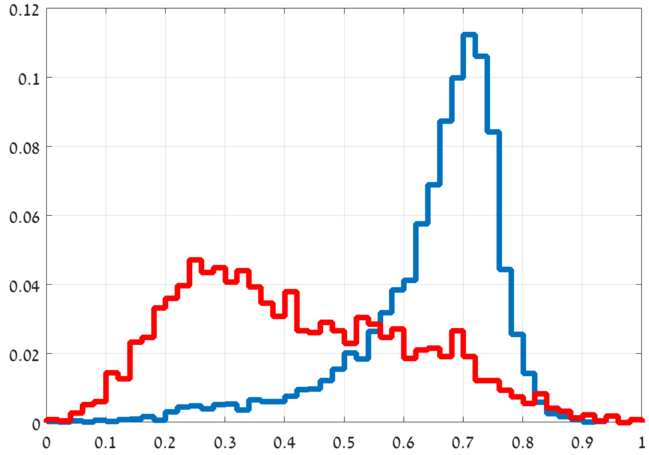

Fig. 2: A histogram of the 5th MFC coefficient. Red: cry events, blue: other events.

III. Logistic regression

The *logistic regression* classifier [11] is a simple supervised algorithm, with the advantage of low computational complexity. The logistic regression is a non-linear hypothesis function of the form:

$$h_{\boldsymbol{\theta}}(\boldsymbol{x}) = \frac{1}{1+\exp\left(-\boldsymbol{\theta}^T\boldsymbol{x}\right)}, \tag{1}$$

where $\boldsymbol{x}$ is a d-dimensional feature vector and $\boldsymbol{\theta}$ is a weight vector. In our case, $h_{\boldsymbol{\theta}}(\boldsymbol{x}) \in (0,1)$ predicts the likelihood of a segment to be a cry sound (values close to 1), or a different sound (values close to 0). The decision is made by comparing $h_{\boldsymbol{\theta}}(\boldsymbol{x}) \in (0,1)$ to a threshold value, to obtain a final binary classification $y \in \{0,1\}$, where 1 denotes a cry event. In the training phase of the classifier, a gradient descent algorithm is used to find $\boldsymbol{\theta}$ that minimizes the *cost function*

$$\begin{aligned} \mathbb{E}(\boldsymbol{\theta}) = &-\frac{1}{n}\sum_{j=1}^{n} y^{(j)} \log\left(\frac{1}{1+\exp(-\boldsymbol{\theta}^T\boldsymbol{x}^{(j)})}\right) \\ &-\frac{1}{n}\sum_{j=1}^{n}\left(1-y^{(j)}\right)\log\left(\frac{\exp(-\boldsymbol{\theta}^T\boldsymbol{x}^{(j)})}{1+\exp(-\boldsymbol{\theta}^T\boldsymbol{x}^{(j)})}\right) \\ &+\frac{\lambda}{2n}\sum_{k=1}^{d}{\theta_k}^2, \end{aligned} \tag{2}$$

given a dataset of n labeled samples $\left\{\boldsymbol{x}^{(j)}, y^{(j)}\right\}_{j=1}^{n}$, where λ is a regularization parameter. The $\boldsymbol{\theta}$-minimizer found by the gradient descent algorithm is then assigned to (1) to classify new unlabeled samples.

A. Detection procedure

A schematic block diagram of the logistic-regression-based algorithm is depicted in Figure 3. The input data is divided into consecutive segments of 4096 samples. For each segment a 50-dimensional feature vector is computed. The trained regularized logistic regression is then applied on each feature vector, and the hypothesis function $h_{\boldsymbol{\theta}}(\boldsymbol{x})$ is obtained,

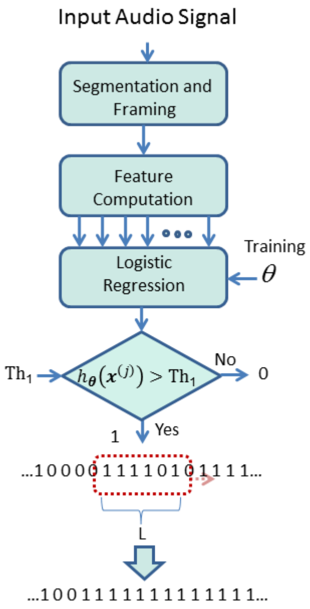

Fig. 3: A schematic block diagram of the logistic regression algorithm

representing an estimation of the posterior probability $p(y|\boldsymbol{x})$, where $y \in \{0, 1\}$ is the sound event to be classified as cry or non-cry and $\boldsymbol{x}$ is the feature vector. Using a threshold value $\mathtt{Th}_1$, an initial decision value for each segment is set according to the following rule:

$$d(n) = \begin{cases} 1, & \text{if } h_{\boldsymbol{\theta}}(\boldsymbol{x}) > \mathtt{Th}_1 \\ 0, & \text{otherwise.} \end{cases} \tag{3}$$

The duration of a single segment is about $93ms$, while most cry events are at least several hundred of milliseconds long. In order to avoid erroneous detections of sections that are too short to be a likely cry event, a smoothing operation is applied to the sequence of initial decisions as follows: a sliding window of length L is applied on the initial sequence of decisions and the smoothed decision $d_s(n)$ for the central segment is updated according to the following rule:

$$d_s(n) = \begin{cases} 1, & \text{if } \sum_{k=-M}^{M} d(n-k) > \mathtt{Th}_2 \\ 0, & \text{otherwise.} \end{cases} \tag{4}$$

where L is odd, $M = (L-1)/2$ and $\mathtt{Th}_2 \in [1, L]$ is a predefined threshold value.

IV. Convolutional Neural Networks

Convolutional neural networks (CNN) [11] have wide applications in the fields of computer vision, natural language processing and many others, especially where huge amounts of data have to be processed and classified. Like ordinary neural networks, they consist of several layers connected by neurons that have learnable weights. Each CNN layer is composed of several filters, applied to outputs provided by the previous layer using the convolution operation. CNNs learn the filters during the training process, which can be thought of a way to generate important features out of the data. Thus, in contrast to traditional classification algorithms, the lack of dependence on prior knowledge is a major advantage of CNNs.

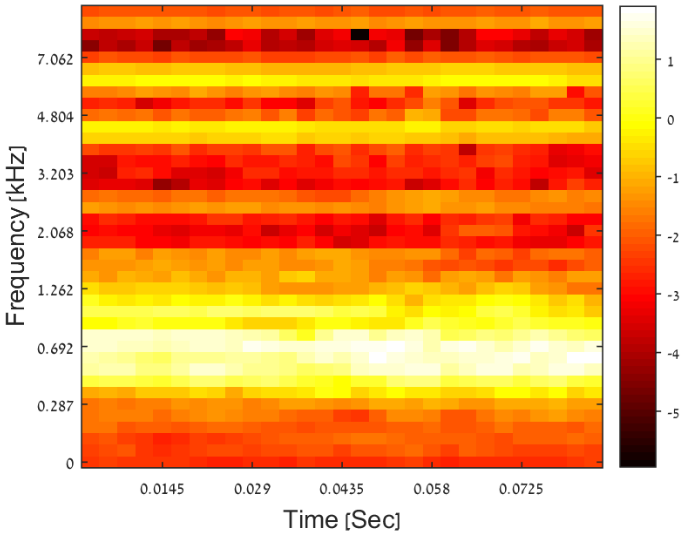

Fig. 4: An LMFB representation of a cry frame. Note that the non-uniform gaps in the frequency axis are due to the logarithmic Mel scale

To work with a CNN classifier, the audio signal is divided into consecutive segments of 4096 samples. Each segment is further divided into frames of 512 samples, with a step size of 128 samples. As the contribution of high frequency bands to the detection of cry signals is limited, a low-pass filter at 11025 Hz is applied. A log Mel-filter bank (LMFB) representation is then produced for each frame, using 40 filters distributed according to the Mel scale in the frequency range $[0, 11025]$ Hz. Given segments of 4096 samples and a step size of 128 samples, this leads to a 40×29 "image" representation of each segment. An example is shown in Figure 4.

The main difference between LMFB and MFCC is that the discrete cosine transform (DCT) of the log-power spectrum is skipped in LMFB representation. This is mainly due to the tendency of the DCT to decorrelate the data, whereas spatial correlation of the input is actually advantageous for a CNN.

The distinctive features of a signal within a frame are mostly due to frequency changes. Thus, our CNN uses convolution layers with "tall" filters: 10×2, 6×2 and 3×2, to achieve high frequency resolution compared to low temporal resolution. Due to the Mel scale, each "pixel" in the LMFB represents a frequency range. To better capture the frequency behavior, small stride values are used. Similarly, the max-pooling layers consist of small blocks, to emphasize the content of correlated frequency bands. The activation function for the CNN is the standard rectifier, corresponding to rectified linear unit (ReLU) layers. The entire CNN architecture is shown in Figure 5.

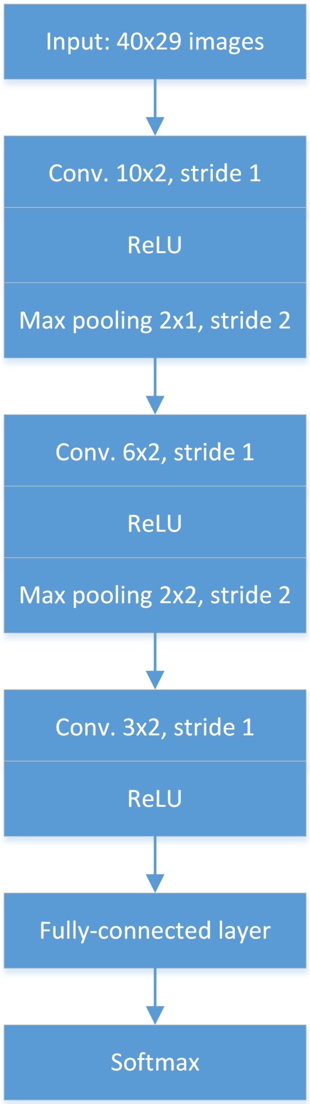

Fig. 5: The CNN architecture.

To train the network, we employed a stochastic gradient descent algorithm with momentum [12]. The gradient in each iteration was evaluated using a mini-batch of 256 frames, over 50 training iterations. A visualization of the filters obtained for the first convolution layer after the training phase is provided in Figure 6. It is evident that at this initial layer the filters capture mostly basic image features such as edges, which correspond to fast transitions in LFMB values.

V. Performance Evaluation

Two important measures for the evaluation are the detection rate and the false-positive rate. The detection rate (also known as sensitivity or recall) is defined as the ratio between the number of true-positive events, i.e. the number of cry events correctly identified, and the total number of cry events in the recording set (true positives and false negatives). The false-positive (or false-alarm) rate is defined as the ratio between the number of false positives (non-cry events identified erroneously as cry events) and the total number of non-cry events in the recording set (including true negatives). Thus, if TP, TN, FP and FN denote true positives, true negatives, false positives and false negatives, respectively, then the detection rate is $\mathrm{TP}/(\mathrm{TP}+\mathrm{FN})$, and the false-positive rate is $\mathrm{FP}/(\mathrm{FP}+\mathrm{TN})$.

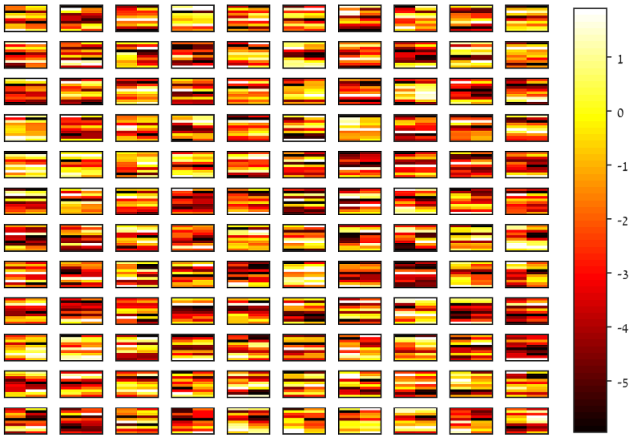

Fig. 6: First convolutional layer filter weights (120 filters, each of dimensions 10×2).

One of the goals of the current study is to construct a platform for conducting psychological research on co-regulatory patterns between a baby and its caregiver, with cry events being a primary variable as a predictor of attachment. Thus, the importance of obtaining a high detection rate is obvious. However, a low false-positive rate is perhaps even more important, in order to avoid the contamination of data with non-related events, which may prevent meaningful conclusions.

Therefore, in the analysis of the cry-detection performance of the logistic regression and the CNN classifiers we focus on the trade-off between the false-positive rate and the detection rate. The performance evaluation was carried out using a receiver operator characteristic (ROC) curve, as shown in Figure 7. Both classifiers were trained using a similar training set of 18000 frames (about 30 minutes), and the ROC curves were obtained using a validation set of two hours. For false-positive rates lower than 5%, the CNN classifier evidently outperforms the logistic regression classifier.

The evaluation results for both classifiers are summarized in Table I. For detection rates of 80%, 85% and 90%, the false-positive rates of the CNN classifier are lower than the corresponding rates of the logistic regression classifier. However, the performance is similar and reversed for higher detection rates. Keeping the false-positive rate at a fixed value of 1.0%, a detection rate of 82.5% is yielded for the CNN classifier, versus 81.0% and 65.0% for the logistic regression classifier, with and without the smoothing procedure, respectively.

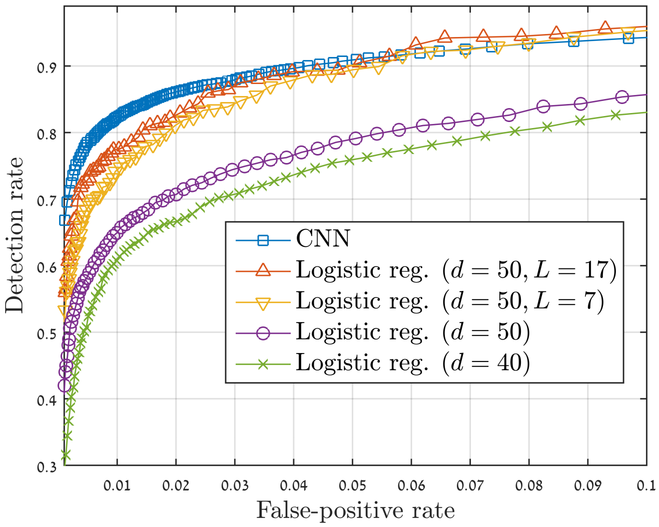

Fig. 7: ROC curves for the logistic regression and the CNN classifiers.

Classifier/Detection rate	80%	85%	90%	95%
Logistic Regression	0.9%	2.1%	4.3%	9.0%
CNN	0.7%	1.6%	4.2%	12.0%

TABLE I: A summary of the false-positive rates for a given detection rate among the two classifiers.

VI. CONCLUSIONS

In this work, two machine-learning algorithms were proposed for the detection of baby cry in audio recordings: a logistic regression classifier and a more complex CNN classifier. The results show a considerable advantage of the CNN classifier compared to the logistic regression classifier. As CNNs are naturally suited for large training datasets and for multi-class classification, we plan to train a CNN classifier to detect various types of domestic sounds in addition to cry signals.

ACKNOWLEDGMENT

The authors would like to thank the staff of the Signal and Image Processing Lab (SIPL), Technion, and the staff of the Vision and Image Sciences Lab (VISL), Technion, for their support.

REFERENCES

[1] D. Stowell, D. Giannoulis, E. Benetos, M. Lagrange, and M. D. Plumbley, “Detection and classification of acoustic scenes and events,” *IEEE Transactions on Multimedia*, vol. 17, no. 10, pp. 1733–1746, Oct 2015.

[2] D. Barchiesi, D. Giannoulis, D. Stowell, and M. D. Plumbley, “Acoustic scene classification: Classifying environments from the sounds they produce,” *IEEE Signal Processing Magazine*, vol. 32, no. 3, pp. 16–34, May 2015.

[3] L. L. LaGasse, A. R. Neal, and B. M. Lester, “Assessment of infant cry: Acoustic cry analysis and parental perception,” *Mental Retardation and Developmental Disabilities Research Reviews*, vol. 11, no. 1, pp. 83–93, 2005.

[4] L. Beckes, H. IJzerman, and M. Tops, “Toward a radically embodied neuroscience of attachment and relationships,” *Frontiers in Human Neuroscience*, vol. 9, p. 266, 2015.

[5] J. Saraswathy, M. Hariharan, S. Yaacob, and W. Khairunizam, “Automatic classification of infant cry: A review,” in *2012 International Conference on Biomedical Engineering (ICoBE)*, Feb 2012, pp. 543–548.

[6] G. Varallyay, “The melody of crying,” *International Journal of Pediatric Otorhinolaryngology*, vol. 71, no. 11, pp. 1699–1708, Nov. 2007.

[7] R. Cohen and Y. Lavner, “Infant cry analysis and detection,” *2012 IEEE 27th Convention of Electrical and Electronics Engineers in Israel (IEEEI 2012)*, pp. 2–6, 2012.

[8] H. IJzerman et al., “A theory of social thermoregulation in human primates,” *Frontiers in Psychology*, vol. 6, no. 464, 2015.

[9] A. M. Noll, “Cepstrum pitch determination,” *The Journal of the Acoustical Society of America*, vol. 41, no. 2, pp. 293–309, 1967.

[10] X. Huang, A. Acero, and H.-W. Hon, *Spoken Language Processing: A Guide to Theory, Algorithm, and System Development*. Prentice Hall PTR, 2001.

[11] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer Science+Business Media, 2006.

[12] K. P. Murphy, *Machine Learning: A Probabilistic Perspective*. MIT Press, 2012.