ROBUST NON-NEGATIVE BLOCK SPARSE CODING FOR ACOUSTIC NOVELTY DETECTION

Ritwik Giri, Arvindh Krishnaswamy, and Karim Helwani

Amazon Web Services, Inc., Palo Alto, CA

{ritwikg, arvindhk, helwk}@amazon.com

ABSTRACT

In this paper we address the problem of detecting previously unseen novel audio events in the presence of real-life acoustic backgrounds. Specifically, during training, we learn subspaces corresponding to each acoustic background, and during testing the audio frame in question is decomposed into a component that lies on the mixture of subspaces and a supergaussian outlier component. Based on the energy in the estimated outlier component a decision is made, whether or not the current frame is an acoustic novelty. We compare our proposed method with state of the art auto-encoder based approaches and also with a traditional supervised Nonnegative Matrix Factorization (NMF) based method using a publicly available dataset - A3Novelty. We also present results using our own dataset created by mixing novel/rare sounds such as gunshots, glass-breaking and sirens, with normal background sounds for various event to background ratios (in dB).

1. INTRODUCTION

1.1. Background

Novel audio event detection has a number of important applications. For example, in surveillance systems, detecting unusual events using audio can nicely complement video based approaches. This is especially true in cases where there is not sufficient illumination, or in the presence of visual occlusions where the performance of video surveillance is impaired. But novelty detection poses interesting challenges as well.

One difficulty is that not all potential audio events can be pre-determined, pre-recorded and labeled. We need to deal with unseen novel sounds and transients. But most of the current literature we have seen on audio surveillance systems [1, 2, 3] propose fully supervised learning methods: along with acoustic background data they also require labeled audio examples corresponding to audio events. We have also seen applications of this supervised audio event detection in consumer products, for example Alexa Guard [1]: this feature enables Echo devices to detect specific sounds that the user selects such as smoke alarm, glass breaking sound, carbon monoxide alarms etc. But these fully-supervised systems can not detect unseen novel audio events. Therefore, researchers are working on unsupervised techniques as well to detect novel audio events [4, 5, 6, 7, 8].

1.2. Related Work

Even though novelty detection is a relatively new problem in the audio signal processing community, this topic has been well-researched in other data modalities and fields such as medical diagnosis [9, 10], damage inspection [11, 12] electronic IT security [13], and video surveillance systems [14]. In [15, 16] authors grouped several novelty detection techniques in two major categories - statistical approaches and neural network based approaches.

Statistical approaches depend on properties of the normal background audio data, and, during training, either fit a model or a probability distribution function over the data. During testing they exploit this pre-trained model to determine if a test sample belongs to the learned distribution or not. These methods have been well researched and have been applied to several novelty detection applications successfully such as in handwriting detection, the recognition of cancer, failure detection in jet engines, and fMRI analysis. In the context of acoustic novelty detection, in [7], the authors have introduced Gaussian Mixture Model (GMM) and Hidden Markov Model (HMM) based methods for detecting audio novelties in realistic acoustic backgrounds such as in 1) smart-home environments, 2) ATM settings, and 3) general-purpose security settings. In [4] authors proposed a one-class Support Vector Machine (OC-SVM) based unsupervised method for real-time detection of novel events in the context of audio surveillance. Recently in [17], the authors proposed a non-negative matrix under-approximation (NMU) method to perform novel-sound detection for unhealthy machineries.

Neural network based approaches, specifically autoencoder based approaches, have recently gained attention for novelty detection in both audio and in other modalities. The main working principle of these approaches lies in training an autoencoder using normal/expected data, and during testing checking if the network is struggling to encode and decode the test data accurately. I.e., if the system produces a high reconstruction error compared to some threshold, it is then considered novel input. In [18], the authors proposed a denoising autoencoder structure using both feedforward units and LSTM units for acoustic novelty detection task and showed significant improvement of performance over both GMM and OC-SVM based methods.

1.3. Contribution

In this article, we propose a robust non-negative block sparse coding based technique to detect novel sounds, such as gunshots, glass-breaking, sirens etc, in different real life acoustic backgrounds, such as in a bus, cafe, beach, city center, metro station and grocery store. During training, we learn subspaces corresponding to each acoustic background using supervised Nonnegative Matrix Factorization (S-NMF). During testing the audio frame in question is decomposed into a component that lies on the mixture of subspaces and a supergaussian outlier component. Because we

[1] https://www.cnet.com/news/alexa-guard-goes-live-lets-your-echo-listen-for-trouble-amazon-home-security/

use a separate estimator for the outlier modeled as a supergaussian random variable, our approach is robust to minor deviations in learned acoustic backgrounds and also actual backgrounds during testing unlike in [17], where no such constraint has been imposed on novel sound estimate. We also create a challenging dataset by mixing novel sounds with real life acoustic scenes for different event to background ratios. Finally, we compare the results of our method, with state-of-the-art methods using a publicly available dataset: A3Novelty. We also compare them using our own dataset and show that our proposed method either matches the performance of the state-of-the-art methods or outperforms them in different cases.

The rest of the article is organized as follows: In Section 2 the proposed method is presented in detail, in Section 3 a brief description of both the datasets, that have been used in this article is given, in Section 4 we present evaluation results of our proposed method and other competing methods over previously mentioned two datasets and finally Section 5 concludes the paper and talks about some future research directions.

2. PROPOSED METHOD

2.1. Training Stage: Learning Background Subspaces

First, an offline training stage is needed to learn the corresponding dictionaries as subspaces, for D different types of acoustic backgrounds such as in a bus, in a cafe, in a city center etc., by solving the following optimization problem D times, for $d = 1...D$.

$$\mathbf{W}_{N_d}, \mathbf{H}_{N_d} = \arg \min_{\mathbf{W}\geq 0, \mathbf{H}\geq 0} \mathrm{KL}(\mathbf{N}_d|\mathbf{WH}) + \mu|\mathbf{H}|_1, \quad (1)$$

where $\mathbf{N}_d \in K$(number of bins) $\times$ L(number of frames) is the magnitude of the d^{th} training background sound STFT representation, and $KL(\cdot|\cdot)$ denotes the KL divergence. This optimization problem can be solved in an iterative manner using multiplicative updates [19]. To avoid scaling indeterminacies, normalization constraint is applied, such that columns of $\mathbf{W}_{N_d}$ have unit norm. After the training stage, the mixture of subspaces corresponding to D different types of acoustic backgrounds can be represented by the concatenated dictionary matrix, $\mathbf{W} = [\mathbf{W}_{N_1}, ..., \mathbf{W}_{N_d}, ..., \mathbf{W}_{N_D}]$.

2.2. Testing Stage: Robust Non-negative Sparse Coding

2.2.1. Basic Model

During testing stage, we will decompose the i^{th} test frame STFT magnitude in the following manner,

$$\mathbf{v_i} = \mathbf{W}\mathbf{h_i} + \mathbf{r_i}, \quad (2)$$

where, $\mathbf{v_i} \in K$(number of bins) $\times$ 1 is the magnitude of the i^{th} frame STFT representation, $\mathbf{r_i} \in K$(number of bins) $\times$ 1 is the outlier term and $\mathbf{h_i} \in M$(number of bases in $\mathbf{W}$) $\times$ 1 is the i^{th} activation vector.

2.2.2. Structured Sparsity in Activation

Since $\mathbf{W}$ is an overcomplete dictionary, a sparseness constraint over the activation vector $\mathbf{h_i}$ is typically employed, leading to a standard sparse coding framework. Since we are operating on STFT magnitude feature space, non-negativity constraint over $\mathbf{h_i}$ will also be employed. During testing, it is reasonable to assume that the acoustic background is not changing drastically, hence instead of solving the above mentioned sparse coding problem for every frame, we will solve for $L = 5$ frames simultaneously. Hence the model will become,

$$\mathbf{V} = \mathbf{WH} + \mathbf{R} \quad (3)$$

where $\mathbf{V}, \mathbf{H}, \mathbf{R}$ are matrices now with $L = 5$ columns/ frames.

For our problem in hand, a more structured sparsity constraint has been identified as useful. Since our acoustic background dictionary $\mathbf{W}$ is essentially a concatenated version of D subdictionaries, we argue that similar block structure can also be expected in each frame of the activation matrix, i.e., $\mathbf{H}_{(:,l)}$. Intuitively, this represents the fact that during testing if a basis vector of a specific subdictionary contributes to represent the testing frame, the other basis vectors of that specific subdictionary will also contribute. Hence in the l^{th} activation vector, a block sparsity structure can be expected. To impose this structural constraint, following regularization term is included in the cost function,

$$\Psi(\mathbf{H}) = \sum_{l=1,..5} \sum_{g_i \in G} \log(||\mathbf{H}_{(g_i,l)}||_1 + \epsilon) \quad (4)$$

where, $G = [g_1, ..., g_D]$ represents the D background subspaces. In literature, this regularization term is also known as the log-ℓ_1 measure [20]. We will also assume that since the acoustic background is not changing within 5 frames the same subdictionary will be used to explain the data over these 5 frames. Hence, structured regularizer over the activation matrix will become,

$$\Psi(\mathbf{H}) = \sum_{g_i \in G} \log(||vec(\mathbf{H}_{(g_i,:)})||_1 + \epsilon). \quad (5)$$

Where, $vec(\mathbf{H}_{(g_i,:)})$ is the vectorized format of submatrix $\mathbf{H}_{(g_i,:)}$. In our work, we keep the hyperparameter $\epsilon = 10^{-4}$ fixed for all our experiments.

2.2.3. Structured Sparsity in Outliers

As discussed above, we employ a supergaussianity/ sparse constraint on the outlier term $\mathbf{R}$, to capture the novel event. This intuition of using a supergaussian random variable to model outliers is motivated from several well known literatures on robust regression [21, 22, 23], which have used heavytailed/supergaussian distributions, such as student's t distribution, to model the outliers in the data.With the supergaussianity assumption, we make sure the that the model mismatch error (tends to be smaller) will not get absorbed in our outlier estimate. For the outlier matrix $\mathbf{R}$, we will again employ log-ℓ_1 regularizer with group sparsity constraint, but in this case each frame/ column of $\mathbf{R}$ will be a group. This can be interpreted as, if one frame is representing a novel event, all of the frequency bins of that frame will have the opportunity to be active. Hence, the regularizer on $\mathbf{R}$ is,

$$\Pi(\mathbf{R}) = \sum_{l=1,..,5} \log(||\mathbf{R}_{(:,l)}||_1 + \epsilon). \quad (6)$$

2.2.4. Derivation of Multiplicative Updates

After combining the model mismatch error, and the two regularizers on activation matrix and outlier matrix, resulting cost function that we will minimize is,

$$\hat{\mathbf{H}}, \hat{\mathbf{R}} = \arg \min_{\mathbf{H}\geq 0, \mathbf{R}\geq 0} \|\mathbf{V} - \mathbf{WH} - \mathbf{R}\|_F^2 + \lambda\Psi(\mathbf{H}) + \mu\Pi(\mathbf{R}). \quad (7)$$

To solve the optimization problem with the non-negativity constraint on $\mathbf{H}$ and $\mathbf{R}$, we follow the block co-ordinate descent framework based multiplicative update rules proposed in [24] and derive the following update rules,

$$\mathbf{H}^{t+1} = \mathbf{H}^t \otimes \frac{\mathbf{W}^T\mathbf{V}}{\mathbf{W}^T(\mathbf{W}\mathbf{H}^t + \mathbf{R}^t) + \frac{\lambda}{\mathbf{S}^t+\epsilon}}, \tag{8}$$

where, $\mathbf{S}^t_{(g_i,l)} = ||vec(\mathbf{H}^t_{(g_i,:)})||_1$, for $l = 1, ..., 5$ number of frames.

$$\mathbf{R}^{t+1} = \mathbf{R}^t \otimes \frac{\mathbf{V}}{\mathbf{W}\mathbf{H}^{t+1} + \mathbf{R}^t + \frac{\mu}{\mathbf{P}^t+\epsilon}}, \tag{9}$$

where, $\mathbf{P}^t_{(k,l)} = ||\mathbf{R}^t_{(:,l)}||_1$, for $k = 1, ..., K$ and $l = 1, ..., 5$ number of frames.

These multiplicative updates are performed for 100 iterations, and the estimated value of the outlier matrix $\mathbf{R}$ is further used for an adaptive thresholding based decision function, discussed in the next subsection. We refer to our proposed approach as Robust Non-negative Block Sparse Coding (RNBSC) in later sections.

A similar robust NMF formulation has been considered before in [25], where the authors used a similar framework to learn robust subspaces for a face recognition task. But they did not consider any sparsity on the activation matrix $\mathbf{H}$ and furthermore, no structured sparsity constraint was considered on outlier matrix $\mathbf{R}$.

2.3. Adaptive Thresholding

For the adaptive thresholding step, we compute outlier estimate and model mismatch error estimate for each frame, i.e. for i^{th} frame, outlier estimate $r(i) = \|\mathbf{R}(:,i)\|_2$ and model mismatch error, $e(i) = \|\mathbf{V}(:,i) - \mathbf{W}\mathbf{H}(:,i) - \mathbf{R}(:,i)\|_2$. We also perform causal moving average of both e and r over previous 10 frames to smoothen the estimates.

Since the background level can vary a lot, instead of using a single value as the outlier threshold, we use an adaptive threshold concept, where $\theta = \beta \times \text{median}(e(1:N))$ is the threshold for each audio clip, where N is the total number of frames in each audio clip during evaluation. Finally, threshold θ is applied over time series of outlier estimate r, to obtain a binary signal, i.e. novelty or no novelty.

3. DATASETS

3.1. A3Novelty Database

The A3Novelty corpus[2] consists of 56 hours of recording and was recorded in a laboratory of the Univerita Politecnica delle Marche. These recordings were performed during both day and night time to account for different acoustic backgrounds. A variety of novel sounds, such as screams, falls, alarms, breakages of objects etc., were played back randomly using a loudspeaker during these recordings to generate backgrounds with novel sounds.

In the original A3Novelty database the audio recordings were segmented in sequences of 30 seconds. Authors of [8], randomly selected 300 sequences from the background material to construct the training set (150 minutes) and 180 sequences from background with novel sounds to compose the evaluation set (90 minutes). We have used the same database for our evaluation purposes which is publicly available[3].

[2]http://www.a3lab.dii.univpm.it/research/a3novelty

[3]http://a3lab.dii.univpm.it/webdav/audio/

3.2. Own Evaluation Database (OED)

Since the A3Novelty database was recorded indoors, it does not account for highly non-stationary acoustic backgrounds, hence, detecting impulsive novel sounds from a stationary acoustic background becomes comparatively easier. To tackle this issue, we created our own database by mixing novel audio events i.e. gunshots, glass breaking and sirens (obtained from publicly available resources[4]) with acoustic background audio recordings obtained from DCASE 2016 challenge [26], which has 6 different acoustic backgrounds: beach, bus, cafe, grocery store, city center and metro station. In our application, Event to Background ratio (EBR) is defined globally, i.e., the gain is computed from the average energy over the whole background (10 secs clip) and the event signal, to raise a global EBR. As discussed in [4], by using this approach we created signals which are good representative of real life audio events. We created these mixtures for different EBRs: 0, 5, 10, 15, 20 dB.

For each of the acoustic background (total: 6) we created 60 audio clips with 3 different type of novel sounds (gunshot, glass breaking, siren). Each clip is of length 10 s, resulting in total of 60 mins of test audio for each EBR. We also included 300 audio clips (each of 10 s) of just acoustic background. During mixing of the audio events, we generated the true labels where the resolution is 1 s, i.e., for each 10 s long clip we generate a 10 dimensional vector as true label. Along with the evaluation set, for training purposes we randomly selected segments of acoustic background recordings (disjoint of evaluation set) from the same DCASE challenge recordings, totaling to 90 mins of training data.

4. EXPERIMENTAL RESULTS

4.1. Competing Methods

In [18, 8], it has been shown that recently proposed Autoencoder (AE) based approaches perform significantly better than previously proposed statistical model based approaches such as GMM [7], HMM [7], and OC-SVM [4]. Hence, we choose to compare our proposed method with two AE based methods. We also compare against S-NMF based approach to illustrate the usefulness of robust outlier term in Equation 3. Since methods, that use any look ahead information will not be feasible for real time novelty detection application, we don't compare with structures with BLSTM units.

- **DAE-MLP:** Denoising AE with Feed Forward structure 257-512-257 and input is corrupted with Gaussian noise (std: 0.1).
- **DAE-LSTM:** Denoising AE with LSTM units in hidden layer (257-512-257) and input is corrupted with Gaussian noise (std: 0.1).
- **S-NMF:** Supervised NMF based approach, where the thresholding is done on mismatch error, i.e., $e(i) = \|\mathbf{V}(:,i) - \mathbf{W}\mathbf{H}(:,i)\|_2$.
- **RNBSC:** Proposed approach.

For a fair comparison, same adaptive thresholding approach has been employed for all competing algorithms.

4.2. Setup

We use Short Time Fourier Transform (STFT) based time-frequency representation of the audio clips and operate on STFT

[4]https://freesound.org/

Table 1: Results over A3Novelty Dataset (1 sec)

Methods	Precision (%)	Recall (%)	F Score (%)
DAE-MLP	95.00	97.43	96.20
DAE-LSTM	97.49	**100**	**98.73**
S-NMF	97.22	89.74	93.33
RNBSC (Proposed)	**100**	97.43	98.70

Table 2: Results over OED (10 secs)

Methods	Precision (%)	Recall (%)	F Score (%)
DAE-MLP	80.00	82.22	81.09
DAE-LSTM	74.29	80.27	77.17
S-NMF	78.51	76.11	77.29
RNBSC (Proposed)	**84.34**	**85.28**	**84.81**

magnitude spectra feature space. For OED we use frame size of 32 ms and a frame step 8 ms. All the audio materials have sampling fequency of 16 KHz. We use FFT size of 512, hence our feature space is $\frac{512}{2} + 1 = 257$ dimensional. For A3Novelty database following [8], 30 ms frame length and frame step of 10 ms are used.

For AE based approaches we have also tried Compression AE structures i.e., with less number of hidden units than input units. But we found out that Denoising AE structures perform better than traditional AEs, supporting results presented in [8]. Hence we only include results for DAEs. For our proposed method RNBSC, all the hyper parameters have been chosen empirically by maximizing F-score over a small held out dev set (10 % of test set) and they are as follows: $\lambda = 0.001, \mu = 0.01, \beta = 4$. For each subdictionary, representing one acoustic background, 50 basis vectors was used.

For all our experiments we use segment based performance metrics i.e., Precision, Recall and F-score, following the standard scoring techniques to evaluate sound event detection systems presented in [27]. For A3Novelty database, we use segment size of 1 s to evaluate all the algorithms, whereas for OED we use both 1 s segment and 10 s segment to score the system outputs. We found out that for EBR higher than 0 dB, recall of all the systems significantly improves. Further tuning/ increasing of β is required for those cases to reduce the false positives (increase precision). For that reason for evaluations on OED we only include testing material of 0 dB EBR. Proposed method and the competing methods have been trained separately for two datasets.

4.3. Results

In Table 1, we report the evaluation results of all competing algorithms over A3Novelty corpus. As discussed above, the lack of variability in acoustic background makes this corpus relatively easier to detect novelties, hence all the competing algorithms produce F-score over 90%. DAE-LSTM and our proposed method RNBSC performs the best among 4 methods (DAE-LSTM does slightly better (0.03%) than the proposed method). Our results using DAE-LSTM is also comparable (our reported result is better) to what was reported in [18] using DAE-LSTM structure for this corpus.

In Table 2 and Table 3 we report evaluation results over OED for all competing methods using 10 s segment and 1 s segment respectively. For both the cases, our proposed approach RNBSC performs the best and outperforms AE based approaches. We also report results using S-NMF and show the usefulness of the robustness, i.e., the extra supergaussian outlier term.

Fig. 1 shows the outlier estimate (r) and model mismatch error (e) for an audio clip with cafe acoustic background, and a novel gunshot sound. We highlight in Fig. 1, the position of the gun shot in the spectrogram. In the bottom figure of Fig. 1, we clearly see that the novel sound is being captured in the outlier estimate and not leaking in to the model mismatch error. Because of the supergaussian/sparse constraint, for all other times energy of the outlier term is close to zero.

Table 3: Results over OED (1 sec)

Methods	Precision (%)	Recall (%)	F Score (%)
DAE-MLP	72.31	74.72	73.49
DAE-LSTM	70.95	70.55	70.75
S-NMF	75.08	67.78	71.24
RNBSC (Proposed)	**77.17**	**81.67**	**79.35**

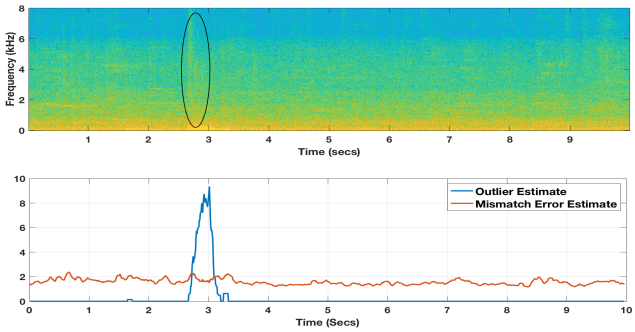

Figure 1: Spectrogram (top), Outlier and Mismatch Error estimate using RNBSC (Bottom) for an audio clip in cafe with Gunshot (novel sound)

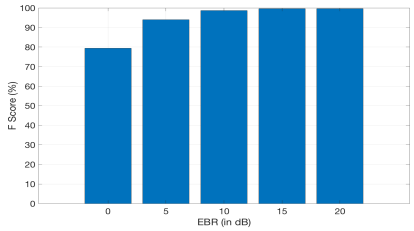

Figure 2: F Score using RNBSC over OED (1 sec) for different EBR

Finally in Fig. 2, we show F score measures of our proposed method for different Event to Background Ratio.

5. CONCLUSION

We have presented a novel unsupervised approach for acoustic novelty detection, using robust non-negative block sparse coding. Previous state-of-the-art autoencoder based approaches solve the problem by modeling only the normal acoustic background, and they detect novel sounds only when the reconstruction/ model mismatch error is above a certain threshold. Our approach on the other hand explicitly models the novel sound using a supergaussian random variable and thresholds on the energy of the expected value of that random variable to detect acoustic novelties. This makes our system much more robust in highly non-stationary acoustic backgrounds, as shown by our empirical results over OED, which has 6 different acoustic backgrounds.

6. REFERENCES

[1] M. Valera and S. A. Velastin, "Intelligent distributed surveillance systems: a review," *IEE Proceedings-Vision, Image and Signal Processing*, vol. 152, no. 2, pp. 192–204, 2005.

[2] G. Valenzise, L. Gerosa, M. Tagliasacchi, F. Antonacci, and A. Sarti, "Scream and gunshot detection and localization for audio-surveillance systems," in *Advanced Video and Signal Based Surveillance, 2007. AVSS 2007. IEEE Conference on*, pp. 21–26, IEEE, 2007.

[3] J. T. Geiger and K. Helwani, "Improving event detection for audio surveillance using gabor filterbank features," in *Signal Processing Conference (EUSIPCO), 2015 23rd European*, pp. 714–718, IEEE, 2015.

[4] S. Lecomte, R. Lengellé, C. Richard, F. Capman, and B. Ravera, "Abnormal events detection using unsupervised one-class svm-application to audio surveillance and evaluation," in *Advanced Video and Signal-Based Surveillance (AVSS), 2011 8th IEEE International Conference on*, pp. 124–129, IEEE, 2011.

[5] F. Aurino, M. Folla, F. Gargiulo, V. Moscato, A. Picariello, and C. Sansone, "One-class svm based approach for detecting anomalous audio events," in *Intelligent Networking and Collaborative Systems (INCoS), 2014 International Conference on*, pp. 145–151, IEEE, 2014.

[6] R. Bardeli and D. Stein, "Uninformed abnormal event detection on audio," in *Speech Communication; 10. ITG Symposium; Proceedings of*, pp. 1–4, VDE, 2012.

[7] S. Ntalampiras, I. Potamitis, and N. Fakotakis, "Probabilistic novelty detection for acoustic surveillance under real-world conditions," *IEEE Transactions on Multimedia*, vol. 13, no. 4, pp. 713–719, 2011.

[8] E. Marchi, F. Vesperini, F. Eyben, S. Squartini, and B. Schuller, "A novel approach for automatic acoustic novelty detection using a denoising autoencoder with bidirectional lstm neural networks," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, pp. 1996–2000, IEEE, 2015.

[9] L. Tarassenko, P. Hayton, N. Cerneaz, and M. Brady, "Novelty detection for the identification of masses in mammograms," 1995.

[10] L. Clifton, D. A. Clifton, P. J. Watkinson, and L. Tarassenko, "Identification of patient deterioration in vital-sign data using one-class support vector machines," in *Computer Science and Information Systems (FedCSIS), 2011 Federated Conference on*, pp. 125–131, Citeseer, 2011.

[11] Y.-H. Liu, Y.-C. Liu, and Y.-J. Chen, "Fast support vector data descriptions for novelty detection," *IEEE Transactions on Neural Networks*, vol. 21, no. 8, pp. 1296–1313, 2010.

[12] C. Surace and K. Worden, "Novelty detection in a changing environment: a negative selection approach," *Mechanical Systems and Signal Processing*, vol. 24, no. 4, pp. 1114–1128, 2010.

[13] A. Patcha and J.-M. Park, "An overview of anomaly detection techniques: Existing solutions and latest technological trends," *Computer networks*, vol. 51, no. 12, pp. 3448–3470, 2007.

[14] M. Markou and S. Singh, "A neural network-based novelty detector for image sequence analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 10, pp. 1664–1677, 2006.

[15] M. Markou and S. Singh, "Novelty detection: a review—part 1: statistical approaches," *Signal processing*, vol. 83, no. 12, pp. 2481–2497, 2003.

[16] M. Markou and S. Singh, "Novelty detection: a review—part 2:: neural network based approaches," *Signal processing*, vol. 83, no. 12, pp. 2499–2521, 2003.

[17] Y. Kawaguchi, T. Endo, K. Ichige, and Hamada, "Non-negative novelty extraction: A new non-negativity constraint for nmf," in *Acoustic Signal Enhancement (IWAENC), 2018 IEEE International Workshop on*, IEEE, 2018.

[18] E. Marchi, F. Vesperini, S. Squartini, and B. Schuller, "Deep recurrent neural network-based autoencoders for acoustic novelty detection," *Computational intelligence and neuroscience*, vol. 2017, 2017.

[19] J. Le Roux, F. J. Weninger, and J. R. Hershey, "Sparse nmf–half-baked or well done?," *Mitsubishi Electric Research Labs (MERL), Cambridge, MA, USA, Tech. Rep., no. TR2015-023*, 2015.

[20] A. Lefevre, F. Bach, and C. Févotte, "Itakura-saito nonnegative matrix factorization with group sparsity," in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, pp. 21–24, IEEE, 2011.

[21] D. E. Tyler, "Robust statistics: Theory and methods," 2008.

[22] K. L. Lange, R. J. Little, and J. M. Taylor, "Robust statistical modeling using the t distribution," *Journal of the American Statistical Association*, vol. 84, no. 408, pp. 881–896, 1989.

[23] Y. Jin and B. D. Rao, "Algorithms for robust linear regression by exploiting the connection to sparse signal recovery," in *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, pp. 3830–3833, IEEE, 2010.

[24] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Advances in neural information processing systems*, pp. 556–562, 2001.

[25] R. Zhao and V. Y. Tan, "Online nonnegative matrix factorization with outliers," in *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*, pp. 2662–2666, IEEE, 2016.

[26] A. Mesaros, T. Heittola, and T. Virtanen, "Tut database for acoustic scene classification and sound event detection," in *Signal Processing Conference (EUSIPCO), 2016 24th European*, pp. 1128–1132, IEEE, 2016.

[27] A. Mesaros, T. Heittola, and T. Virtanen, "Metrics for polyphonic sound event detection," *Applied Sciences*, vol. 6, no. 6, p. 162, 2016.