DOI 10.1016/j.bspc.2017.07.024

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*

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.
Keywords: neonatal intensive care unit, vocalization detection, noise reduction, spectral subtraction, non-negative matrix factorization.

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, 5, 6, 7, 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 ef-

**Corresponding author:
e-mail: ganna.raboshchuk@upc.edu (Ganna Raboshchuk)

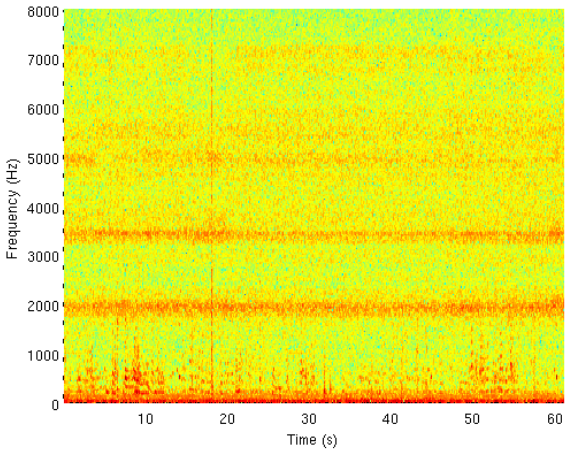

Fig. 1. Spectrogram of an audio sample with the typical ventilation noise.

fect, 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 of studies dealing with the related task of voice/speech activity detection have been reported in the literature, e.g., [13, 14, 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 Figure 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 Section 4 and Section 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]:

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

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:

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

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

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

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...3}[1], $\beta = 0$ {0...1} 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...1} and $\beta = 0$ {0...1}; for SS with MCRA, $\alpha, \beta, \alpha_d, \alpha_s, \alpha_p$ are equal to, correspondingly, 1 {0...1}, 0.01 {0...0.1}, 0.2 {0.2...0.95}, 0.9 {0.7...0.95} and 0.1 {0.01...0.7}.

2.2. Non-negative matrix factorization

Non-negative matrix factorization (NMF) was first presented in its basic form in [20] and since then it has proven to be useful in various pattern recognition areas, such as automatic speech recognition [21] and acoustic event detection [22]. Basically, NMF is a linear decomposition technique that attempts to approximate an input non-negative matrix V as a product of two non-negative matrices, i.e.

$$V_{F\times N} \approx W_{F\times R} \cdot H_{R\times N}, \quad (4)$$

where $R \leq F$ controls the rank of the approximation. In audio signal processing, NMF is usually applied to the spectrogram of the signal [23], and F and N correspond to the number of frequency bins and number of frames, respectively. The columns of W are usually referred to as bases, and the rows of H as their corresponding weights or activations in time.

The optimization problem of minimizing the divergence between the input matrix and its approximation needs to be solved:

$$\arg\min_{W,H} D(V||WH) + \lambda|H|_1 \qquad W,H \geq 0 \quad (5)$$

where D is a cost function (in this work, the Kullback-Leibler divergence), and the parameter $\lambda \geq 0$ is used to impose a sparsity constraint on the activations, thus favouring solutions with fewer bases activated at a given time. The minimization is achieved by updating W and H with multiplicative factors (derived using the gradient descent algorithm) until convergence [24].

A supervised NMF approach is used where the bases matrix W_t is trained beforehand on the training data. In the general case, when S sound sources are considered, a bases matrix is trained for each source separately and a global bases matrix is constructed via concatenation $W_t = [W_1; ...; W_S]$. At the source separation step the bases matrix is fixed and only the activations matrix is estimated $H = [H_1; ...; H_S]$, i.e. the optimization problem is:

$$\arg\min_{H} D(V||W_tH) + \lambda|H|_1 \qquad H \geq 0 \quad (6)$$

In the basic case, the spectrum of each source can be obtained by multiplication of the source bases by the corresponding activations, i.e.

$$\hat{V}_i = W_iH_i, \qquad i \in [1..S]. \quad (7)$$

Commonly, an approach similar to Wiener filtering is applied to reconstruct each source:

$$\hat{V}_i = \frac{W_iH_i}{\sum_i W_iH_i} \otimes V, \quad (8)$$

where multiplication $\otimes$ and division operations are element-wise [23].

The considered binary vocalization detection is, in principle, a two-source problem with the two sources corresponding to vocalization and non-vocalization classes. The global bases matrix $W_{train} = [W_V; W_{NV}]$ consists of the bases trained for each class, respectively. The enhanced audio signal is then reconstructed using only the vocalization spectra $\hat{V}_V$ and the phase of the original input audio.

In our work, the implementation of NMF described in [24] is used, with the following parameter setup: the input matrix V is a magnitude spectrogram computed on Hann-windowed frames of 32 ms length with 16 ms shift. We train $R = 25$ {25...100} bases per class, where each base corresponds to a vector of dimension $F \times 1$. The sparsity parameter λ is set to 0.01 {0...2}. At training and testing time we use up to 25 iterations.

2.3. Combined approach

We want to exploit the complementarity that may exist between the SS and NMF algorithms, by investigating several combinations of the techniques.

Firstly, we try the combination in which SS and NMF are applied consecutively. In this case, the audio data are previously processed by SS in order to attenuate the ventilation noise, and then this enhanced audio is used as training data for NMF. Alternatively, NMF is applied prior to SS processing.

Secondly, we employ NMF to obtain the noise spectrum estimate $|\hat{D}(n,k)|^\gamma$ for SS technique. Contrary to NMF based preprocessing, where the spectrum is reconstructed for vocalizations, here we obtain the reconstructed spectrum $\hat{V}_{NV}$ for non-vocalizations or, in other words, the irrelevant sounds. Each column n of the reconstructed non-vocalization spectral matrix $\hat{V}_{NV}$, which corresponds to the time frame n, is assigned to the vector $|\hat{D}(n,k)|^\gamma$ in (1). The advantage of this approach is that the noise estimate is assumed to be more accurate.

[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.

3. Detection system

The input signal is split into frames using a Hamming window with the frame length of 30 ms and the frame shift of 10 ms. 16 Frequency-Filtered Logarithmic FilterBank Energy (FF-LFBE) features [25] along with their 16 first temporal derivatives were extracted from each frame. Therefore, the dimension of the feature vector is 32.

A Gaussian Mixture Model (GMM) based detector was used, consisting of a model for vocalization and a model for non-vocalization. Each model is a single Gaussian probability density function with diagonal covariance matrix as, in our experiments, this provided better detection performance than using more mixture components. With the likelihoods obtained from the two models, each frame is classified either as vocalization or non-vocalization. The decision threshold is chosen based on the Equal Error Rate (EER) criterion, assuming that both types of errors are equally important at the frame level.

In addition to GMMs, which is a generative classification approach, we also perform experiments employing a discriminative Support Vector Machines (SVM) based classifier. SVMs aim at maximizing the margin between the classes and have the advantage of using only the training samples that are closest to the decision surface, which can be beneficial when a limited amount of training data is available [26]. In this work, both linear and Radial Basis Function (RBF) kernels are employed. Before being fed to the classifier, the input features are mean-variance normalized; the mean and variance values calculated on the training data are also applied to the testing data.

Optionally, smoothing (via majority voting) is applied to the string of output labels. The length of the smoothing window was optimized with regards to the detection performance and is equal to 31 frames.

4. Evaluation setup

The presented experimental evaluations were performed using an audio database acquired in the NICU of Hospital Sant Joan de Déu Barcelona. The database contains ten recording sessions carried out both in the morning and in the afternoon (later recordings were not possible). The recordings were made in a NICU room designated for intensive care of very preterm newborns, which is equipped with four incubators. The number of sessions recorded in each incubator site was roughly the same. Each recording session was made with a different newborn, which allowed us to capture the variability due to the equipment used (including ventilation equipment). As the amount of activities that take place in the NICU can be very large, a set of acoustic scenarios, which mostly correspond to the daily nursery care related activities (e.g. changing a diaper, measuring temperature), was selected for recording. A given session contains a subset of those selected scenario recordings, each around 1-2 min long.

Two electret unidirectional microphones connected to a 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, approximately 50 cm above it, usually pointing to the centre of the room. More information about the database acquisition, as well as a general acoustic description of the NICU can be found in [27].

The experiments were carried out with the part of the recorded database that was annotated and has a total duration of 40.2 min. The vocalization sounds are present 56.7% of this time. Only the recordings acquired with the microphone outside the incubator were used to keep homogeneous experimental conditions, and also because this microphone is closer to the vocalization sources. The original 44.1 kHz recordings were downsampled to 16 kHz.

A 10-fold cross-validation scheme was applied in order to obtain more statistically relevant results, where on each fold 9 recording sessions were used for training and 1 session for testing. The overall metric scores were obtained by aggregating the results over all 10 folds. Note that a cross-validation scheme was also applied for NMF processing, where 9 sessions were used for bases training, which were applied to perform separation over 1 testing session.

For every pre-processing scheme, the classification models were re-trained on the data obtained after that pre-processing.

The detection performance was evaluated at the frame level. The Missing Rate (MR) and the False Alarm Rate (FAR) metrics were used, which are defined as:

$$MR = \frac{N_M}{N_V}, \qquad FAR = \frac{N_{FA}}{N_{NV}}, \tag{9}$$

where N_M and N_{FA} are the number of misclassified frames for vocalization and non-vocalization class, respectively, and N_V and N_{NV} are the total number of vocalization and non-vocalization frames, respectively.

Table 1. Vocalization detection performance obtained by the baseline system. In bold are the best scores for each column.

Number of Gaussians	No post-processing	Smoothing	
	Evaluation metrics (%)		
	MR = FAR	MR	FAR
1	**32.90**	**29.64**	**29.68**
2	35.40	31.19	31.23
4	36.55	32.17	32.49
8	39.33	34.15	37.23

5. Experimental results

The baseline GMM-based system performance is presented in Table 1 as a function of the number of Gaussian components used. The EER, which corresponds to both MR and FAR metrics having the same value, is reported when no post-processing (i.e. smoothing) is applied. It can be seen that the increase in the number of Gaussians seems to be detrimental to detection performance; therefore only one Gaussian was used in subsequent experiments. Furthermore, in all cases smoothing the classifier output improved the detection results in terms of both metrics, yielding up to 12% relative improvement in the best case. Taking into account the temporal context using a number of adjacent frames, smoothing discards sporadic miss and false alarm

Table 2. Vocalization detection performance obtained by the GMM-based system with a pre-processing step. In bold are the best scores for each column.

Pre-processing	No post-processing	Smoothing	
	Evaluation metrics (%)		
	MR = FAR	MR	FAR
None	32.90	29.64	29.68
SS standard	33.92	30.75	31.44
SS average	31.20	27.46	27.82
SS + MCRA	29.39	26.34	25.65
NMF basic	33.56	29.99	29.97
NMF Wiener	32.12	29.11	28.26
SS → NMF	31.99	27.44	27.96
NMF → SS	**28.31**	**24.44**	**25.26**
SS + NMF	33.80	30.54	30.19

errors. The system performance will be further compared to the smoothed baseline.

Table 2 shows the detection performance of the system when different pre-processing schemes are applied prior to detection. Several of the proposed schemes are able to improve the baseline results.

First of all, the results for the SS and NMF techniques applied separately are presented. It can be seen that applying the standard SS leads to a performance loss (by 3.74% and 5.93%, relatively, in terms of MR and FAR, respectively). This may be explained by the fact that some of the recordings contain vocalization sounds at the beginning and the obtained noise estimate is not accurate, which may have caused distortion of vocalizations. This explanation is also justified by the optimal parameter values obtained ($\alpha = 0.01, \beta = 0$) which basically corresponds to doing almost no subtraction.

On the other hand, SS using the average spectrum noise estimate (*SS average*) and SS with the MCRA algorithm for the noise estimation (*SS + MCRA*) are both able to improve the baseline result due to the better noise estimate obtained. In the case of *SS average* the relative improvement is 7.35% and 6.27% in terms of MR and FAR, respectively, showing that the average noise estimate is able to represent the ventilation noise. It is also reflected in the higher optimal value of $\alpha = 0.2$. And as *SS + MCRA* pre-processing results in a more accurate noise estimate, it yields even higher relative improvement: 11.13% in terms of MR and 13.58% in terms of FAR metric scores. While the noise estimate in *SS average* mainly captures ventilation noise, the noise estimate in *SS + MCRA*, due to its continuous updating, is able also to capture some non-stationary non-vocalization sounds, and hence *SS + MCRA* provides better results.

As for NMF-based pre-processing the gain is not so obvious. Employing the basic technique for vocalizations reconstruction (*NMF basic*) doesn't bring any improvement to the baseline result; conversely, a relative loss of 1.18% in terms of MR and of 0.98% in terms of FAR is obtained. On the other hand, NMF with Wiener-like reconstruction (*NMF Wiener*), which provides better sound quality [28], improves the results, but to a small

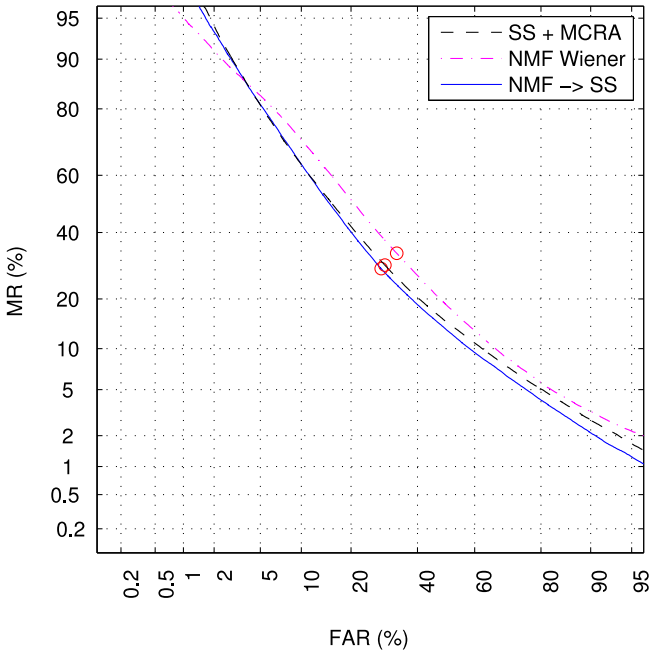

Fig. 2. The DET graphs for the three best performing setups. Circles correspond to EER points.

extent: by 1.79% and 4.78%, relatively, in terms of MR and FAR, respectively. The reason for NMF not performing well may be the fact that strong ventilation noise and other sounds are present in the training data of both vocalizations and non-vocalizations, thus reducing the discriminative power of the trained bases.

As mentioned above, the two techniques target different types of noises: SS is more suitable for reduction of stationary noises, while NMF may deal with non-stationary noises. The fact that SS showed better results on our data correlates well with the observation that the stationary ventilation noise is vastly present in all the recordings and, therefore, influences the detection performance more strongly than do other interfering sounds.

The last part of the table contains the detection results for different technique combinations: when SS and NMF are applied consecutively to the audio signal (*SS → NMF* and *NMF → SS*) and, also, when NMF is used to obtain the noise estimate for SS (*SS + NMF*). Note that the best setups of SS and NMF techniques are used for the audio-based combinations, namely, *SS + MCRA* and *NMF Wiener*. For *SS + NMF* pre-processing, *NMF Wiener* is used for the noise estimation, the parameters of SS are set to $\alpha = 0.2, \beta = 0$ and the frame length is set to 64 ms. In the rest of the cases the optimal parameter setups obtained for each technique separately are kept.

The best detection results are obtained when SS is applied to the audio signal pre-processed with NMF (*NMF → SS*); in this case the relative improvement achieved is 17.54% in terms of MR and 14.89% in terms of FAR. This confirms the complementarity of the techniques in terms of the types of sounds they are targeting. The detection results for the alternative pre-processing sequence (*SS → NMF*) are worse than using SS

Table 3. Vocalization detection performance obtained by the SVM-based system. In bold are the best scores for each column.

Pre-processing, kernel	No post-processing		Smoothing	
	Evaluation metrics (%)			
	MR	FAR	MR	FAR
None, linear	30.65	37.60	27.83	36.59
None, RBF	30.67	37.24	27.93	36.08
NMF → SS, RBF	**25.09**	**35.2**	**22.07**	**33.94**

alone (only 7.42% and 5.80% relative improvement in terms of MR and FAR, respectively, compared to the baseline results). This may be due to the fact that SS processing introduces a musical noise to the output audio which, as occurs with ventilation noise, is not beneficial for bases training. It can also be seen that the *SS + NMF* combination does not outperform the baseline setup. This may be attributed in part to the fact that the processing window length used in SS is not optimal for NMF.

The Detection Error Tradeoff (DET) graphs are presented in Figure 2 for the best performing setups of SS, NMF, and their combination when no post-processing is applied. It can be seen that the combination of the two techniques outperforms each one of them individually at all the operational points of the curve except for the ones where FAR is very low.

In Table 3 we provide the detection results for the SVM-based classification, with both linear and RBF kernels. Either no pre-processing or the $NMF \rightarrow SS$ pre-processing, which gave the best results for GMM-based classifier, is applied. For linear SVM, the parameter C, which controls the trade-off between the training error and the margin, is set to 1e−4 {1e−5...1}. For SVM with RBF kernel, this parameter is equal to $C = 0.05$ {1e−4...1}, and the parameter γ of RBF is set to 0.001 {0.0001...0.25}.

It can be seen that there is no significant difference in the detection results for the two types of SVM kernel functions on our data, and the RBF kernel only slightly outperforms the linear one. I.e., with smoothing post-processing, the total error (MR + FAR) for linear kernel is equal to 64.42%, while for RBF kernel it is equal to 64.01%. Similarly to the GMM-based system, these results are improved when the pre-processing step is added, although the overall improvement is somewhat smaller in this case. In particular, the relative improvement in terms of MR and FAR is equal to 20.98% and 5.93%, respectively.

Comparing the results for the two types of classification models, a generative GMM and a discriminative SVM, it can be seen that SVM-based system does not outperform the GMM-based one. The total error for the GMM-based and SVM-based systems is equal to 49.7% and 56.01%, respectively. Perhaps, this is due to an observed strong overlap of the vocalization and non-vocalization classes in the feature space.

6. Conclusions

This paper presents our work on vocalization sound detection for a new and challenging application: automatic analysis of the acoustic environment of a preterm infant in a NICU. The work focused on the pre-processing step of non-vocalization sounds reduction. It has been shown that the detection system benefits from introducing the enhancement step, though the obtained detection error is still relatively high due to the complexity of the detection problem in a real-world NICU environment and the scarcity of data.

Binary detection of vocalizations is the first step towards a correlation study and shall be followed by detection of each type of relevant vocalization sounds. Future work could entail detecting higher intensity vocalizations (i.e., foreground speech and shouts) or parental voices, as these sounds are supposed to affect a preterm baby the most. Improvements could also be made to other steps of the detection system, e.g. other feature extraction schemes might allow better discrimination between the vocalization and non-vocalization classes.

Acknowledgments

This study was supported in part by the Spanish government (TEC2012-38939-C03-02, TEC2015-69266-P) as well as by the European Regional Development Fund (ERDF/ FEDER). The authors are grateful to Vanessa Sancho Torrents and Francisco Alarcón Sanz for their work on the database annotation.

References

[1] S. M. A. Hassanein, N. M. El Raggal, A. A. Shalaby, Neonatal nursery noise: practice-based learning and improvement, Journal of Maternal-Fetal and Neonatal Medicine 26 (4) (2013) 392–395.

[2] M. D. Livera, B. Priya, A. Ramesh, P. N. Suman Rao, V. Srilakshmi, M. Nagapoornima, A. G. Ramakrishnan, M. Dominic, Swarnarekha, Spectral analysis of noise in the neonatal intensive care unit, Indian journal of pediatrics 75 (3) (2008) 217–222.

[3] C. Krueger, S. Wall, L. Parker, R. Nealis, Elevated sound levels within a busy NICU, Neonatal network 24 (6) (2005) 33–37.

[4] L. Gray, M. K. Philbin, Effects of the neonatal intensive care unit on auditory attention and distraction, Clinics in perinatology 31 (2) (2004) 243–260, vi.

[5] A. Salavitabar, K. K. Haidet, C. S. Adkins, E. J. Susman, C. Palmer, H. Storm, Preterm infants' sympathetic arousal and associated behavioral responses to sound stimuli in the neonatal intensive care unit, Advances in neonatal care 10 (3) (2010) 158–166.

[6] E. M. Wachman, A. Lahav, The effects of noise on preterm infants in the NICU, Archives of Disease in Childhood - Fetal and Neonatal Edition 96 (4) (2010) F305–F309.

[7] Committee on Environmental Health, Noise: a hazard for the fetus and newborn, Pediatrics 100 (4) (1997) 724–727.

[8] S. N. Graven, J. V. Browne, Sleep and brain development: the critical role of sleep in fetal and early neonatal brain development, Newborn and Infant Nursing Reviews 8 (4) (2008) 173–179.

[9] K. A. Thomas, A. Uran, How the NICU environment sounds to a preterm infant: update, The American journal of maternal child nursing 32 (4) (2007) 250–253.

[10] P. Kuhn, C. Zores, T. Pebayle, A. Hoeft, C. Langlet, B. Escande, D. Astruc, A. Dufour, Infants born very preterm react to variations of the acoustic environment in their incubator from a minimum signal-to-noise ratio threshold of 5 to 10 dBA, Pediatric research 71 (4) (2012) 386–392.

[11] E. McMahon, P. Wintermark, A. Lahav, Auditory brain development in premature infants: the importance of early experience, Annals of the New York Academy of Sciences 1252 (2012) 17–24.

[12] M. Caskey, B. Stephens, R. Tucker, B. Vohr, Importance of parent talk on the development of preterm infant vocalizations, Pediatrics 128 (5) (2011) 910–916.

[13] S. Morita, M. Unoki, X. Lu, M. Akagi, Robust voice activity detection based on concept of modulation transfer function in noisy reverberant

environments, Journal of Signal Processing Systems 82 (2) (2015) 163–173.

[14] S. Graf, T. Herbig, M. Buck, G. Schmidt, Features for voice activity detection: a comparative analysis, EURASIP Journal on Advances in Signal Processing 2015 (1) (2015) 1–15.

[15] I. Hwang, H.-M. Park, J.-H. Chang, Ensemble of deep neural networks using acoustic environment classification for statistical model-based voice activity detection, Computer Speech & Language 38 (2016) 1–12.

[16] A. Temko, C. Nadeu, Acoustic event detection in meeting-room environments, Pattern Recognition Letters 30 (14) (2009) 1281–1288.

[17] W. Wang (Ed.), Machine Audition: Principles, Algorithms and Systems, 1st Edition, IGI Global, 2010.

[18] M. Berouti, R. Schwartz, J. Makhoul, Enhancement of speech corrupted by acoustic noise, in: Proc of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Vol. 4, 1979, pp. 208–211.

[19] P. C. Loizou, Speech Enhancement: Theory and Practice, 2nd Edition, CRC Press, 2013.

[20] D. D. Lee, H. S. Seung, Learning the parts of objects by non-negative matrix factorization, Nature 401 (6755) (1999) 788–791.

[21] J. T. Geiger, J. F. Gemmeke, B. Schuller, G. Rigoll, Investigating NMF speech enhancement for neural network based acoustic models, in: Proc of Interspeech, 2014, pp. 2405–2409.

[22] A. Mesaros, T. Heittola, O. Dikmen, T. Virtanen, Sound event detection in real life recordings using coupled matrix factorization of spectral representations and class activity annotations, in: Proc of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2015, pp. 151–155.

[23] F. Weninger, J. Le Roux, J. R. Hershey, S. Watanabe, Discriminative NMF and its application to single-channel source separation, in: Proc of Interspeech, 2014, pp. 865–869.

[24] P. O'Grady, B. Pearlmutter, Convolutive non-negative matrix factorisation with a sparseness constraint, in: Proc of the 16th IEEE Signal Processing Society Workshop on Machine Learning for Signal Processing, 2006, pp. 427–432.

[25] C. Nadeu, D. Macho, J. Hernando, Time and frequency filtering of filter-bank energies for robust HMM speech recognition, Speech Communication 34 (12) (2001) 93–114.

[26] R. O. Duda, P. E. Hart, D. G. Stork, Pattern Classification, 2nd Edition, Wiley, 2001.

[27] G. Raboshchuk, C. Nadeu, B. Muñoz Mahamud, A. Riverola de Veciana, S. Navarro Hervas, On the acoustic environment of a neonatal intensive care unit: initial description, and detection of equipment alarms, in: Proc of Interspeech, 2014, pp. 2543–2547.

[28] P. Smaragdis, Convolutive speech bases and their application to supervised speech separation 15 (1) (2007) 1–12.