# Cry-based Detection of Developmental Disorders in Infants

Amit Oren, Avi Matzliach, Rami Cohen

Signal and Image Processing Lab (SIPL)
Andrew and Erna Viterbi Faculty of Electrical Engineering
Technion - Israel Institute of Technology
Technion City, Haifa 3200003, Israel
{amito, avi.m, rc}@campus.technion.ac.il

Hagit Friedman

Faculty of Health and Social Sciences
Department of Nursing
University of Haifa
Haifa 3498838, Israel
hmts@netvision.net.il

***Abstract*—Developmental disorders are a group of neurological conditions originating at childhood, which involve serious impairments in language, learning, and motor skills. An early detection of developmental disorders is crucial, as it enables early intervention (e.g., speech-language and occupational therapy) that may reduce neurological and functional deficits. In this work, we develop a tool for an early identification of developmental disorders in infants based on their cry. We use signal-processing tools to extract distinguishing cry features (e.g., pitch and formants), and exploit their correlation with the risk of having developmental disorders. To estimate this risk, we train a $k$-NN machine-learning system. Performance is evaluated against a database of diagnosed infants, with 89% accuracy in cross-validation testing.**

## I. Introduction

Neuro-developmental disorders are impairments in the growth and development of the brain or the central neurologic system. These impairments are characterized in damage to personal, social and occupational skills. As of 2008, approximately 15% of children in the United States have been diagnosed with a developmental disorder, compared to only 12.8% in 1997 [1]. It is widely accepted that early intervention can profoundly improve the quality of life of children at risk and their families [2]. Many developmental impairments, such as language and social delays associated with autism, can be identified as early as 18 months of age. However, over 39% of children in the United States with Autism Spectrum Disorder (ASD) are not diagnosed before the age of 6 [3]. This may result in a delayed intervention, making the prognosis less likely to improve. To enable an early and effective intervention, reliable tools for the detection of early signs of developmental disorders are of utmost importance.

As the production of crying requires the coordination of different areas of the brain, irregularities in the cry signal may indicate a neurological insult. For example, the brainstem controls the laryngeal muscles, the tongue and the lungs through the vagal complex. These three organs shape the form of the cry signal, hence their malfunction will produce an irregular cry signal [4]. Previous researches have shown that the infant cry signal can be used as a diagnosis tool for health and developmental status. In [5], major differences were seen between the cries of healthy infants and infants with asphyxia. In [6], it was shown that infants at risk for ASD produced a significantly different cry from healthy infants. Several approaches have been made at developing an automated system for characterizing pathologies in infants' cries. In [7], hidden Markov models are used for the automatic classification of various types of infants' cries. A feed-forward neural network was used in [8] to classify normal and pathological cries of deaf infants. A similar approach was proposed in [9] using general regression neural networks. However, these methods require complex classifiers and long training time.

In this work, we introduce a cry-based tool for a convenient and early diagnosis of infants. This tool is aimed at being simple and cost-effective, and it simply requires samples of an infant cry. Ultimately, it may serve as a first screening layer, indicating whether parents should consult an expert. We first provide a mechanism for the extraction of vocal features from infant cry signal. We then use a $k$-NN based machine learning system to estimate an infant developmental status. We demonstrate that certain features are remarkably more indicative of developmental disorders than others. By an appropriate weighting of the features, good accuracy in detecting developmental disorders in infants is achieved.

The paper is structured as follows. In Section II, we discuss the methods used to extract features from the cry signals. Section III describes the machine-learning system for infant cry classification and performance evaluation. Finally, conclusions are drawn in section IV.

## II. Methods

The cry signal is composed of crying utterances of varying length, where in between the infant pauses to inhale. As opposed to speech signals, where through an analysis of voiced and unvoiced parts words and syllables can be recognized, a healthy cry signal usually contains a majority of voiced frames. For improved analysis, we divide the cry acoustic features into *frame* and *segment* features. Frame features are extracted from 15[ms] frames, whereas segment features are extracted from

0.3[s] segments with 50% (0.15[s]) overlap. The time lengths of the frames and segments are justified later in this section.

### A. Databases

Two cry signal databases were used in this paper:

1) Infant cry database, collected by H. Friedman of the Department of Nursing in the University of Haifa. This database consists of cry signals of 25 infants aged 34 – 70 gestational weeks. Each infant is classified as either healthy (no developmental disorder) or impaired (suffers from a developmental disorder).
2) The Chillanto database [10]. This database consists of normal cries, cries of deaf infants, asphyxiated infants, infants suffering from hunger, or infants in pain.

### B. Frame features

The following features are extracted from each frame. The number of samples in a frame is denoted by $N$.

*1) Pitch frequency:* The cry sound is elicited due to periodic vibration of the vocal cords. The frequency of these vibrations is known as the *pitch frequency* $f_0$ of the cry. We use the *windowed autocorrelation* method [11] for pitch detection. For improved pitch detection reliability, a time frame with at least 3 pitch periods is needed. As the typical pitch range in infants is 200[Hz]–450[Hz], we work with frames of $3/200$[s] = 15[ms], each containing at least 3 pitch periods.

*2) Formants:* The physical barriers of the oral and nose cavity constitute a sound box, shaping the spectral shape of the cry signal. The flow of air through them adds dominant frequencies, called *Formants*, corresponding to the resonant frequencies of the oral and nose cavities. We extract the first 3 formants ($F1$, $F2$ and $F3$) based on the line-spectral pair representation [12]. Typical formant values for healthy infants are approximately 1100[Hz] for $F1$, 3300[Hz] for $F2$ [4] and 3500[Hz] for $F3$.

*3) Spectral centroid:* We perform a fast Fourier transform (FFT) on each frame to calculate its *Spectral Centroid* (SC). Assuming that the frame FFT vector is $\boldsymbol{f}$, and the corresponding frequencies vector is $\boldsymbol{x}$, SC is calculated as:

$$\mathrm{SC[Hz]} = \frac{1}{\mathrm{N}} \frac{\sum_{\mathrm{i}=0}^{\mathrm{N}-1} |\mathrm{f_i}| \cdot \mathrm{x_i}}{\sum_{\mathrm{j}=0}^{\mathrm{N}-1} |\mathrm{f_j}|}. \tag{1}$$

SC gives an estimate of the spectral content of the frame. A typical SC value for healthy infants is approximately 1000[Hz].

*4) Short time energy:* Again using the frame FFT vector $\boldsymbol{f}$, we calculate the short time energy of the frame, $E$ defined as:

$$E = \frac{1}{N} \sum_{i=0}^{N-1} |f_i|^2. \tag{2}$$

*5) Quarterly frequencies:* Using the short time energy of the frame, we find the frequencies above which 25%, 50% and 75% of the energy resides. Those are dubbed as the first, second and third *quarterly frequencies* of the frame energy spectrum.

*6) Mel-Frequency Cepstrum Coefficients (MFCC):* To better estimate the spectral envelope of the signal, we extract the MFC coefficients. In the Mel representation, the frequency axis is scaled to match the Mel logarithmic scale, which simulates better the way pitch is perceived in human ears. The Periodogram of each frame is multiplied by a series of triangular band-pass filters, where each filter matches a different frequency on the Mel-frequency scale. The logarithm of the total spectral energy of each filter is computed, and discrete cosine transform (DCT) is performed to obtain the MFCC [12].

*7) Linear Predictive Coding (LPC) Coefficients:* LPC coefficients represent the spectral envelope of the signal using a linear predictive model. They are the coefficients of a forward linear predictor, obtained by minimizing the prediction error of the original signal in the least squares sense. LPC coefficients are mostly used for speech compression and encoding, as the spectral envelope can be efficiently represented by a small number of coefficients. As this is a robust method for speech processing, we found it also plausible for cry signal phenomena analysis. In this work, the first 3 LPC coefficients are extracted from each frame.

### C. Segmentation and segment length

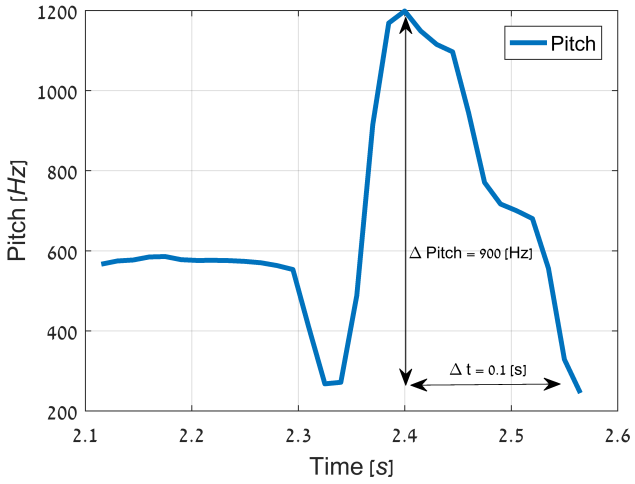
Due to phenomena occurring in time intervals longer than frame length, we use features extracted from segments as well. For example, the *pitch contour* tracks the pitch (detected in the frame level) over time, where variations in the pitch contour through a cry utterance may indicate possible disorders [1]. Thus, it is important to extract features from segments in addition to frames.

There are several consideration in choosing the segment length. First, the segment length should be reasonably long compared to the frame length, to capture time-varying phenomena correctly. Second, to avoid discarding utterance parts due to segmentation, the segment length should be a divisor of the typical utterance length. Finally, it is desired that the segment length is a multiple of the frame length, such that no frame parts are discarded. By empirically testing several segment length values, a segment length value of 0.3[s] resulted in good detection of time-varying phenomena with a minimal loss of utterance parts due to segmentation. In particular, for this choice of segment length, each segment contains $0.3[\mathrm{s}]/15[\mathrm{ms}] = 20$ frames, such that no frame parts are discarded due to segmentation.

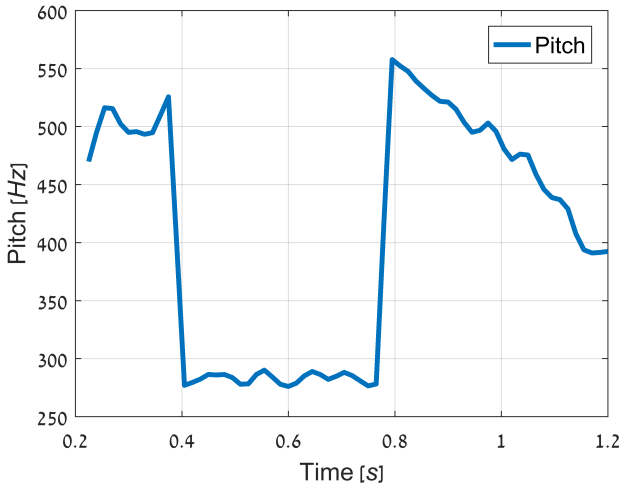
### D. Segment features

Previous studies have shown a correlation between certain pitch patterns (e.g., rapid variations) and possible developmental disorders in infants [1]. In the rest of this sub-section, we describe several segment features aimed at identifying such patterns.

*1) Glide:* The glide feature is defined as a steep rise or fall of $f_0$ of at least $600[\mathrm{Hz}]/0.1[\mathrm{s}]$ [1]. This justifies choosing segment length of 0.1[s], as a shorter length would potentially




(a) An example of the glide phenomenon.




(b) An example of the vibrato phenomenon.

Fig. 1. Pitch-related features.

reduce glide detection. The extraction of glide from a cry segment is done by splitting the utterance into sub-segments where the pitch contour is constantly rising or constantly falling. The average slope of the pitch contour, denoted $\Delta p$, is then calculated from each sub-segment by:

$$\Delta p = \frac{|\Delta f_0|}{\Delta t}. \tag{3}$$

Where $\Delta f_0$ is the total change in $f_0$ in the sub-segment, and $\Delta t$ is the duration of the sub-segment. If $\Delta p \geq {}^{600[\text{Hz}]}/_{0.1[\text{s}]}$, the sub-segment is marked as containing a glide. Otherwise, the sub-segment is not marked as containing a glide. Fig. 1(a) shows a steep falling glide in the pitch contour, with a fall of approximately $900$[Hz] in $0.1$[s].

*2) Vibrato:* Vibrato is defined as *rapid* falling and rising of $f_0$. To detect vibrato in a segment, the sizes of sub-segment groups containing runs of more than 2 positive/negative pitch differences (larger than 3Hz) are summed. This sum in then normalized by a maximal empirical value, and dubbed as the *vibrato intensity* of the segment. An example of the vibrato feature is shown in Fig. 1(b).

*3) Cry melody:* Cry melody describes the general trend of the pitch contour; whether it rises, falls or flat. Utterance containing a majority of one melody might indicate a developmental disorder [1]. The identification of cry melodies is done by using derivatives of the pitch contour. A positive valued derivative indicates a rising trend in the pitch contour, where a negative valued derivative a falling trend. We begin by smoothing the segment pitch vector using a moving average. Local extremum points are then extracted from the smoothed vector. The difference between the pitch of each two adjacent extremum points is calculated. A positive difference of over $50$[Hz] corresponds to a rising melody, whereas a negative difference of over $50$[Hz] corresponds to a falling melody. Between these thresholds, the melody is considered as flat.

*4) Cry mode:* Cry mode describes a continuous temporal state in a cry segment, where the pitch contour is in either in a certain range, or cannot be clearly detected (e.g., the signal is aperiodic). Two modes are extracted in this work: *phonation*, where the pitch is up to $750$[Hz], and *hyperphonation*, where the pitch is above $1000$[Hz] [1].

## III. Cry classification system

### A. Cry classification

The cry signals in the database are diagnosed as belonging to either 'healthy' or 'impaired' infants. Our aim is to train a machine-learning system, for the detection of developmental disorders in infants based on features extracted from their cry signals. We use the $k$-NN algorithm with $k = 5$ that provided the best classification results. In the training phase, each frame is represented as a 22-dimensional feature vector, based on the features described in Section II-B and Section II-D. The entire system architecture is shown in Fig. 2.

For improved classification results, the RELIEFF iterative feature selection algorithm [13] is used. This algorithm provides a weight for each feature, measuring its contribution to the correct classification of training samples. As shown in Fig. 3, the most prominent features are the short time energy of the frame, the third formant, the vibrato feature and the segment cry melodies (falling, rising, flat).

### B. Performance evaluation

To evaluate the system performance, we tested the system against both databases using $n$-fold balanced cross-validation. To reduce correlation in the training set, the cry signals are divided into disjoint sets, such that a cry signal does not appear in both sets. In our tests, the system classified correctly $89\%$ of the infants. The percentage of cry signals falsely detected as 'impaired' while tagged as 'healthy' is approximately $9\%$, whereas the percentage of cry signals falsely detected as 'healthy' while tagged as 'impaired' is negligible. For comparison, when non-overlapping segments are used, the

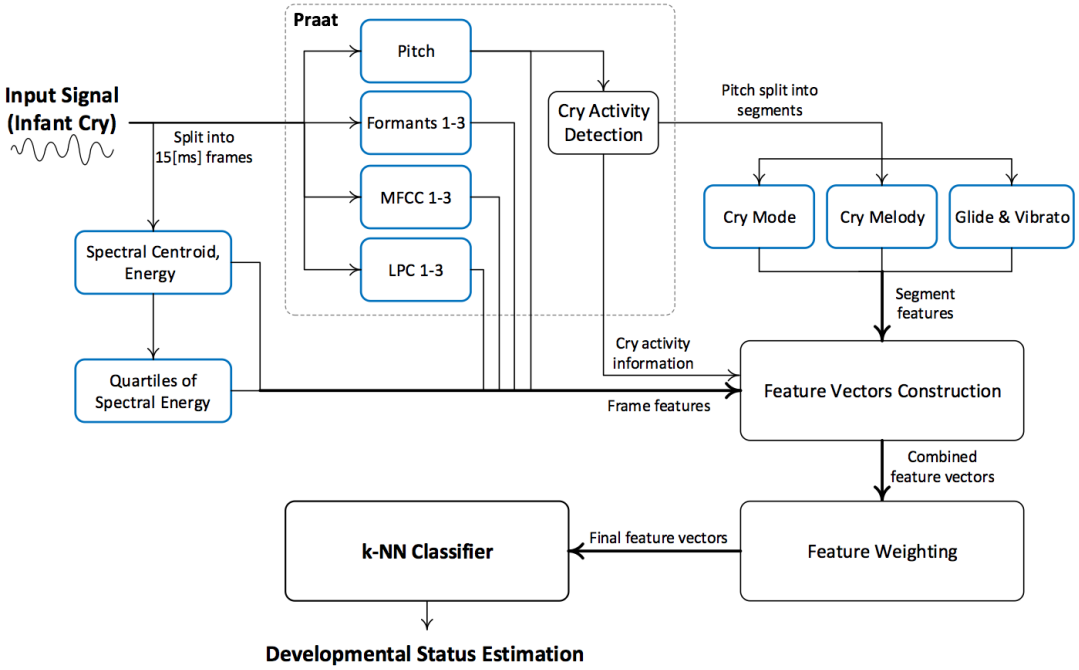



Fig. 2. Infant cry classification - block scheme.

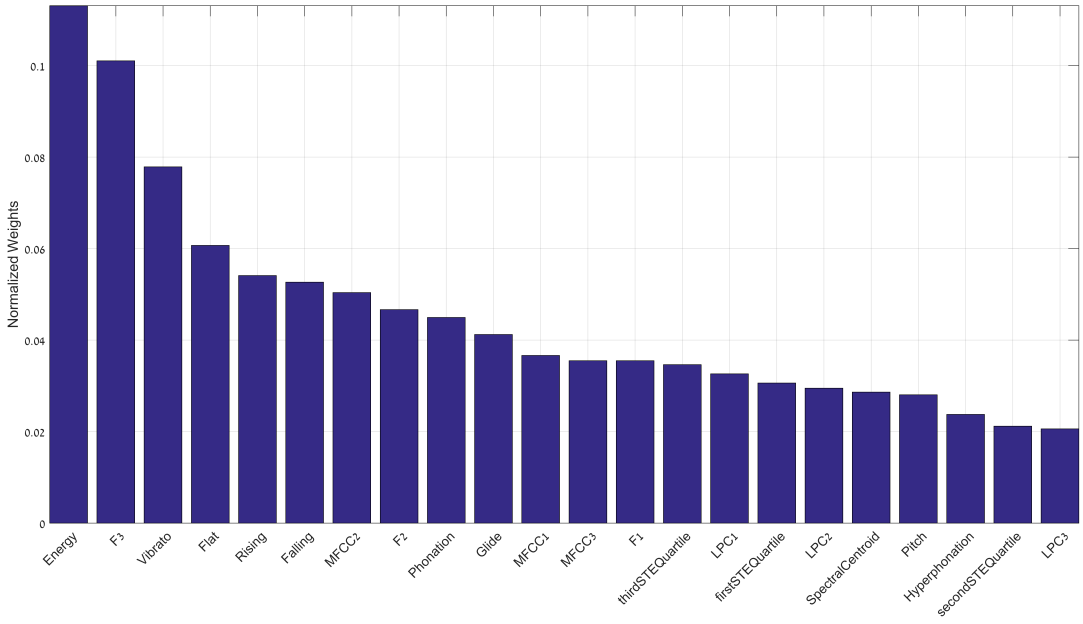



Fig. 3. Mixed frame and segment features after RELIEFF.

accuracy is 82.7%. This indicates the importance of the use of overlapping segments, yielding additional segment features and improving classification performance.

## IV. Conclusions

In this work, we presented a system for the detection of health disorders in infants, based on their crying. The system is based on extracting temporal and spectral features out of a cry signal, followed by a $k$-NN classifier. Performance evaluation shows approximately 90% accuracy on a database

of diagnosed infants. The results demonstrate the potential of cry analysis for an early detection of developmental disorders. As a possible extension, it is suggested to consider the use of additional machine-learning algorithms such as support vector machines (SVM).

## V. Acknowledgments

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

## References

[1] J. Soltis, "The signal functions of early infant crying." *The Behavioral and brain sciences*, vol. 27, no. 4, pp. 443–458; discussion 459–490, 2004.

[2] L. K. Koegel, R. L. Koegel, K. Ashbaugh, and J. Bradshaw, "The importance of early identification and intervention for children with or at risk for autism spectrum disorders." *International journal of speech-language pathology*, vol. 16, no. 1, pp. 50–6, 2014.

[3] "NCHS data brief no. 97," Center for Disease Control and Prevention (The CDC), Tech. Rep., 2012.

[4] 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.

[5] K. Michelsson, P. Sirvio, and O. Wasz-Hockert, "Pain cry in full-term asphyxiated newborn infants correlated with late findings," *Acta Paediatrica Scandinavica*, vol. 66, no. 5, pp. 611–616, 1977.

[6] S. J. Sheinkopf, J. M. Iverson, M. L. Rinaldi, and B. M. Lester, "Atypical Cry Acoustics in 6-Month-Old Infants at Risk for Autism Spectrum Disorder," *Autism Research*, vol. 5, no. 5, pp. 331–339, 2012.

[7] D. Lederman et al., "On the use of hidden Markov models in infants' cry classification," *The 22nd Convention of Electrical and Electronics Engineers in Israel*, pp. 350–352, 2002.

[8] J. O. Garcia and C. A. R. Garcia, "Mel-frequency cepstrum coefficients extraction from infant cry for classification of normal and pathological cry with feed-forward neural networks," *Proceedings of the International Joint Conference on Neural Networks*, vol. 4, 2003.

[9] M. Hariharan, R. Sindhu, and S. Yaacob, "Normal and hypoacoustic infant cry signal classification using time-frequency analysis and general regression neural network," *Computer Methods and Programs in Biomedicine*, vol. 108, no. 2, pp. 559–569, 2012.

[10] C. A. R. Garcia, "Baby Chillanto infant cry database," INAOE, Mexico.

[11] P. Boersma, "Accurate Short-Term Analysis of the Fundamental Frequency and the Harmonics-To-Noise Ratio of a Sampled Sound," *Proceedings of the Institute of Phonetic Sciences*, vol. 17, pp. 97–110, 1993.

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

[13] I. Kononenko, E. Šimec, and M. Robnik-Šikonja, "Overcoming the myopia of inductive learning algorithms with RELIEFF," *Applied Intelligence*, vol. 7, no. 1, pp. 39–55, 1997.