Infant Cry Analysis and Detection

Rami Cohen[1]
[1]Signal and Image Processing Lab (SIPL)
Department of Electrical Engineering
Technion - Israel Institute of Technology
Technion City, Haifa 32000, Israel
rc@tx.technion.ac.il

Yizhar Lavner[1,2]
[2]Department of Computer Science
Tel-Hai Academic College, Upper Galilee, Israel
yizhar.lavner@gmail.com

***Abstract*—In this paper we propose an algorithm for automatic detection of an infant cry. A particular application of this algorithm is the identification of a physical danger to babies, such as situations in which parents leave their children in vehicles. The proposed algorithm is based on two main stages. The first stage involves feature extraction, in which pitch related parameters, MFC (mel-frequency cepstrum) coefficients and short-time energy parameters are extracted from the signal. In the second stage, the signal is classified using the k-NN algorithm and is later verified as a cry signal, based on the pitch and harmonics information. In order to evaluate the performance of the algorithm in real world scenarios, we checked the robustness of the algorithm in the presence of several types of noise, and especially noises such as car horns and car engines that are likely to be present in vehicles. In addition, we addressed real time and low complexity demands during the development of the algorithm. In particular, we used a voice activity detector, which disabled the operation of the algorithm when voice activity was not present. A database of baby cry signals was used for performance evaluation. The results showed good performance of the proposed algorithm, even at low SNR.**

I. Introduction

Infant crying can be considered a biological alarm system [1], and it is the first means of communication for newborns. Infant crying signals distress or needs, calls for the attention of parents or caregivers and motivates them to alleviate the distress [2].

Infant crying is characterized by its periodic nature, i.e. alternating cry utterances and inspirations. Each utterance or burst sound is produced by a rapid flow of air through the larynx, eliciting the repeated opening and closing of the vocal folds, which in turn generates periodic excitation. This excitation is transferred through the vocal tract to produce the cry sound, which normally has a fundamental frequency (pitch) of 250-600 Hz [2].

The acoustic signal of an infant's cry is of great importance since it contains valuable information about their physical and physiological condition, such as health, weight, identity, gender and emotions [3].

In the past thirty years, different studies that have analyzed the acoustic characteristics of the infant or baby cry have been carried out, for various purposes and applications. These include, for example, recognition of potential neurological insults or the medical status of newborns [4], discrimination between normal and hearing impaired babies [5], enhancing social robots' behavior in childhood education settings [6], or identification of babies solely by the sound of their cry [7].

In this paper, we describe an analysis of the cry sound of infants and present an algorithm for cry detection, which is aimed at alerting parents in situations of potential physical danger, for example, on occasions of infants being left alone in closed apartments or vehicles.

The proposed algorithm is based on two main stages. The first stage involves feature extraction, in which pitch related parameters, MFC (mel-frequency cepstrum) coefficients and short-time energy parameters are extracted from the signal. In the second stage, the signal is classified using k-NN and later verified as a cry signal.

The rest of the paper is organized as follows. In Section II, the characteristic features of the cry signal and the procedures for their computation are detailed, and the algorithm of cry detection is described. Performance evaluation of the proposed algorithm is presented in Section III. Finally, conclusions are drawn in Section IV.

II. Methods

A. Database

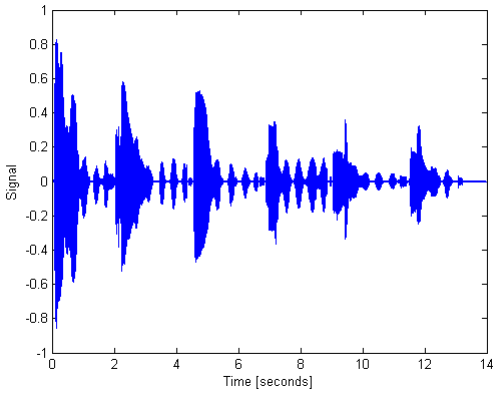
The cry signals used in this paper were obtained from a database created by G. Varallyay [8]. This database consists of cry signals of babies ranging in age between $1-2$ years. An example of a baby cry signal is shown in Figure 1, with its corresponding spectrogram.

In order to evaluate the performance of the proposed algorithm in a noisy environment, we used several types of noise, including engines, passers-by, motorcycles and speech signals, obtained from several databases. We focused on noise sounds that are likely to be present in the vicinity of vehicles.

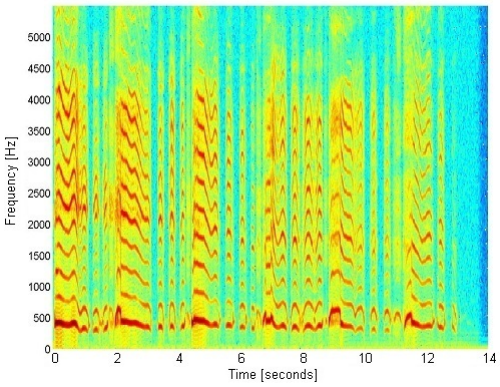
B. Signal Analysis and Feature Extraction

The following features were extracted from the baby cry signals:

978-1-4673-4681-8/12/$31.00 ©2012 IEEE



(a) Signal waveform



(b) Spectrogram

Figure 1: Baby cry example

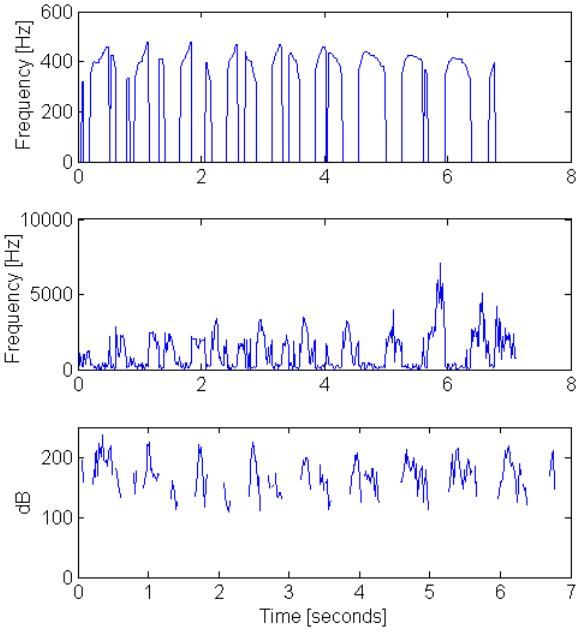


Figure 2: Example for features of a cry signal. Top: pitch, middle: HF, bottom: HAPR

1) Pitch frequency: The fundamental frequency f_0 is important for classification purposes. The pitch detection algorithm is based on a combination of the cepstral method [9] for coarse-pitch period detection, and the cross-correlation method [10] for refining the results.

2) Short-time energy: The short-time energy (STE) of a signal $x[n]$, using an analysis frame of N-samples length (beginning at $n = N_0$), is defined as:

$$E[N_0] = \frac{1}{N} \sum_{N_0}^{N_0+(N-1)} x^2[n] \tag{1}$$

3) Mel-Frequency Cepstrum Coefficients (MFCC): MFCC [11] provide a representation of the short-term power spectrum of a signal. These coefficients are obtained by multiplying the short-time Fourier Transform (STFT) of each analysis frame by a series of M triangularly-shaped ideal band-pass filters, with their central frequencies and widths arranged according to a mel-frequency scale. The total spectral energy $E[i]$ contained in each filter is computed and a Discrete Cosine Transform (DCT) is performed to obtain the MFCC sequence:

$$MFCC(l) = \frac{1}{M} \sum_{i=0}^{M-1} \log(E(i)) \cdot \cos\left(\frac{2\pi}{M}\left(i + \frac{1}{2}\right) \cdot l\right) \tag{2}$$

for $l = 0, ..., M-1$.

4) Harmonicity Factor: The Harmonicity Factor (HF) provides an estimate of the presence of harmonics in each analysis frame. A pre-defined number of peaks is found in the magnitude of the DFT for each analysis frame. Then, the result of the mod operation of each of the highest k peak frequencies (h_i) with the pitch frequency f_0 is obtained. The sum of mods yields the harmonicity factor:

$$HF = \sum_{i=1}^{k} h_i \bmod f_0 \tag{3}$$

As this parameter tends to 0, the signal is considered as more harmonic. The cry signal has a harmonic nature, as can be seen in Figure 1b, so small values of this parameter are expected for bursts of cry signals.

5) Harmonic-to-Average Power Ratio: In almost all cry bursts, the signal contains strong harmonic components in at least part of the burst duration. The harmonic-to-average power ratio (HAPR) is a spectral feature [12] that determines the ratio of the harmonic component power and the average spectral power. The first step in obtaining the HAPR is identifying the highest peaks around the harmonic frequencies in the DFT magnitude of each frame, where the harmonics are multiples

of the pitch frequency. Denoting the power component around the m^{th} harmonic as $|X(2\pi m f_0, t)|^2$, the HAPR is described as:

$$HAPR(t, M)[dB] = \frac{1}{M} \sum_{m=2}^{M} 10log_{10} \frac{|X(2\pi m f_0|^2}{|P_x(t)|^2} \quad (4)$$

where f_0 is the normalized pitch frequency of the t^{th} frame, and the average spectral power for this frame is defined as:

$$P_x(t) = \frac{1}{N} \sum_{k=0}^{N-1} |X(\omega_k), t|^2 \quad (5)$$

with N the length of the DFT, and $\omega_k \triangleq 2\pi k/N$.

6) Burst Frequency: This parameter is computed from the DFT of the short-time energy signal, and measures the frequency of bursts that appear in the STE graph. The logic behind this feature is that an infant cry tends to be periodic. Therefore, the maximal peak of the DFT of the STE should indicate this periodicity.

7) Rise-time and Fall-time of the short-time energy: Rise-time and fall-time are determined by averaging the rise-time and fall-time for each of the highest 3 peaks of the STE.

Some of the features (1, 4 and 5) are shown in Figure 2.

C. Cry Detection Algorithm

The aim of the detection algorithm is to classify each incoming segment of a stream of input audio signals as 'cry', 'no cry', or as 'no activity'. The algorithm analyzes the signal at various time-scales (*segments* of several seconds, *sections* of about 1 second, and *frames* of several tens of msec), in order to achieve a low false-negative error rate.

In order to work only on meaningful parts of the signals, we employed a statistical model-based voice activity detector (VAD) [13]. This VAD was used in order to determine when our baby cry detection algorithm should analyze the input signal. That leads to reduction in power consumption in cases such as a microphone in a car that mostly receives no signal or only noise.

The proposed algorithm is composed of three main stages:

i) **Voice Activity Detector (VAD)** for detecting sections with sufficient audio activity.
ii) **Classification** Using k-nearest neighbours (k-NN) algorithm, in which each frame is classified either as a cry sound ('1'), if close enough to cry training samples, or as 'no cry' ('0').
iii) **Post-processing** for validating the classification stage in order to reduce false-negative errors.

The detailed detection algorithm works as follows (Figure 3):

1) The signal is divided into consecutive and overlapping segments, each of 10 seconds, with a step of 1 second.
2) For each segment, a VAD is applied and the amount of activity is calculated. Each segment is further divided into sections of 1 second, with an overlap of 50% . If the activity duration of a given section is below a pre-defined threshold (30% in our evaluation experiments), the section is considered as having insufficient activity, and is classified as 'no cry' or '0'. If the activity is above the threshold, the section is divided into short-time frames (with a duration of 32 msec and a hop size of 16 msec in our evaluation).
3) The following parameters are computed for each frame:
   - The MFC coefficients
   - The pitch frequency
   - Harmonicity factor (HF)
   - Harmonic-to-average power ratio (HAPR)
4) Each frame is classified either as 'cry' or as 'no cry', based on its MFCC using a k-NN classifier. For each section, if at least half of the frames are classified as 'cry', the whole section is considered as 'cry'. Otherwise, it is considered as 'no cry'.
5) In cases in which the section is classified as 'no cry', a post-processing stage is conducted, where the following conditions are checked for the frames included in the section:
   - The pitch frequency is between 300 Hz and 600 Hz for at least K_f frames;
   - The HF is below 100 Hz for at least M_f frames;
   - The HAPR of at least N_f frames is above 100 dB and below 300 dB.

   where K_f, M_f and N_f are pre-defined thresholds. If at least two of the above conditions are satisfied, the section is considered as 'cry'. Otherwise, it is considered as 'no cry'.
6) For each segment, if at least M_f sections are classified as 'cry', the whole segment is classified as 'cry'.

III. Performance Evaluation

We evaluated the performance of the proposed algorithm, also testing its robustness in the presence of noise. A signal consisting of randomly distributed baby cry samples, as well as noise samples, was used. The signal's duration was 370 seconds, of which 50% was baby cry samples. The signal was contaminated with additive white Gaussian noise (AWGN) of varying variances, in order to obtain desired levels of SNR.

We measured both the *section detection rate*, which is defined as the fraction of the correctly classified sections (1 second each) of baby cry, and the *segment detection rate*, which measures the fraction of correctly classified segments (10 seconds each) in which baby cry is present.

The results are shown in Figure 4. As can be seen, our algorithm provides good performance even at low SNR, with a detection rate of almost 100% in high SNR, for both sections and segments.

In addition, it can be seen that the use of segments improves the robustness of the proposed algorithm by 3-8%. This improvement is of great importance, since it reduces

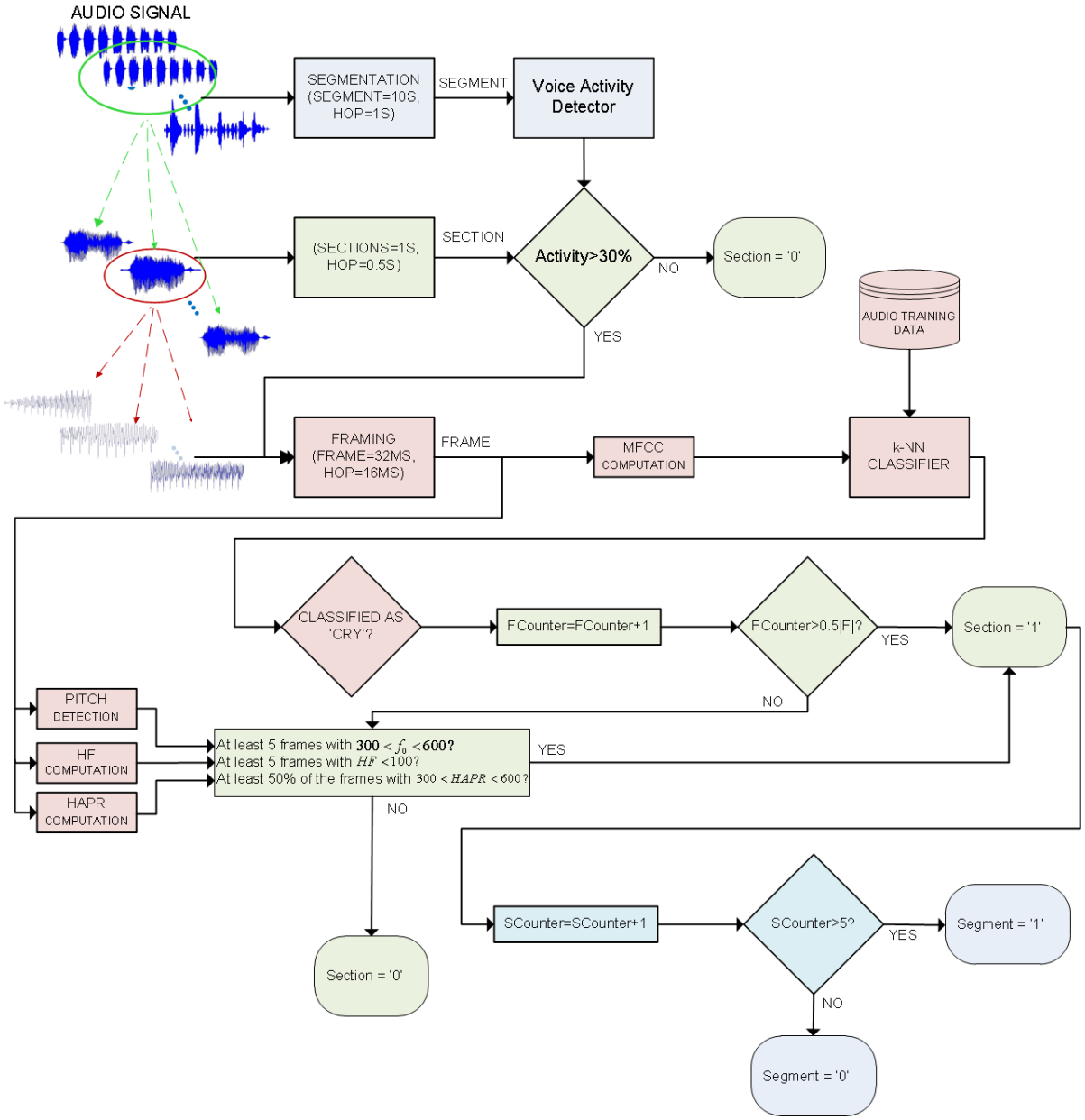


Figure 3: Baby cry detection algorithm - block scheme

the percentage of false-negatives. The false-positive rate was negligible.

IV. Conclusions

We present an efficient algorithm for detecting infant cry signals. The algorithm is based on three decision levels in different time-scales: a frame level, in which each frame (tens of msec) is classified either as 'cry' or 'no cry', based on its spectral characteristics; sections of a few hundred msec; and segments of several seconds for which the final decision is obtained according to the number of 'cry' sections they contain.

The multiple time-scale analysis and decision levels are aimed at providing a classifier with very high detection rate, while keeping a low rate of false positives. The results of the performance evaluation, with infant cry recordings as well as other natural sounds such as car engines, horn sounds and speech, demonstrate both high detection rate and robustness in the presence of noise, even at low SNR values. Another advantage of the algorithm is its simplicity: it is based on a small number of features, which are relatively simple to implement.

In a future research we plan to extend the evaluation of the proposed algorithm, using a broader set of cry and noise

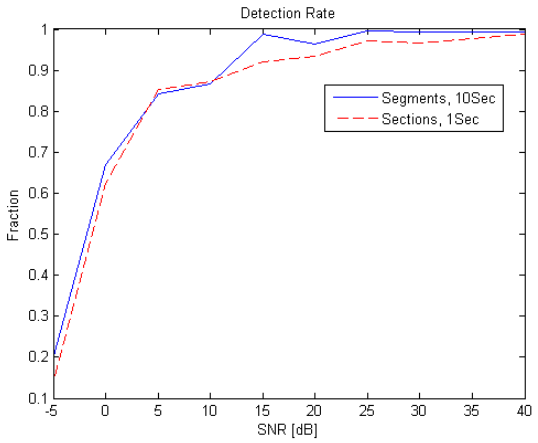


Figure 4: Baby cry detection algorithm - results

signals.

ACKNOWLEDGMENT

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

REFERENCES

[1] L. T. Singer and P. S. Zeskind, editors. *Biobehavioral Assessment of the Infant*, pages 149–166. Guilford Press, 2001.

[2] L. L. LaGasse, R. Neal, and B. M. Lester. Assessment of infant cry: acoustic cry analysis and parental perception. *Mental Retardation and Developmental Disabilities Research Reviews*, 11(1):83–93, 2005.

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

[4] J.O. Garcia and C.A. Reyes Garcia. Mel-frequency cepstrum coefficients extraction from infant cry for classification of normal and pathological cry with feed-forward neural networks. In *Proceedings of the International Joint Conference on Neural Networks*, volume 4, pages 3140–3145, July 2003.

[5] G. Varallyay. The melody of crying. *International Journal of Pediatric Otorhinolaryngology*, 71(11):1699 – 1708, 2007.

[6] P. Ruvolo and J. Movellan. Automatic cry detection in early childhood education settings. In *7th IEEE International Conference on Development and Learning (ICDL)*, pages 204–208, Aug. 2008.

[7] A. Messaoud and C. Tadj. A cry-based babies identification system. In *Proceedings of the 4th international conference on Image and signal processing*, ICISP'10, pages 192–199, 2010.

[8] G. Varallyay. Crysamples, http://sirkan.iit.bme.hu/ varallyay/crysamples.htm, 2009.

[9] A. M. Noll. Cepstrum pitch determination. *The Journal of the Acoustical Society of America*, 41(2):293–309, 1967.

[10] D. Gerhard. Pitch extraction and fundamental frequency: History and current techniques. Technical report, University of Regina, Canada, 2003.

[11] A. Klautau. The MFCC. Technical report, Signal Processing Lab, UFPA, Brasil, 2005.

[12] T. van Waterschoot and M. Moonen. Fifty years of acoustic feedback control: State of the art and future challenges. *Proceedings of the IEEE*, 99(2):288 –327, Feb. 2011.

[13] J. Sohn, N. S. Kim, and W. Sung. A statistical model-based voice activity detection. *Signal Processing Letters, IEEE*, 6(1):1 –3, Jan. 1999.