AUTOMATIC INFANT CRY DETECTION

G. Várallyay Jr., András Illényi, Zoltán Benyó
1Budapest University of Technology and Economics, Dept. of Control Engineering and Information Technology, Budapest, Hungary
2Budapest University of Technology and Economics, Dept. of Telecommunications and Media Informatics, Budapest, Hungary

Abstract: Cry detection can be defined as a procedure where the voiced crying sounds are selected from the recording. The most difficult part of the cry detection is to recognize the inspiratory sounds and separate them from the voiced sounds. In addition, sound recordings may come from different places and recorded with several devices, in this way the method of the cry detection has to be universal. Authors created the Extended Harmonic Product Spectrum method to classify the spectral structure of a given signal. Based on this new method authors developed the Automatic Infant Cry Detection (AICD) system to detect voiced cry sounds in any kind of recording.

Keywords: Infant cry, cry detection, Extended Harmonic Product Spectrum

I. INTRODUCTION

Not only crying sounds can be found in a recording of an infant cry. For example, the infant takes an inspiration between two voiced crying and there can be shorter-longer pauses as well. During these pauses the background noises might be heard in the recording. The recording device might have its own noise. The inspiration can be quiet or audible. It can be placed before or after the voiced crying sound. The infant can suspend the voiced sound or reduce it. The sound of crying can be high-pitched or low-pitched, nasal, veiled, reedy, woody, etc. Many further attributes could be listed in connection with the crying sound.

In a 60 s long recording 8-10 pieces of voiced cries can be found on an average. Cry detection can be defined as a procedure where the voiced crying sounds are selected from the recording. As there are many different kinds of cries, and there might be misleading sounds in the recording as well (background noises, inspiratory sounds, etc.) the cry detection was performed manually in most of the research teams as [1], [2]. For example in 1982 Hirschberg and Szende, or in 1999 Michelsson and Michelsson applied spectrographic analysis of the infant cry, and they selected the voiced crying sounds manually after determining a visual spectrogram from the recording [3], [4]. In the last decade some teams have started applying speech detection software, but generally these software can be used with limitations as the speech and cry signals have differences [5], [6].

The most difficult part of the cry detection is to recognize the inspiratory sounds and separate them from the voiced sounds. In addition, sound recordings may come from different places and recorded with several devices, in this way the method of the cry detection has to be universal. It will be shown that effective cry detection can be executed with limitations and considerations both in the time and the frequency domains.

II. METHODS

From a simplified view the goal of the speech detection is to detect the boundaries of each word, accordingly in the cry detection the start and the end points are to be found of each crying segments. A common attribute of the words that they have a relative big energy, in this way they can be detected by applying a well-chosen energy threshold [7], [8]. In case of cry recordings by seeking for the high energy parts not only the crying segments but also the inspiratory sounds, louder background noises are found.

To create an effective Automatic Infant Cry Detection (AICD) system authors recommend inspecting the spectral content along with the energy content of the recordings. While the crying segments are typically harmonic signals (i.e. having the fundamental frequency and its subharmonics in the spectrum), generally the noise signals (e.g. the inspiratory sounds) have less regular spectral structure [9]. In this study the well-known Short-Time Energy Function was applied to obtain the energy content of the recordings. The spectral content was determined with the extension of the Harmonic Product Spectrum (HPS) method.

A. Short-Time Energy Function

The Short-Time Energy (STE) function of an audio signal is defined as:

$$E_n = \frac{1}{N} \sum_{m} [x(m) \cdot w(n-m)]^2$$

where $x(m)$ is the discrete time audio signal, $n$ is time index of the short-time energy, and $w(m)$ is a rectangle window, i.e.

$$w(n) = \begin{cases} 1, & 0 \leq n \leq N - 1, \\ 0, & \text{otherwise} \end{cases}$$
It provides a convenient representation of the amplitude variation over the time [10]. It is fact that values of $E_n$ for the unvoiced (i.e. coughing, silence, etc.) components are in general significantly smaller than those of the voiced (i.e. the real crying) components [11]. It can be used as the measurement to distinguish audible sounds from silence when the signal-to-noise ratio is high. The loudness of the crying segments is typically decreasing at the end, in this way the only analysis of the energy content would issue in losing more quiet parts of the crying segments. There are also some cases when the start and the end are louder than the mid of the crying segment, in these cases the only analysis of the energy would issue in cutting the segment to two pieces.

Moreover a main task of the effective AICD is to find the voiced crying sounds and to separate the inspiratory sounds from them. Authors found that in many cases the inspiratory sounds were stuck to the voiced crying sounds in this way the energy function could not distinguish between them. A subsidiary method is needed to analyze the spectral content of the recordings as well to be able to detect and cancel the inspiratory sounds.

B. Extended Harmonic Product Spectrum

The Harmonic Product Spectrum (HPS) is a robust algorithm to determine the fundamental frequency of a multimodal signal [12]. The HPS extracts the fundamental frequency directly from the signal spectrum by decimating the input spectrum by integer factors and computing their product (see Fig. 1.). The input parameter of the HPS is $N$, which refers how many decimated spectrums to determine for the calculation. The primary point for choosing the value of $N$ is the expected number of the subharmonics. Authors found that $N = 9$ is an optimal value for the infant cry in general.

![Fig. 1. Illustration of the Harmonic Product Spectrum (HPS) method.](image)

In 2009 Várallyay stated that the HPS may be capable for describing the spectral content of crying sounds [13]. To prove this statement he utilized the following: if the spectrum is harmonically rich, the HPS will result one enhanced peak at the fundamental frequency [7]. Beyond the position of the HPS peak, some other attributes of the harmonic product spectrum might be informative. It is worthy of note that in case of noise-like signals several peaks can be expected in the harmonic product spectrum, not only one. He defined two new parameters ($H_{\text{max}}$ and $F_{\text{width}}$) by extending the HPS to classify the regularity of the structure of the spectrums (regular structure means harmonic structure). $H_{\text{max}}$ is the intensity of the biggest peak in the product spectrum, $F_{\text{width}}$ means the bandwidth of the product spectrum at the level of $10^{-4}H_{\text{max}}$, see Fig. 2.

Várallyay found that:

- The higher the HPS peak was, the more regular the structure of the original spectrum was, and vice versa.
- The narrower the HPS bandwidth was, the more regular the structure of the original spectrum was, and vice versa.

C. Comparing the energy and the spectral methods

On Fig 3. the outputs of the STE and the EHPS are shown in case of a 12 s long recording. Every method gives information about the cry recording from different aspects. There are five voiced crying sounds in this recording between 0.1-1.3; 1.7-2.7; 3.1-3.8; 7.2-9.7; and 10.2-11.6 s. The 4th has smaller amplitude than the others have. There are audible inspiratory sounds after the 1st, the 2nd, the 4th and the 5th voiced crying sounds.

The Short-Time Energy Function is capable to detect loud voiced crying sounds, while the detection of quiet ones (as between 7.2 and 9.7 s) is less efficient. The $Y$-axis is normalized between 0 and 1. The quiet parts of the recordings have small $E_n$ values (<0.1) while the loud inspiratory sounds and crying segments have bigger $E_n$ values (>0.2).

$H_{\text{max}}$ and $F_{\text{width}}$ obtained by the EHPS refer to the regularity of the spectrum of a short-time crying window. As the range carrier of $H_{\text{max}}$ could overstride more magnitude orders it is logarithmized and normalized between 0 and 1. $H_{\text{max}}$ is at high level (>0.8) continuously in case of crying segments and having significant start

![Fig. 2. The outputs of the original HPS (left) and the Extended HPS (right) methods.](image)
and end points, in this way it can be applied for automatic cry detection with high efficiency. Usually $F_{\text{width}}$ has smaller values (<15 Hz) at the place of the crying segments, but it is very sensitive to the noises in the recording, thus its curve is less continuous.

D. Guidelines for the Automatic Infant Cry Detection

To create the Automatic Infant Cry Detection system authors utilized the following experiments [14]:

- The voiced crying sounds have a greater duration than 250 ms.
- The distance between the inspiratory sounds and the crying segments can be even less than 100 ms, which results that the maximum window length shouldn’t exceed 50 ms.
- In general the energy of the voiced crying sounds and the inspiratory sounds is greater than the background noises.
- The spectrum of the voiced crying sounds has more regular structure than the inspiratory sounds have.
- The recordings might come from different places and devices, in this way the energy and/or the spectral thresholds have to be determined separately for each recording.
- The calculation of the energy function needs less time than the EHPS.

According to the results from Fig. 3., the Automatic Infant Cry Detection should be implemented with the application of $H_{\text{max}}$ obtained from the EHPS. Although authors recommend applying a pre-selection by the energy function as the first main step of the automatic cry detection, in this way the analysis of the spectral content can be focused only on these pre-selected parts. By using the $F_{\text{width}}$ at the last steps of the automatic cry detection it is possible to recognize and separate the inspiratory sounds from the detected crying segments.

E. Main steps of the Automatic Infant Cry Detection

The Automatic Infant Cry Detection was implemented in MATLAB. To illustrate the main steps of the AICD a short recording is shown which contains three voiced crying sounds (Fig. 4/A.). There are two audible inspiratory sounds (before the 2\textsuperscript{nd} and the 3\textsuperscript{rd} voiced crying sounds), a coughing and a sudden noise.

After the DC component extraction and normalization an energy threshold was determined to focus only on the interesting parts of the recording (Fig. 4/F.). The parts shorter than 200 ms were eliminated (Fig. 4/G.). The start and end points of the remaining signals were revised (Fig. 4/J.) and the $F_{\text{width}}$ was applied to find and cancel the inspiratory sounds which had stuck to the voiced sounds (Fig. 4/M.). The spectral contents of the remaining parts were investigated with $H_{\text{max}}$ to find the clear, voiced crying sounds of the recording (Fig. 4/P.)

III. RESULTS

A. Detected voiced crying sounds

With the developed AICD authors detected 2780 voiced crying sounds from 366 recordings. The 95\% of the detected sounds were between 0.3 and 0.2 s in point of their duration. The mean value of the duration of the voiced crying sounds was $0.79 \pm 0.54$ s and the median was 0.91 s. The total length of the recordings was 8753 s, while the total duration of the detected voiced crying sounds segments was 2535 s.

B. Exactness

To test the exactness of the developed Automatic Infant Cry Detection three different recordings (from different
All the detected voiced crying sounds were devoid of the inspiratory sounds.

Regarding to the boundaries of the detected sounds there were a 0.024/0.006 s difference at the start/end points between the manually and the automatically selected sounds on an average.

IV. DISCUSSION AND CONCLUSION

The recognition and the separation of the audible inspiratory sounds are critical tasks and specialty in the cry detection. These inspiratory sounds can be found both before and after the voiced crying sounds.

The Extended Harmonic Product Spectrum method is capable to classify the spectral content of cries, and to distinguish the crying segments from the inspiratory sounds as well. $H_{max}$ is at high level continuously in case of crying segments and having significant start and end points, in this way it can be applied for automatic cry detection with high efficiency. Authors recommend using the $F_{vibh}$ to avoid selecting the inspiratory sounds.

Highly effective automatic cry detection can be accomplished in consideration of the energy content, the spectral content and limitations according to the experienced features of the infant cry. Within these limitations authors recommend utilizing the minimal duration of the crying segments, the wide range of the amplitude of the crying segments and the minimal distance between the inspiratory sounds and the crying segments.

The implemented Automatic Infant Cry Detection can be downloaded from the File Exchange at the MathWork's website from December 2009 [15].

ACKNOWLEDGEMENTS

The authors thank all the hospitals involved in the data collection. Special thanks for their help to Zsolt Farkas and Gábor Katona chief doctors from the Heim Pál Hospital, and Zsolt Szabó chief doctor from the Borsod County Hospital.

This research has been supported by National Office for Research and Technology (NKTH MEC-07-1-2009-0275), Hungarian Scientific Research Foundation (OTKA-T69055) and National Technical Developmental Committee (OMFB-01116/2007).

REFERENCES