AUTOMATIC EXTRACTION OF SPONTANEOUS CRIES OF PRETERM NEWBORNS IN NEONATAL INTENSIVE CARE UNITS

Sandie Cabon
Univ Rennes, Inserm, LTSI - UMR 1099
Rennes, France
sandie.cabon@univ-rennes1.fr

Bertille Met-Montot
Univ Rennes, Inserm, LTSI - UMR 1099
Rennes, France
bertille.met-montot@univ-rennes1.fr

Fabienne Porée
Univ Rennes, Inserm, LTSI - UMR 1099
Rennes, France
fabienne.poree@univ-rennes1.fr

Olivier Rosec
Voxygen
Rennes, France
olivier.rosec@voxygen.fr

Antoine Simon
Univ Rennes, Inserm, LTSI - UMR 1099
Rennes, France
antoine.simon@univ-rennes1.fr

Guy Carrault
Univ Rennes, Inserm, LTSI - UMR 1099
Rennes, France
guy.carrault@univ-rennes1.fr

***Abstract*—Cry analysis has been proven to be an inescapable tool to evaluate the development of preterm infants. However, to date, only a few authors proposed to automatically extract spontaneous cry events in the real context of Neonatal Intensive Care Units. In fact, this is challenging since a wide variety of sounds can also occur (e.g., alarms, adult voice). In this communication, a new method for spontaneous cry extraction from real life recordings of long duration is presented. A strategy based on an initial segmentation between silence and sound events, followed by a classification of the resulting audio segments into two classes (cry and non-cry) is proposed. To build the classification model, 198 cry events coming from 21 newborns and 439 non-cry events, representing the richness of the clinical sound environment were annotated. Then, a set of features, including Mel-Frequency Cepstral Coefficients, was computed in order to describe each audio segment. It was obtained after Harmonic plus Noise analysis which is commonly used for speech synthesis although never applied for newborn cry analysis. Finally, six machine learning approaches have been compared. K-Nearest Neighbours approach showed an accuracy of 94.1%. To experience the precision of the retained classifier, 412 hours of recordings of 23 newborns were also automatically processed. Results show that despite a difficult clinical context an automatic extraction of cry is achievable. This supports the idea that a new generation of non-invasive monitoring of neuro-behavioral development of premature newborns could emerge.**

***Index Terms*—audio processing, spontaneous cries, prematurity, newborns, Neonatal Intensive Care Units, neuro-behavioral development, Harmonic plus Noise Analysis**

I. Introduction

Prematurity is the leading cause of poor neonatal health [1]. Indeed, this is the main cause of mortality, pathology contraction and developmental disorders. Although a dedicated care (i.e., Neonatal Intensive Care Units (NICU), physiological monitoring and specialized medical staff) had been introduced to alleviate these difficulties, solutions to improve the follow-up of the newborns are still sought, especially regarding the evaluation of the neuro-behavioral maturation. To meet this clinical need, the Digi-NewB project, funded by the European Union programme for Research and Innovation Horizon2020, aims to develop a new generation of monitoring system using non-invasive modalities, such as microphones and cameras.

Analyses of spontaneous cries have been shown to be an informative tool to assess the development of premature newborns [2]. More precisely, some authors discussed relationship between frequency content and increasing gestational age [3] and found that fundamental frequency F_0 is generally higher in preterm than in full-term newborns at term-equivalent ages [4], [5]. However, in these studies, authors focus their analyses either after a manual extraction of cry events [4] or on short recordings [3], [5]. To move towards continuous monitoring, techniques to automatically retrieve spontaneous cries have to be proposed. Extracting newborn cries is challenging due to the fact that in NICU, several sounds from other sources (e.g., alarms, adult voices) can occur. Besides, conditions of recording differ regarding the gestational age (GA) and PostMenstrual Age (PMA) of each newborn (e.g., type of room and bed). To date, only a few studies tackled this problem, either using Convolutional Neural Networks (CNN) [6] or Hidden Markov Models (HMM) [7], [8]. Nevertheless, in these studies, no audio recording of long duration was processed.

In this communication, we present a new strategy to automatically extract spontaneous cry of newborns from long duration recording. Then, the process is evaluated in two steps: the first being by comparing performances of six machine learning approaches and the second, after the deployment of the best one on a larger dataset.

II. Methods

Our approach is rooted in the unsupervised segmentation method proposed in [9]. This method leans on energy thresholding where two thresholds are automatically estimated on each recording by the mean of the Otsu method [10].

When applied to an audio recording containing only cries, this approach will return only cry intervals. However, when applied on long recordings, irrelevant sounds segments (e.g., alarms, adult voices) are also extracted. Thus, we decided to apply it, as a first step, to extract segments of interest from silence periods. Only segments with a duration between 250 milliseconds and 5 seconds are retained. Then, in order to provide a relevant characterization of behavioral development, segments of cries have to be automatically recognized among them. Hence, a workflow based on segmentation and machine learning was investigated, as presented in Figure 1. To construct this workflow, four steps were involved: annotation of a relevant database, feature extraction, dimensionality reduction and selection of a classifier.

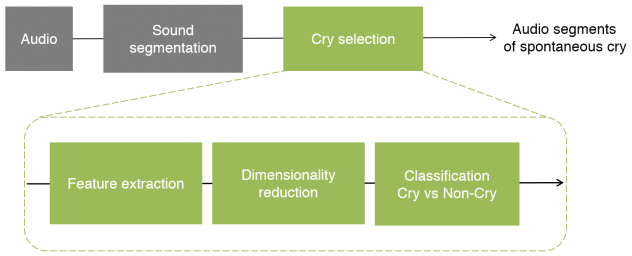

Fig. 1. Global workflow of cry extraction.

A. Annotated database

This communication belongs to the DigiNewB project. During the project, multimodal data (video, audio, clinical signs and electrophysiological) have been recorded in six french hospitals, representing more than 37 000 hours of recording. From this database, 27 recording representing a large part of the diversity encountered in NICU were selected. They involve fourteen boys and seven girls born between 27+5 and 41+4 GA and recorded between 28+5 and 41+5 PMA. Audio recordings were acquired at 24 kHz using an omni-directional microphone (FG-23329-P07) marketed by Knowles Acoustic, on/in different types of beds (i.e., radiant warmer, cradles and incubators). Microphone was placed between 30 centimeters and 1 meter of the newborn head. In closed bed, it was placed at newborn' feet while for open bed, it was set at his/her head.

First, audio segments were extracted using the segmentation step. Then, 637 segments were manually labeled as cry or as non-cry segments. We chose to annotate segments containing only one type of event, meaning that segments with overlapping sounds (e.g., cry with adult voice) were not selected. Figure 2 summarizes the annotated data.

A total of 198 cries events were annotated. Since the auditory differences between vocalizations and cries are sometimes subjective, only obvious cries were annotated as such. All other sounds have been considered as non-cry events, resulting into 439 segments. In this category, several events were included: vocalizations (e.g., coing), other baby noises (e.g., coughing or hiccups), adults' voices, alarms from devices and background noises. The most diverse segments were selected. Thus, men and women voices, several types of alarms, many background noises coming from the adults activity (e.g., doors opening/closing, packaging friction, water flowing from the tap) as well as from devices (e.g., ventilator support airflow, bed adjustment noises), were selected.

Fig. 2. Overview of the annotated audio data.

B. Feature extraction

In recent works, authors mainly used Mel-Frequency Cepstral Coefficients (MFFCs) to describe audio signals for classification [2]. In some cases, preprocessing steps were first applied in order to reduce the effect of noise, such as beamformer [6] or signal decomposition [7]. Coefficients were then computed frame by frame. Finally, each frame was classified by taking into account adjacent frames either using CNN [6] or using HMM [7], [8].

In this work, we chose to integrate noise by modeling audio segments using Harmonic plus Noise Model (HNM), commonly applied in speech synthesis [11]. HNM analysis is known to be more suitable for quasi-harmonic signals, which is the case of baby cries and vocalizations, adult voices and alarms. In addition, to perform the analysis, it is necessary to limit the analysis in a certain frequency band that we chose to adapt to cry analysis [150-750 Hz]. The underlying hypothesis of these choices is that an analysis focused on extracting characteristics relevant to cry analysis will give discriminating features for classification in case of other types of sounds. The principle of HNM analysis is to create a synthetic signal $s(t)$ composed of harmonic $h(t)$ and noise $n(t)$ parts that fit the original signal such as:

$$s(t) = h(t) + n(t) \tag{1}$$

In fact, the spectrum of voiced speech signal can be divided into two bands bounded by a maximum voiced frequency, a time varying parameter. The lower band is related to the harmonic part and the upper band to the noise part. The harmonic part $h(t)$ is modeled as a sum of harmonics such as:

$$h(t) = \sum_{k=1}^{K(t)} A_k(t) \cos(k\theta(t) + \Phi_k(t)) \tag{2}$$

where $A_k(t)$ and $\Phi_k(t)$ are respectively amplitude and phase at time t of the k-th harmonic. $K(t)$ represents the time-varying number of harmonics included in the harmonic part. On its part, the noise part $n(t)$ is supposed to have been obtained by filtering a white Gaussian noise $u(t)$ by a time-varying,

normalized all-pole filter $f(t,\tau)$ and multiplying the result by an energy triangular-like envelope function $e(t)$, as follows:

$$n(t) = e(t)[f(t,\tau) * u(t)] \quad (3)$$

To fit the model, an initial estimation of F_0 must be provided. In our case, we chose to estimate it by the mean of Continuous Wavelet Transform, as proposed in [12]. Once the signal is modeled, the spectral conversion of $h(t)$ is performed to extract MFCCs. The analysis is performed over an audio segment on frames of 5 milliseconds without overlap. The audio segment is then summarized by median values of each feature coming from HNM analyses of each frame.

Features provided by the HNM analysis had been supplemented by two features of the time domain:

- Total duration in seconds of the segment since some type of sounds may last longer (e.g., adult speech) or shorter (e.g, beep) than typical cries;
- Zero Crossing Rate ZCR, shown to be useful to distinguish alarms from cries [13]:

In total, 73 features are computed. They are synthesized in Table I.

TABLE I
LIST OF THE COMPUTED FEATURES FOR CLASSIFICATION

Type of feature	Estimation Method	Number of instances
Fundamental frequency	HNM	1
Number of harmonics	HNM	1
Harmonic amplitudes	HNM	18
Harmonic phases	HNM	14
Gain	HNM	1
Filter coefficients	HNM	20
Cepstral coefficients	HNM	16
Zero Crossing Rate	ZCR	1
Duration	Duration	1

C. Dimensionality reduction

A common problem in classification is overfitting. This occurs when a classifier corresponds too closely or exactly to a particular set of data and may therefore fail to fit additional data. This can be due to a too high number of features describing each sample. To prevent this situation and reduce the dimension of our feature set, Principal Component Analysis (PCA) was applied. Data are first standardized to avoid the prevalence of a feature in the dimensionality reduction process. Then, there are two ways of performing dimension reduction with PCA, either by keeping a share of the total variance or by targeting a number of principal components. In our case, we chose to keep the principal components that represent 95% of the total variance.

D. Classifiers

Recent works proposed to exploit temporality for classification (e.g., HMM). In this paper, we chose to extract segments of interest first and summarized them by median values of the feature set. From this feature set, we proposed to train and compare six commonly known classification approaches: Linear Discriminant Analysis, Logistic Regression, K-Nearest Neighbours, Random Forest, Multi-Layer Perceptron and Support Vector Machine.

Our goal being to identify cry segments from others, a binary classification was performed. A training/validation set and a testing set were defined, respectively composed of 60% and 40% of the dataset. On the training/validation set, a k-fold cross validation (k = 3) was performed in order to tune parameters of the classifiers. Parameters were tuned within the objective to reach the highest precision. In that case, the lowest false positive rate was sought, i.e. recover as few non-cry segments as possible.

III. RESULTS

In this section, several results are presented. First, the outcomes of the dimensionality reduction process are discussed. Then, regarding classifiers, retained parameters during the training/validation step and performances on the testing set of each model are presented. To finish, the whole method is evaluated after deployment on a new set of data. Experiments were conducted using scikit-learn 0.22.1 and Python 3.7.3.

A. Dimensionality reduction

As a first step, data have been projected into a 2-dimensional space for visualization using PCA. Resulting graph is presented in Figure 3.

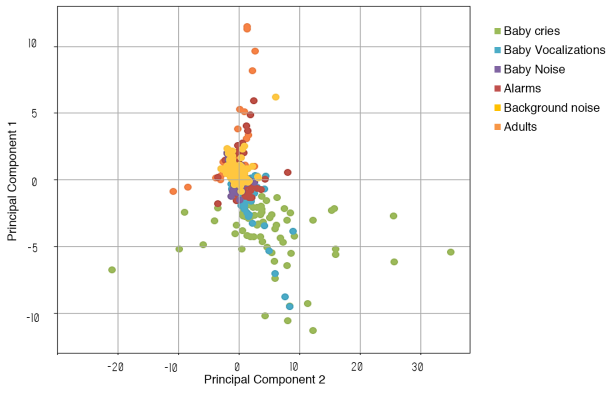

Fig. 3. Visualization of the dataset using the first two principal components.

One can see that baby cries are mostly located on the bottom of the graph with some of the baby vocalizations while other sounds tend to be in the upper band. This observation is reassuring regarding our classification purpose although we may already notice that it will be difficult to discriminate between baby cries and vocalizations on the sole basis of these two dimensions.

In a second time, we projected the feature set into principal components that represent 95% of the total variance, resulting into a projected feature set of 41 components.

B. Classification results

1) Tuning of the parameters on the training/validation set: For each classifier, several parameters and hyper-parameters

have been tested in order to reach the highest precision during cross-validation. A summary of these tests is reported in Table II. For each method, best parameters are marked in bold.

TABLE II
PARAMETERS TESTING SUMMARY. FINAL SELECTING SETS OF PARAMETERS ARE MARKED IN BOLD.

Method	Parameters
KNN	Number of neighbours ∈ [1, 3, 5, **11**, 15] Distance: Manhattan or **Euclidean**
LDA	Solver ∈ [singular value decomposition, **least squares solution**, eigenvalue decomposition]
LR	Cut-off ∈ [**0.1**, 0.2, 0.5, 0.7]
RF	Number of trees ∈ [5, 10, 20, 50, 100, **300**] Quality split criterion: gini or **entropy**
MLP	Number of hidden layer ∈ [**1**, 2, 5] Number of perceptron per layer ∈ [2, 5, 10, **20**, 30] Activation function ∈ [identity, logistic sigmoid, hyperbolic tan, **rectified linear unit**]
SVM linear	No additional parameter
SVM polynomial	degree ∈ [1, 2, **3**, 4]
SVM Gaussian	margin ∈ [0.01, 0.1, **1**, 10, 100, 10^3, 10^4] gamma ∈ [0.0001, 0.001, 0.01, 0.1, 1, **5**, 10, 100]

2) Results on the test set: Classification results are reported in Figure 4. A good generalization is observed for all models since values are quite stable between metrics.

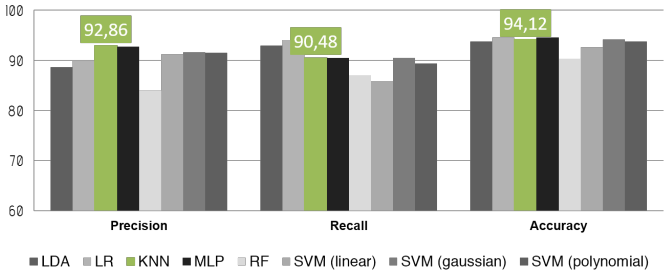

Fig. 4. Performances (in %) of cry selection on the test set for each machine learning approach by maximizing the precision in the learning phase.

The best precision score is obtained by KNN and reaches 92.9%. The highest recall score is obtained by LR, with 94.1%. Accuracies are high (above 90.2%) for all classifiers. The results with MLP are also really good since a precision of 92.7%, a recall of 90.48% and an accuracy of 94.5% are reached. With regard to our objective, a classifier that has learned to be precise on cry selection seems to be the most reasonable choice. Indeed, the goal is to extract cries in order to assess newborn evolution during hospitalization, in other words, over the long term. In that case, there is no need to get all the cries provided that only cries are extracted. The results of the KNN are in line with this. In fact, it carries a high precision while keeping a high recall value. This means that there is also a low chance of missing cries with this model.

C. Evaluation of the model during deployment

In this section, the precision of the model during deployment is assessed. For that purpose, the method was applied on 42 recordings of 23 newborns (10 girls and 13 boys) of several GA (from 25+6 to 40+3) and PMA (from 28+1 to 41+3). The median duration of recordings was about 8 hours, giving a total of 412 hours. After the segmentation step, 495534 sound events were obtained. It was reduced to 5409 segments using the KNN model. If we take a closer look, they represent a total duration of 1 hour and 48 minutes. Segments that were automatically classified as cries were then verified and percentages of good classifications and misclassifications were computed in regard to two metrics: number of segments and duration of these segments. It has been reported in Figure 5.

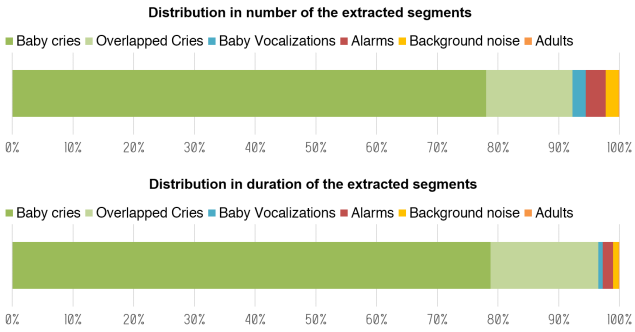

Fig. 5. Cry extraction results during the deployment. Percentages of good classifications and misclassifications regarding number (top) and duration (bottom) of the extracted segments.

Firstly, among the segments, 4221 revealed to be true positives, giving a precision of 78%. The second most represented type of segments is the one on which another sound event happened in the same time than a cry (14.2% of overlapped cry). If percentages are recomputed by integrating overlapped cries to cries, precision reaches 92.1% and actual errors rate drops at 7.9%.

Secondly, alarms represent 3.3% of the misclassified segments, followed by baby vocalizations (2.1%), background noise (2.1%) and adults (0.1%). However, these misclassifications represent only 4 minutes. In relation to the 412 hours analyzed, it gives a rate of 0.02%.

To complete these observations, we can also report that processing duration depends on the content of each recording, notably because of the time necessary to perform HNM analysis for classification. It is directly linked to the number of segments issued from the segmentation step and their duration. Here, computational duration ranged from 45 minutes to 6 hours with a mean of 3 hours for all 8-hours recordings.

IV. DISCUSSION AND CONCLUSION

In this paper, a process was proposed to automatically extract cry events from long audio recordings of preterm newborns. First, a segmentation of the periods of interest was performed and then, a classification approach for cry selection has been developed. This approach was a multi-faceted problem since a relevant database had first to be manually

collected and annotated, the right ensemble of features had to be defined and classifiers had to be correctly trained.

As a first step, a high number of segments (637 segments during the training and 5409 segments during the deployment) has been manually annotated. It is very important in the establishment of a database and constitutes a real contribution for further method comparisons. In fact, this dictionary contains a wide variety of sound events coming from real conditions of NICU in several hospitals. Additionally, in this study, a particular attention has been paid to collect cries of newborns of different gestational ages (from extremely preterm to full-term) and postmenstrual ages (from 28+1 to 41+5 weeks). Hence, the proposed method fits a multitude of clinical configurations and to the entire newborn population.

Secondly, for the first time, Harmonic plus Noise analysis, generally used in speech synthesis, has been applied within the objective of cry extraction. Features obtained from this analysis have proven their effectiveness since high classification performances were obtained. Indeed, performances of the KNN classifier during the testing reached 92.9% of precision and 90.5% of recall, with an accuracy of 94.1%. This is in adequation with the literature where accuracies of 89.2% was reported with HMM [8] and 86.6% with CNN [6].

Thirdly, an automatic classification of sounds has been performed on real life data of long duration with the proposed method. About 412 hours of recordings were analysed and we showed that the classifier performed well. Indeed, a precision of 92.1% was retrieved when including cries and overlapped cries. Additionally, computation times may allow an application in near-real time.

To date, it is now possible to automatically extract a large number of cries. This will help to resolve several limitations in cry extraction, that raised up during this work. Indeed, a 7.9% error rate was reported since some alarms, baby vocalization, background noises and adults were misclassified. To reinforce the model, the annotated database could be updated by adding these segments. It may also be relevant to perform the HNM analysis simultaneously in an upper band ($>$750 Hz) to enhance the feature set. In fact, some of the alarm segments that were misclassified have a frequency above 750 Hz and thus, could be characterized using HNM. It could also be relevant to discard background noises that occupy a broad band of frequencies as well as to detect cry events with overlapping sounds.

Future prospects concern clinical analysis. Once cries are extracted, they can be analyzed in order to evaluate the neuro-behavioral development of the newborn. As an example, the evolution of the fundamental frequency can be studied. To go further, an automatic search of the relevant band of frequencies to perform this estimation will have to be proposed. In fact, although the frequency band [150-750 Hz] has been proven effective for cry extraction, it may not be the case for a precise analysis of the melody of some cries. In fact, over a cry the fundamental frequency may be superior to 750 Hz, such as in hyperphonation cries. Moreover, the impact of overlapping sounds on cry characterization will have to be studied.

To finish, studies of the literature have not proposed a strict definition of a cry. To do so, it might be instructive to build on the work of Golub *et al.* which proposed a physio-acoustic model of the infant cry [14].

Ultimately, these results show that it is possible to imagine a non-invasive generation of monitoring to assist clinicians in their evaluation of neuro-behavioral development of preterm newborn. In addition, this work may be the basis to carry out new studies regarding the vocal development of the newborn, such as regarding the impact of intubation.

ACKNOWLEDGMENT

Results incorporated in this publication received funding from the European Union's Horizon 2020 research and innovation program under grant agreement no. 689260 (Digi-NewB project)

REFERENCES

[1] World Health Organization, "Born too soon: the global action report on preterm birth," 2012.

[2] S. Cabon, F. Porée, A. Simon, O. Rosec, P. Pladys, and G. Carrault, "Video and audio processing in paediatrics: a review," *Physiological Measurement*, vol. 40(2), pp. 1–20, 2019.

[3] C. Manfredi, L. Bocchi, S. Orlandi, L. Spaccaterra, and G. P. Donzelli, "High-resolution cry analysis in preterm newborn infants," *Medical Engineering & Physics*, vol. 31, no. 5, pp. 528–32, 2009.

[4] Y. Shinya, M. Kawai, F. Niwa, and M. Myowa-Yamakoshi, "Preterm birth is associated with an increased fundamental frequency of spontaneous crying in human infants at term-equivalent age," *Biology Letters*, vol. 10, no. 8, 2014.

[5] S. Orlandi, C. A. R. Garcia, A. Bandini, G. Donzelli, and C. Manfredi, "Application of pattern recognition techniques to the classification of full-term and preterm infant cry," *Journal of Voice*, vol. 30, no. 6, pp. 656–663, 2016.

[6] D. Ferretti, M. Severini, E. Principi, A. Cenci, and S. Squartini, "Infant cry detection in adverse acoustic environments by using deep neural networks," in *2018 26th European Signal Processing Conference (EUSIPCO)*, 2018, pp. 992–996.

[7] L. Abou-Abbas, C. Tadj, C. Gargour, and L. Montazeri, "Expiratory and inspiratory cries detection using different signals' decomposition techniques," *Journal of Voice*, vol. 31, no. 2, pp. 259.e13 – 259.e28, 2017.

[8] G. Naithani, J. Kivinummi, T. Virtanen, O. Tammela, M. J. Peltola, and J. M. Leppänen, "Automatic segmentation of infant cry signals using hidden Markov models," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2018, no. 1, pp. 1–14, 2018.

[9] S. Orlandi, C. Manfredi, L. Bocchi, and M. Scattoni, "Automatic newborn cry analysis: A non-invasive tool to help autism early diagnosis," in *Engineering in Medicine and Biology Society (EMBC), 2012 Annual International Conference of the IEEE*. IEEE, 2012, pp. 2953–2956.

[10] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 9, no. 1, pp. 62–66, 1979.

[11] Y. Stylianou, "Harmonic plus noise models for speech, combined with statistical methods, for speech and speaker modification," *Ph. D thesis, Ecole Nationale Superieure des Telecommunications*, 1996.

[12] S. Orlandi, A. Guzzetta, A. Bandini, V. Belmonti, S. D. Barbagallo, G. Tealdi, S. Mazzotti, M. L. Scattoni, and C. Manfredi, "AVIM—a contactless system for infant data acquisition and analysis: Software architecture and first results," *Biomedical Signal Processing and Control*, vol. 20, pp. 85–99, 2015.

[13] G. Várallyay, "Future prospects of the application of the infant cry in the medicine," *Periodica Polytechnica Electrical Engineering*, vol. 50, no. 1-2, pp. 47–62, 2006.

[14] H. L. Golub and M. J. Corwin, "A physioacoustic model of the infant cry," in *Infant crying*. Springer, 1985, pp. 59–82.