Biomedical Signal Processing and Control 39 (2018) 390–395

Contents lists available at ScienceDirect

Biomedical Signal Processing and Control

journal homepage: www.elsevier.com/locate/bspc

Pre-processing techniques for improved detection of vocalization sounds in a neonatal intensive care unit

Ganna Raboshchuk[a,*], Climent Nadeu[a], Sergio Vidiella Pinto[a], Oriol Ros Fornells[a], Blanca Muñoz Mahamud[b], Ana Riverola de Veciana[b]

[a] *TALP Research Center, Department of Signal Theory and Communications, Universitat Politècnica de Catalunya, Barcelona 08034, Spain*
[b] *Neonatology, Hospital Sant Joan de Déu, Esplugues de Llobregat, Barcelona 08950, Spain*

ARTICLE INFO

Article history:
Received 6 September 2016
Received in revised form 26 June 2017
Accepted 20 July 2017
Available online 30 August 2017

Keywords:
Neonatal intensive care unit
Vocalization detection
Noise reduction
Spectral subtraction
Non-negative matrix factorization

ABSTRACT

The sounds occurring in the noisy acoustical environment of a Neonatal Intensive Care Unit (NICU) are thought to affect the growth and neurodevelopment of preterm infants. Automatic sound detection in a NICU is a novel and challenging problem, and it is an essential step in the investigation of how preterm infants react to auditory stimuli of the NICU environment. In this paper, we present our work on an automatic system for detection of vocalization sounds, which are extensively present in NICUs. The proposed system reduces the presence of irrelevant sounds prior to detection. Several pre-processing techniques are compared, which are based on either spectral subtraction or non-negative matrix factorization, or a combination of both. The vocalization sounds are detected from the enhanced audio signal using either generative or discriminative classification models. An audio database acquired in a real-world NICU environment is used to assess the performance of the detection system in terms of frame-level missing and false alarm rates. The inclusion of the enhancement pre-processing step leads to up to 17.54% relative improvement over the baseline.

1. Introduction

Most premature infants receive specialized medical care in Neonatal Intensive Care Units (NICUs) during the first several weeks or even months of life, which is crucial for their survival. A typical NICU environment is acoustically very rich, with diverse sounds produced both by human activities and by multiple biomedical equipment [1,2] contributing to high sound levels [3]. It has been recognized that such a noisy NICU environment may compromise normal growth and neurodevelopment of preterm infants [4–8] as the immature brain may not be able to adapt and respond normally to loud, randomly produced sounds of variable intensity taking place in a NICU [9].

The effects of a NICU acoustic environment on a preterm infant could be revealed by the infant's reactions to auditory stimuli from it, which can be investigated by relating the presence of particular sounds (i.e., sound identities and their situation in time) with the preterm physiological variables. Note that in such investigation the sounds are not produced artificially, but occur naturally in the NICU environment and are the ones actually perceived by the preterm infant. A study of this kind can complement greatly the work already reported in the literature, in which only the sound pressure level is considered without taking into account the spectro-temporal properties and identity of sounds (e.g., in [10]).

To carry out a statistical correlation study that uses the sound identities, large amounts of labelled audio data are required, which can only be obtained through automatic detection from audio signals. In this paper, we address the detection of vocalizations, which encompass all sounds produced through a vocal tract, either by infant or adult (i.e., speech, cries, laugher, cough, etc.). These sounds are those most frequently occurring in a NICU environment and that may affect a preterm baby [11,12]. For instance, newborns demonstrate a clear preference for the maternal voice [4], which can have a calming effect, while shouts or cries may affect the newborn in a negative manner. The aim of the detection system developed is to automatically label temporal regions within the input audio where a vocalization sound is present, i.e., to specify the start and end time of each vocalization occurrence without specifying its particular type.

The acoustic analysis of the audio data collected in the NICU shows that speech (i.e. foreground and background voices) is the predominant type of vocalization in that environment. A multitude

[*] Corresponding author.
E-mail address: ganna.raboshchuk@upc.edu (G. Raboshchuk).

http://dx.doi.org/10.1016/j.bspc.2017.07.024
1746-8094/© 2017 Elsevier Ltd. All rights reserved.

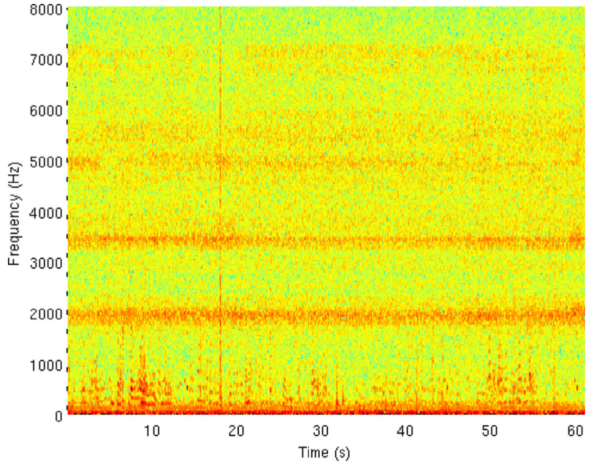

Fig. 1. Spectrogram of an audio sample with the typical ventilation noise.

of studies dealing with the related task of voice/speech activity detection have been reported in the literature, e.g., [13–15] to cite a few. However, there are factors specific to this task in the NICU acoustic environment. Due to the rich multisource nature of that environment, various sound events usually take place simultaneously. Considering vocalizations, the temporal overlaps with other sounds are even more probable due to their extensive presence. Moreover, a specific type of noise produced by the ventilation equipment that supports breathing in neonates which spreads over a wide frequency range is strongly present in the recordings. A spectrogram of one of the typical samples of the ventilation noise is given in Fig. 1. 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 from a recording session an appropriate type of ventilation is used, and this fact introduces a great deal of variability to the data. As the performance of the detection systems is known to deteriorate significantly in the presence of background noise or temporal overlaps between sound sources [16,17], the NICU environment makes vocalization detection quite challenging.

In this paper, in order to obtain a more robust detection, we address the above-mentioned factors by including a pre-processing step that is based on the following techniques:

(1) spectral subtraction, to attenuate the stationary ventilation noise;
(2) non-negative matrix factorization, which is more suitable for audio enhancement in case of non-stationary noises, to segregate vocalizations from the other interfering sounds and noise.

This study compares the performance of the detection system when different pre-processing schemes based on the techniques and their combinations are applied prior to detection, and selects the scheme yielding the best detection results. The usefulness of the pre-processing step is evaluated when either a generative or a discriminative classification approach is used. To our knowledge, this is the first work where the employed enhancement techniques are applied in the context of a NICU acoustic environment.

The rest of the paper is organised as follows. Section 2 provides details on how the pre-processing step of the detection system is implemented and briefly describes the enhancement techniques used, and Section 3 contains description of the detection system itself. The evaluation setup and experimental results are presented in Sections 4 and 5, respectively.

2. Enhancement techniques

2.1. Spectral subtraction

Spectral subtraction (SS) algorithm is the classical tool used for audio denoising where an additive model is assumed, i.e. the noise-corrupted input signal $y(n)$ is composed of the clean signal $x(n)$ and the additive noise signal $d(n)$; that is $y(n) = x(n) + d(n)$. Then, the clean signal spectrum $\hat{X}(n, k)$ can be estimated by subtracting an estimate of the noise spectrum $\hat{D}(n, k)$ from the noisy signal spectrum $\hat{Y}(n, k)$ as follows [18]:

$$|\hat{X}(n,k)|^{\gamma} = \begin{cases} |\hat{Y}(n,k)|^{\gamma} - \alpha|\hat{D}(n,k)|^{\gamma}, \\ \quad \text{if } |\hat{Y}(n,k)|^{\gamma} > (\alpha+\beta)|\hat{D}(n,k)|^{\gamma} \\ \beta|\hat{D}(n,k)|^{\gamma}, \ \text{otherwise} \end{cases} \tag{1}$$

where n and k are, correspondingly, the frame and the frequency bin index, $\gamma = 1$ yields magnitude and $\gamma = 2$ yields power spectrum subtraction, α is the subtraction factor, which controls the amount of noise to be subtracted, and $0 < \beta \ll 1$ is the spectral floor parameter, which controls the amount of residual and perceived musical noise. This approach is referred to as SS using oversubtraction (because usually $\alpha \geq 1$) [19].

The use of a proper noise estimate $\hat{D}(n, k)$ is crucial for the quality of the enhanced signal. Often, it is obtained once from the first frames of the input audio. But since the annotation data are not available, it is not guaranteed that there are no vocalization sounds present in that beginning segment. On the other hand, since ventilation noise is stationary and is present throughout the recording, we propose using the average spectrum of the whole input signal as noise estimate.

Alternatively, the noise estimate can be obtained and updated throughout the input signal, taking into account the probability of the presence of speech. Such an approach is able to better deal with highly non-stationary noise environments. In this work, we employ the Minima-Controlled Recursive-Averaging (MCRA) algorithm [19], so the mean-square estimate of the noise power spectrum is obtained recursively as follows:

$$\begin{aligned} |\hat{D}(n,k)|^{\gamma} &= \alpha_d(n,k)|\hat{D}(n-1,k)|^{\gamma} \\ &\quad + (1-\alpha_d(n,k))|\hat{Y}(n,k)|^{\gamma}, \end{aligned} \tag{2}$$

where $\alpha_d(n, k)$ is a smoothing factor defined as

$$\alpha_d(n,k) = \alpha + (1-\alpha)p(n,k). \tag{3}$$

Here, $p(n, k)$ is the speech-presence probability which is calculated using the ratio of the smoothed (with a smoothing factor α_s) noisy signal spectrum to its local minimum. This ratio is compared to a threshold δ yielding a binary speech-presence probability estimate, which is further smoothed over time with a smoothing factor α_p.

In this study, the following parameter setup is used: the processing is performed on Hann-windowed half-overlapped 64 ms frames with $\gamma = 2$. For standard SS, $\alpha = 0.01$ $\{0\ldots3\}$,[1] $\beta = 0$ $\{0\ldots1\}$ and the noise estimate is obtained from the first 7 frames of the audio recording (which roughly corresponds to 200 ms); for SS with the average spectrum noise estimate, $\alpha = 0.2$ $\{0\ldots1\}$ and $\beta = 0$ $\{0\ldots1\}$; for SS with MCRA, α, β, α_d, α_s, α_p are equal to, correspondingly, 1 $\{0\ldots1\}$, 0.01 $\{0\ldots0.1\}$, 0.2 $\{0.2\ldots0.95\}$, 0.9 $\{0.7\ldots0.95\}$ and 0.1 $\{0.01\ldots0.7\}$.

[1] The range of values on which each parameter was optimized using grid search is shown in curly brackets. Note that the parameter tuning was not exhaustive and there may be more optimal parameter configurations, but, as observed during tuning, no large improvement should be expected and the general relation between the technique performance will hold.