On the Acoustic Environment of a Neonatal Intensive Care Unit: Initial Description, and Detection of Equipment Alarms

Ganna Raboshchuk1, Climent Nadeu1, Omid Ghahabi1, Sergi Solvez1, Blanca Muñoz Mahamud2, Ana Riverola de Veciana2, and Santiago Navarro Hervas2

1TALP Research Center, Department of Signal Theory and Communications, Universitat Politècnica de Catalunya, Barcelona, Spain
2Hospital Sant Joan de Déu Barcelona, Esplugues de Llobregat, Barcelona, Spain
{ganna.raboshchuk, climent.nadeu, omid.ghahabi, sergi.solvez}@upc.edu, {bmunozm, ariverola, snavarroh}@hsjdbcn.org

Abstract
The acoustic environment of a typical neonatal intensive care unit (NICU) is very rich and may contain a large number of different sounds, which come either from the equipment or from the human activities taking place in it. There exists a medical concern about the effect of that acoustic environment on preterm infants, since loud sounds or particular sounds may be harmful for their further neurological development. In this work, first of all, an initial description of the acoustic characteristics of the NICU has been carried out using a set of diverse recordings produced with microphones placed both inside and outside an incubator. Then, the work has focused on detection of the most relevant types of sounds. In this paper, after describing the recorded database and the acoustic environment, preliminary experiments for detection of the acoustic alarms of devices are reported. The proposed detection system is based on Deep Belief Networks (DBN). The experimental results show that the DBN-based system is able to achieve better results than a baseline GMM-based system.

Index Terms: acoustic event detection, acoustic analysis, neonatal intensive care unit, deep belief networks

1. Introduction
Newborns delivered at a gestational age of 24-32 weeks (very low birth weight preterms) commonly have health problems and must be admitted to the Neonatal Intensive Care Unit (NICU), which, in most of the cases, is crucial for their survival. The increased survival and reduced neonatal morbidity of preterm infants in the past three decades has not always been accompanied by an improvement in their neurological development [1]. The negative or stressful environmental impact of NICUs on the developing brain has been widely documented [2, 3, 4, 5].

At present it is known that when a premature passes from an ideal intrauterine environment to an environment with multiple unexpected stimuli (light, noise, proprioceptive stimuli) this may have a negative effect on its neurodevelopment. Inadequate, unexpected noises replace natural hearing placental stimulation [6]. Usually, to respond to so many extraneous stimuli this, first of all, an initial description of the acoustic characteristics of the NICU has been carried out using a set of diverse recordings produced with microphones placed both inside and outside an incubator. Then, the work has focused on detection of the most relevant types of sounds. In this paper, after describing the recorded database and the acoustic environment, preliminary experiments for detection of the acoustic alarms of devices are reported. The proposed detection system is based on Deep Belief Networks (DBN). The experimental results show that the DBN-based system is able to achieve better results than a baseline GMM-based system.

Index Terms: acoustic event detection, acoustic analysis, neonatal intensive care unit, deep belief networks

1. Introduction
Newborns delivered at a gestational age of 24-32 weeks (very low birth weight preterms) commonly have health problems and must be admitted to the Neonatal Intensive Care Unit (NICU), which, in most of the cases, is crucial for their survival. The increased survival and reduced neonatal morbidity of preterm infants in the past three decades has not always been accompanied by an improvement in their neurological development [1]. The negative or stressful environmental impact of NICUs on the developing brain has been widely documented [2, 3, 4, 5].

At present it is known that when a premature passes from an ideal intrauterine environment to an environment with multiple unexpected stimuli (light, noise, proprioceptive stimuli) this may have a negative effect on its neurodevelopment. Inadequate, unexpected noises replace natural hearing placental stimulation [6]. Usually, to respond to so many extraneous stimuli in an organized way is difficult for a preterm newborn. An important negative effect of noise in the NICU is the one it has on sleep, which is essential for the neurodevelopment of prema-

The problem is aggravated because in NICUs the maximum noise limits recommended [8] are exceeded frequently [9]. Different ways that have been proposed to deal with this problem can be divided into two groups. The major group of methods is aimed at analyzing and changing the acoustical environment of the NICU (for example, by planning a more rational distribution of the wards [10] or controlling and reducing the activities taking place in it [11]). The other group of methods directly concerns the preterm baby and implies protecting a baby with special accessories (earmuffs) [12, 13], but although a prolonged time of quiet sleep was observed, the positive effect on the physiological variables has not been proven completely, and there exists a risk of tactile overstimulation and negative long-term outcomes (e.g. auditory deprivation) [14]. The line of work pursued in this paper, which consists of analyzing the acoustical environment of a preterm baby, refers to the first group of methods.

The acoustic environment of a preterm baby introduced into a NICU has been the object of a number of reported studies during the last two decades. These works analyze the environment placing a microphone both inside [15] and outside the incubator. Usually, sound is represented only by its intensity level and just a few works analyze sound spectra [16]. To our knowledge, very little studies considered the intensity levels of specific sounds [17] or analyzed specific conditions of the NICU acoustic environment [18]. Unlike most previous works, in this work the whole content of the audio signal is used, not only its intensity level.

The main objective of this work is the analysis of the acoustic environment of the preterm infant in a NICU. In particular, in this paper we present the creation of a (limited) annotated audio database, an initial audio description, and the first results of automatic detection of equipment alarms with Gaussian Mixture Models (GMM) and Deep Belief Networks (DBN) based systems using the produced database. The focus on equipment alarms is due to the large extent they are present in the recordings, and the fact that an automatic monitoring system aiming to detect alarms may help to create a safer environment for a preterm in the NICU.

Deep Belief Networks (DBN) have recently opened a new research line in image, audio and speech processing areas (e.g., [19, 20, 21, 22, 23, 24]). DBNs are originally generative networks which can be trained by a greedy layer-wise algorithm using Restricted Boltzmann Machines (RBMs) [25, 26]. However, by adding a top label layer and using a standard backpropagation algorithm, these generative DBNs can be converted to discriminative ones [26, 27] and, therefore, they are appropriate
we observe sounds which are continuous (like chair moving, and this list is not exhaustive. This number was surprisingly acoustic events happening in the NICU environment were found acquisition process is still in progress. 

In this subsection we provide the results of an initial acoustic analysis using the data that we have collected so far. The data

for classification tasks.

The paper is organized as follows. Section 2 contains the description of the the produced database and the acoustic environment. In Section 3, we define a metric, present the detection systems and discuss the experimental results of the alarms detection task.

2. Database description

2.1. Database acquisition

The acoustic analysis and the experimental evaluations presented in this paper were performed using the database recorded in the NICU of Hospital Sant Joan de Deu (HSJD) (see Figure 1). In total, ten recording sessions were carried out, both in the morning and in the afternoon. The overall duration of the acquired audio data is 108.7 minutes. Two electret unidirectional microphones connected to the Olympus LS-5 Linear PCM Recorder were used to make recordings. One microphone was placed inside the incubator, close to the infant’s ear, and the other one outside the incubator, at approximately 50 cm distance above it.

Each recording session includes a subset of the defined scenarios (in average there are 8.6 samples per scenario). Ten different scenarios were considered: nine of them correspond to the daily nursery care related activities, and, also, a Neutral scenario was defined for including the time periods when no other scenario takes place and the doors of the incubator are closed.

The audio data was manually annotated using the ELAN tool [28]. Each alarm signal (the tone itself) was annotated separately.

2.2. Initial audio description

In this subsection we provide the results of an initial acoustic analysis using the data that we have collected so far. The data acquisition process is still in progress.

A typical NICU environment is characterized by a large diversity of sounds. During our study more than 65 different acoustic events happening in the NICU environment were found and this list is not exhaustive. This number was surprisingly high even for the medical staff working there. Among these events, we identified at least 13 different types of alarms.

The sounds found in the NICU environment have diverse spectro-temporal characteristics. Regarding the time dimension, we observe sounds which are continuous (like chair moving, or drawer) or impulsive (like knocks, steps, or door slam); sounds which are periodic (like alarms, or CPAP noise) or aperiodic (like spray, or plastic wrapping). A great diversity is also present in spectral domain, were we found lower- (<3 kHz; e.g., chair moving), and higher-frequency (e.g., squeaks) sounds, as well as acoustic events whose content is spread across all the frequency range.

With regards to the spectro-temporal characteristics of the alarm sounds, first, they are periodic in time. Each period consists of a signal segment and a silence segment. The signal itself is periodic and has fundamental frequencies distributed up to 5-6 kHz, and most of them are around 1.5 kHz.

In the NICU environment sound sources are extensively overlapped in time. Concerning the alarm sounds, in most cases they overlap with other sounds, but the overlaps between different alarms are not rare either. For example, for the data used throughout the experiments, the statistics of time when several alarms happen simultaneously is the following: 2 alarms - 8.54%, 3 alarms - 0.88%, 4 alarms - 0.14% of total time labelled as alarm.

During the acoustic analysis of the recordings we found two specific types of noise produced by equipment. In most of the cases these noises are simultaneous, but may also happen individually, and are present throughout the recordings. The first type of noise is a narrow-band noise at 15 kHz frequency. That noise has short temporal interruptions and can be considered periodic: 5.5 s of noise are followed by 1-1.5 s of pause. In the vast majority of the cases the noise is stronger outside than inside the incubator. The second type of noise is the ventilation noise, which is usually spread over a wide frequency range. Depending on the recording session (i.e. equipment used) the noise is stronger either inside or outside the incubator. There are several different types of ventilation equipment in the NICU, having noises with different spectral characteristics. Depending on the particular needs of a preterm infant a corresponding type of ventilation is used, thus this introduces a lot of variability to the data.

In general, we have observed a strong inter-session variability caused by differences in: incubators and accompanying equipment used during recording sessions; their position (one of the four positions depicted in Figure 1 were used); nursing procedures depending on the nurse personal style of work and the particular preterm infant needs; acoustic environment in general (number of people in the room, sounding alarms, etc.).

3. Detection systems

3.1. Audio enhancement

The original 44.1 kHz recordings were downscaled to 24 kHz, thereby removing the high-frequency equipment noise at 15 kHz, but still preserving all the relevant information.

In order to deal with the stationary ventilation noise present in the recordings, we applied the Spectral Subtraction (SS) technique together with the Minima-Controlled Recursive-Averaging (MCRA) algorithm for estimation of the noise spectrum [29].

3.2. Baseline system

The baseline result for the alarms detection problem was obtained using a Gaussian Mixture Model (GMM) based classifier. There are two models: the alarms (AL) model, and the non-alarm (NON-AL) model. Each model consists of two Gaussians with diagonal covariance matrix. The HTK toolkit [30]...
was employed for training and testing this GMM-based system. The input signal is successively framed with a Hamming window, being the frame length 30ms and the frame shift 10ms. Concerning the features, 18 frequency-filtered log filter-bank energies (FF-LFBE) [31] along with their 18 first temporal derivatives are extracted from each frame. Therefore, the dimension of the feature vector is 36.

With the likelihoods obtained from the two models, each frame is classified as either AL or NON-AL.

3.3. DBN-based system

DBNs are originally probabilistic generative models with multiple layers of stochastic hidden units above a layer of visible variables which represent an input vector (e.g., see Figure 2). The algorithm treats every two adjacent layers as a Restricted Boltzmann Machine (RBM) network, which is constructed from a layer of binary stochastic hidden units and a layer of stochastic visible units. The output of each RBM is considered as the input to the next RBM.

There is an efficient greedy layer-wise algorithm for learning DBNs [26]. The algorithm consists of three steps. At first, hidden states (h) are computed given visible states (v); then, using those h, v is reconstructed; and, on the third step, h is updated, given the reconstructed v. Finally, the change of connection weights is given as follows,

$$\Delta w_{ij} \approx -\alpha \left( (v_i h_j)_{data} - (v_i h_j)_{recon} \right), \tag{1}$$

where $\alpha$ is the learning rate, $w_{ij}$ represents the connection weight between the visible unit i and the hidden unit j, $\langle \cdot \rangle_{data}$ and $\langle \cdot \rangle_{recon}$ denote the expectations when the hidden state values are driven respectively from the input visible data and the reconstructed data.

In fact, the training process tries to minimize the reconstruction error between the actual input data and the reconstructed one. The parameter updating process is iterated until the algorithm converges. Each iteration is called an epoch. It is possible to perform the parameter update after processing each training example, but it is often more efficient to divide the whole input data (batch) into smaller size batches (minibatch) and to do the parameter update by an average over each minibatch. More theoretical and practical details can be found in [25, 26, 32].

When the unsupervised learning is finished, it can be converted to a discriminative model by adding a label layer (o) on top of the network and doing a supervised backpropagation training. In other words, the unsupervised learning can be considered as a pre-training for the supervised stage. It has been shown [26] that this unsupervised pre-training can set the weights of the network to be closer to a good solution than random initialization and, therefore, avoids local minima when using supervised gradient descent.

Only one hidden layer networks are explored in this paper (see Figure 2). The hidden layer has 512 units. A Gaussian-Bernoulli RBM [32, 27] is employed as the first RBM. Experimentally, the size of each minibatch is set to 20 and the inputs are randomly distributed among minibatches. The feature vector consists of 13 successive frames of 18 filter-bank energies (FBEs) and is labelled with the class of the middle frame. Thus, the dimension of the feature vector is 234. The input vectors are mean-variance normalized before being fed to the network; the mean and variance values calculated on the training data are further applied to the testing data. The training data is balanced with regards to classes by randomly selecting samples of the predominant class.

The learning rate ($\alpha$), the number of epochs (NoE), and the momentum in the unsupervised stage are set, respectively, to 0.00001, 27, and 0.9. The supervised learning is then carried out with $\alpha = 0.001$, NoE=200, and a fixed momentum of 0.9. The weight decay for both unsupervised and supervised stages is set, correspondingly, to $2 \times 10^{-3}$ and $1.2 \times 10^{-7}$.

If the input labels in the training phase are chosen as ($l_1 = 1, l_2 = 0$) and ($l_1 = 0, l_2 = 1$) for AL and NON-AL classes respectively, the final output score in the testing phase will be computed in a Log Likelihood Ratio (LLR) form as follows,

$$LLR = \log(\alpha_1) - \log(\alpha_2),$$

where $\alpha_1$ and $\alpha_2$ represent respectively the output of the first and the second units of the top layer. The LLR is further converted to a hypothesized label by choosing the threshold that gives the minimum $C_{2D+E}$ value at the frame level (see Subsection 3.5).

3.4. Classifier output post-processing

At the output of both classifiers, the resulting time string of hypothesized labels is smoothed to force grouping of frame labels in order to obtain event labels. Within each of the consecutive non-overlapping smoothing windows of odd length the class to which most of the output frames belongs is determined. Then, all the frames within the window are assigned the label of that winner class. For this task, we found a duration of smoothing window equal to 39 frames to be the most appropriate for a GMM-based system and to 11 frames for the DBN-based one.

3.5. Evaluation metrics

The system recognition performance was evaluated using two metrics, that operate at either the frame level or the block level. Let’s denote with AL the alarm label and NON-AL the non-alarm label.

The frame-based or frame-level accuracy metric is defined as one minus the relative system error:

$$FB-ACC = 1 - \frac{\text{FalseAlarms} + \text{Misses}}{\text{TotalFramesNumber}}, \tag{3}$$

where Misses and FalseAlarms are the numbers of misclassified frames corresponding, respectively, to AL and NON-AL reference labels; and TotalFramesNumber is the total number of frames evaluated.

In evaluation campaigns about acoustic event detection that have taken place so far, an event-based metric was defined, typically using the F-score [34, 35]. In those cases, the classification task was not binary but multi-class, and several classes could happen simultaneously. In our application, to define an event-level metric, we could use as event definition either an alarm period or a whole alarm sound. None of them is much meaningful for the purposes of our application, as there are alarm sounds that span over a long time interval. Therefore, we decided to
use a metric with a time span that is a trade-off between them. Inspired by [33], we call it a block-based metric.

To compute that block-based metric, the input audio stream is divided into consecutive non-overlapping blocks of 5s length. For each of them a label (AL or NON-AL) is assigned using the following criterion: the block is labelled as AL in case it has more than one alarm signal; otherwise it is labelled as NON-AL. The basic idea that is being pursued is that neither the staff nor the preterm baby respond to only one alarm signal, but there should occur several of them (we believe from 2 to 4 periods of signal-silence, are enough) in order that the sound is perceived as alarm.

The defined block-based metric is based on the detection cost function \((C_{Det})\) used in NIST evaluations [36], and is computed using the formula:

\[
BB-ACC = 1 - \alpha \cdot \left( \frac{C_{Miss} \cdot P_{Miss|Target} \cdot P_{Target}}{C_{FA} \cdot P_{FA|NonTarget} \cdot P_{NonTarget}} + (C_{FA} \cdot P_{FA|NonTarget} \cdot P_{NonTarget}) \right),
\]

where \(P_{Target} = 0.59\) and \(P_{NonTarget} = 0.41\) are, respectively, the fractions of AL and NON-AL blocks calculated over all the database; \(C_{Miss} = 0.3\) and \(C_{FA} = 0.7\) are estimated application-specific costs of misses and false alarm errors; \(\alpha\) is a normalization factor equal to a fraction of 1 by a score of the system, that is always wrong; and, finally, \(P_{FA|NonTarget}\) and \(P_{Miss|Target}\) are computed by the following expressions:

\[
P_{FA|NonTarget} = \frac{FalseAlarms}{TotalALBlocksNumber},
\]

\[
P_{Miss|Target} = \frac{Misses}{TotalNONALBlocksNumber},
\]

where \(Misses\) and \(FalseAlarms\) are defined similarly to the \(FB-ACC\) metric; and \(TotalALBlocksNumber\) and \(TotalNONALBlocksNumber\) are the total number of reference blocks labelled, respectively, as alarm (AL) and non-alarm (NON-AL).

4. Results

The experiments were carried out with the part of the recorded database that was labelled. The total amount of data used is around 36.5 minutes and 22.29% of this time is labelled as alarms. In total there were 31 files from different recording sessions. Note, only recordings made with the microphone placed outside the incubator were used to keep homogeneous experimental conditions.

As the dataset is small, in order to obtain more statistically relevant results, a 10-fold cross validation scheme was applied: each time, 9 sessions of data were used for training and 1 session for testing.

Table 1 shows the baseline result obtained with the GMM-based system, and demonstrates the importance of both pre- and post-processing steps for the detection accuracy. Namely, the signal enhancement (denoising) itself yields 32.62% relative improvement in terms of \(FB-ACC\) metric and 14.36% in terms of \(BB-ACC\) metric. Meanwhile, smoothing the classifier output provides only 8.02% improvement in terms of \(FB-ACC\) metric, but leads to 54.08% relative improvement in terms of \(BB-ACC\). Finally, the combination of both techniques results in 46.39% and 72.35% relative improvement in terms of \(FB-ACC\) and \(BB-ACC\), respectively. The fact that the accuracies are not high matches actual observations made in the feature space, where the two classes are rather overlapped.

<table>
<thead>
<tr>
<th>Experiment</th>
<th>FB-ACC, %</th>
<th>BB-ACC, %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>53.12</td>
<td>38.15</td>
</tr>
<tr>
<td>Baseline + denoising</td>
<td>70.45</td>
<td>43.66</td>
</tr>
<tr>
<td>Baseline + smoothing</td>
<td>57.38</td>
<td>58.74</td>
</tr>
<tr>
<td>Baseline + denoising + smoothing</td>
<td>77.76</td>
<td>65.75</td>
</tr>
</tbody>
</table>

Table 2 shows the recognition performance of the GMM-based and DBN-based systems. For simplification purposes, the DBN-based system was tuned on the first iteration of the 10-\(\alpha\) cross-validation (where the first recording session was used as the testing data, while all the other sessions were used as the training data), and the system parameters chosen on that first iteration were further used for all the other 9 iterations. Contrarily, for GMM the parameters were tuned on all the 10 sessions. Despite that difference, it can be seen in the table that the DBN-based system outperforms the GMM-based system, especially in terms of the block-based metric.

For both systems the majority of the errors are misses which agrees with the application-specific error weights chosen.

<table>
<thead>
<tr>
<th>System</th>
<th>FB-ACC, %</th>
<th>BB-ACC, %</th>
</tr>
</thead>
<tbody>
<tr>
<td>GMM-based</td>
<td>71.76</td>
<td>65.75</td>
</tr>
<tr>
<td>DBN-based</td>
<td>79.45</td>
<td>71.51</td>
</tr>
</tbody>
</table>

5. Conclusions

In the reported work, the first results obtained from the exploration of the acoustic environment of a preterm in a NICU are presented. First of all, the initial acoustic analysis carried out from a first set of recordings shows the strong audio acoustic diversity of the NICU. Also, after defining a meaningful block-based metric, the alarm detection experimental results show the usefulness of a signal enhancement pre-processing, and the potential of the DBN-based detection system for that task.

Apart from collecting a larger training database to achieve higher detection accuracies, future work will be oriented towards fine-tuning of the DBN system (e.g. including longer context) and to design task-specific features. We also plan to perform a more thorough acoustic analysis of the acoustic environment, and furthermore switch to the multi-class specific-alarm detection problem.

6. Acknowledgements

The authors are grateful to Vanessa Sancho Torrents, Cristina Borras Novell for their help with database production, and to Eva Bargallo for her contributions at the beginning of this work. This work has been supported by the Spanish project SARAI (TEC2010-21040-C02-01).
7. References


