A NOVEL APPROACH FOR AUTOMATIC ACOUSTIC NOVELTY DETECTION USING A DENOISING AUTOENCODER WITH BIDIRECTIONAL LSTM NEURAL NETWORKS

Erik Marchi[1]*, Fabio Vesperini*[2]*, Florian Eyben*[1]*, Stefano Squartini*[2]*, Björn Schuller*[3,4,1]

[1]Machine Intelligence & Signal Processing Group, Technische Universität München, Germany
[2]A3LAB, Department of Information Engineering, Università Politecnica delle Marche, Italy
[3]Chair of Complex and Intelligent Systems, University of Passau, Germany
[4]Department of Computing, Imperial College London, UK

ABSTRACT

Acoustic novelty detection aims at identifying abnormal/novel acoustic signals which differ from the reference/normal data that the system was trained with. In this paper we present a novel unsupervised approach based on a denoising autoencoder. In our approach auditory spectral features are processed by a denoising autoencoder with bidirectional Long Short-Term Memory recurrent neural networks. We use the reconstruction error between the input and the output of the autoencoder as activation signal to detect novel events. The autoencoder is trained on a public database which contains recordings of typical in-home situations such as talking, watching television, playing and eating. The evaluation was performed on more than 260 different abnormal events. We compare results with state-of-the-art methods and we conclude that our novel approach significantly outperforms existing methods by achieving up to 93.4 % F-Measure.

Index Terms— Acoustic Novelty Detection, Denoising Autoencorder, Bidirectional LSTM, Recurrent Neural Networks

1. INTRODUCTION

Novelty detection is a challenging classification task that aims at recognising situations in which unusual events occur. The problem can be treated as a one-class classification task: typically the amount of normal data consists of a very large set, and the normal class can be accurately modelled, whereas the acoustic events lying 'outside' of the class are considered as *novel* events. Several approaches have been proposed for the practical importance of the novelty detection, especially for automatic monitoring systems.

In the past years, many systems have been developed for surveillance applications. Surveillance is carried out, e. g., to ensure public safety or for safety-oriented supervision of private environments where people may live alone. In fact, the increasing desire in public security over the past decades has motivated the installation of sensors such as cameras or microphones in public places (stores, subway, airports, etc.). Thus, the need of unsupervised situation assessment stimulated the signal processing community to experiment with several according automated frameworks. Data-driven classification approaches, relying on a-priori classification of the data, were applied for a successful operation and recognition of the events. Usually, the research in the area of automatic surveillance systems is mainly focused on detecting abnormal events based on the acquired video information. However, the information given by the acoustic signal offers several advantages, such as low computational needs or the fact that the illumination conditions of the space to be monitored do not have an effect on the sound; the same applies for possible occlusion or fast events like shots or explosions. The statistical approach is the most widely used for this problem. It consists of modelling data based on its statistical properties and using this information to estimate whether a test sample comes from the same distribution or not.

1.1. Related work

Statistical and probabilistic approaches are the most commonly used in the field of novelty detection. Novelty detection ranges from automatic recognition of handwriting, the recognition of cancer [1], informatic intrusion detection systems, non-destructive inspection for the analysis of mechanical components [2, 3], to audio segmentation [4], and many others.

In the past decade, a pioneering study investigated the relationship between the degree of novelty of the input data and the corresponding reliability of the outputs from the neural network, and demonstrated its performance with an application to the control of the oil flow in multiphase pipelines [5]. Subsequent works proposed the application of a Compression Autoencoder neural network to detect abnormal CPU data usage [6, 7]. In further works [8–11], the use of compression autoassociators for outlier detection was studied and in [12], the autoassociator was applied for the task of damage classification under changing environmental conditions.

Several studies exist in the field of acoustic event classification applying GMM and HMM to detect human presence (speech, laughter, cough) , animal sounds, sounds of objects [13, 14] and sounds caused by various types of guns [15]. Since the reconstruction of all possible types of abnormal events is not practically doable, the need of new unsupervised machine learning approaches – able to recognise unknown data – showed a rising interest in the research community. However, to our best knowledge, very few studies investigated unsupervised approaches for acoustic novelty detection. Such studies investigated mostly HMM- and GMM-based approaches for acoustic surveillance of abnormal situations [16] and for automatic space monitoring [17].

1.2. Contribution of this work

A novel purely unsupervised approach to acoustic novelty detection is proposed. It relies on auditory spectral features and Denoising Autoencoders (DAEs) with bidirectional Long Short-Term Memory

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement No. 289021 (ASC-Inclusion) and No. 338164 (ERC Starting Grant iHEARu). Correspondence should be addressed to erik.marchi@tum.de.

(BLSTM) recurrent neural networks (RNNs) to detect novel events. The auditory spectral features are processed by an autoencoder, which acts as a one-class classifier. Our approach relies on the reconstruction error that the denoising autoencoder commits trying to reconstruct a novel sound which the network has never seen in the training phase. We compare our results with state-of-the-art methods and conclude that our novel approach significantly outperforms existing methods by achieving up to 93.4 % F-Measure on the test data.

The article is structured as follows: First, a basic description of the denoising autoencoder for acoustic novelty detection is given (Section 2); then, we define the features and the experimental set-up and present evaluation results (Section 4) before concluding the paper in Section 5.

2. DENOISING AUTOENCODER FOR ACOUSTIC NOVELTY DETECTION

This section introduces the basic concepts of autoencoders and describes the basic autoencoder, compression autoencoder and denoising autoencoders.

2.1. Basic Autoencoder

A basic autoencoder – a kind of neural network typically consisting of only one hidden layer –, sets the target values to be equal to the input. Deep neural networks use it during training of hidden layers to find common data representation from the input [18, 19]. Formally, in response to an input example $x \in \mathbf{R}^n$, the hidden representation $h(x) \in \mathbf{R}^m$ is

$$h(x) = f(W_1 x + b_1), \tag{1}$$

where $f(z)$ is a non-linear activation function, typically a logistic sigmoid function $f(z) = 1/(1+\exp(-z))$ applied component-wise, $W_1 \in m \times n$ is a weight matrix, and $b_1 \in \mathbf{R}^m$ is a bias vector.

The network output maps the hidden representation h back to a reconstruction $\tilde{x} \in \mathbf{R}^n$:

$$\tilde{x} = f(W_2 h(x) + b_2), \tag{2}$$

where $W_2 \in n \times m$ is a weight matrix, and $b_2 \in \mathbf{R}^n$ is a bias vector.

Given an input set of examples $\mathcal{X}$, autoencoder training consists in finding parameters $\theta = \{W_1, W_2, b_1, b_2\}$ that minimise the reconstruction error, which corresponds to minimising the following objective function:

$$\mathcal{J}(\theta) = \sum_{x \in \mathcal{X}} \|x - \tilde{x}\|^2 . \tag{3}$$

The minimization is usually realised by stochastic gradient descent as often used in the training of neural networks.

2.2. Compression Autoencoder

In the case of having the number of hidden units m less than the number of input units n, the network is forced to learn a compressed representation of the input. For example, if some of the input features are correlated, then this Compression Autoencoder (CAE) is able to learn those correlations and reconstruct the input data from a compressed representation.

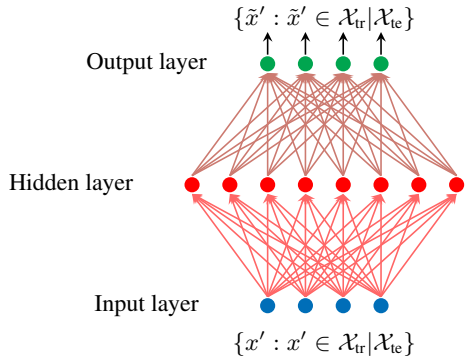

Fig. 1: Structure of the denoising autoencoder (DAE) on the training set $\mathcal{X}_{\text{tr}}$ or testing set $\mathcal{X}_{\text{te}}$. $\mathcal{X}_{\text{tr}}$ contains data of non-novel acoustic events; $\mathcal{X}_{\text{te}}$ consists of *novel* and *non-novel* acoustic events.

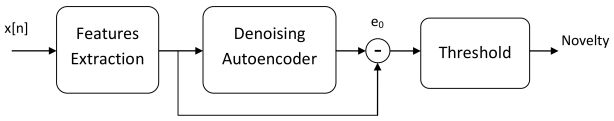

Fig. 2: Block diagram of the proposed acoustic novelty detector with a denoising autoencoder. The features are extracted from the input signal; the reconstruction error between the input and the reconstructed features is then processed by a thresholding block which detects the *novel* or *non-novel* event.

2.3. Denoising Autoencoder

The idea of denoising autoencoders [20] is quite intuitive. In order to force the hidden layer to retrieve more robust features and prevent it from simply learning the identity, the autoencoder is trained to reconstruct the input from a corrupted version of it.

Formally, the initial input x is corrupted by means of additive isotropic Gaussian noise in order to obtain: $x'|x \sim N(x, \sigma^2 I)$. The corrupted input x' is then mapped, as with the basic autoencoder, to a hidden representation

$$h(x') = f(W_1' x' + b_1'), \tag{4}$$

from which we reconstruct a the original signal as follows:

$$\tilde{x}' = f(W_2' x + b_2'). \tag{5}$$

The parameters $\theta' = \{W_1', W_2', b_1', b_2'\}$ are trained to minimise the average reconstruction error over the training set, to have $\tilde{x}'$ as close as possible to the uncorrupted input x, which corresponds to minimising the objective function in Equation 3. In our approach, the training set $\mathcal{X}_{\text{tr}}$ consists of background environmental sounds, and the test set $\mathcal{X}_{\text{te}}$ consists of recordings containing 'abnormal' sounds. The structure of the denoising autoencoder is shown in Figure 1, and a block diagram of the proposed novelty detector is depicted in Figure 2.

2.4. Features

Auditory Spectral Features (ASF) [21] are computed by applying the Short Time Fourier Transformation (STFT) using a frame size of 30 ms and a frame step of 10 ms. Each STFT yields the power spectrogram which is converted to the Mel-Frequency scale using a

filter-bank with 26 triangular filters obtaining the Mel spectrograms $M_{30}(n,m)$. Finally, to match the human perception of loudness, a logarithmic representation is chosen:

$$M^{30}_{log}(n,m) = log(M_{30}(n,m) + 1.0). \quad (6)$$

In addition, the positive first order differences $D_{30}(n,m)$ are calculated from each Mel spectrogram as follows:

$$D_{30}(n,m) = M^{30}_{log}(n,m) - M^{30}_{log}(n-1,m). \quad (7)$$

Furthermore, the frame energy is also included as a feature thus leading to a total number of 54 features. The features are extracted with our open-source audio feature extractor openSMILE [22].

3. BLSTM RECURRENT NEURAL NETWORK AND THRESHOLDING

Suitable types of networks for our purpose are RNNs and Bidirectional RNNs with LSTM units instead of 'usual' non-linear ones. In addition to LSTM memory blocks, we use bidirectional RNNs [23]. A bidirectional RNN can access context from both temporal directions. This is achieved by processing the input data in both directions with two separate hidden layers. Both hidden layers are then fed to the output layer. The combination of bidirectional RNNs and LSTM memory blocks leads to bidirectional LSTM networks [24], where context from both temporal directions is exploited. It has to be noted that using bidirectional LSTM networks makes it impossible to use the system for online processing as a look-ahead buffer is needed. BLSTM networks have been already applied widely in other tasks with remarkable performance. We conducted several preliminary evaluations to find the best network layout by varying the number of hidden layers and their size (i. e., the number of LSTM units for each layer). The best network layout for our RNNs has three hidden layers with 156, 256, and 156 LSTM units, respectively. The best network layout for our BRNNs has six hidden layers (three for each direction) with 216 LSTM units, each. Supervised learning was applied with up to 100 epochs for training the network. Network weights are recursively updated by standard gradient descent with backpropagation of the sum of squared error (SSE). The gradient descent algorithm requires the network weights to be initialised with non zero values; thus we initialise the weights with a random Gaussian distribution with mean 0 and standard deviation 0.1. The input and output layers of the network have 54 units. Thus, the trained autoencoder is able to reconstruct each sample and novel events are identified by processing the reconstruction error with an adaptive threshold. The input x is segmented into sequences of 30 seconds of length. For every time-step, the Euclidean distance between each standardised input feature value and the network's output is computed. The distances are summed up and divided by the number of coefficients in order to represent the reconstruction error of each time-step with a single value. In order to obtain an optimal novelty detection, a threshold θ is then applied to obtain a binary signal. This threshold is proportional to the median of the error signal of a sequence e_0 by a multiplicative coefficient β, constrained to the range from $\beta_{min} = 1$ to $\beta_{max} = 2$:

$$\theta' = \beta * median(e_0(1), ..., e_0(N)). \quad (8)$$

Figure 3 shows the reconstruction error for a given sequence. The figure clearly depicts a low reconstruction error in reproducing 'normal' input such as talking, television sounds and other 'normal' environmental sounds. On the other hand, the denoising autoencoder shows a high reconstruction error when it comes to reproducing novel acoustic events such as a scream, or an alarm.

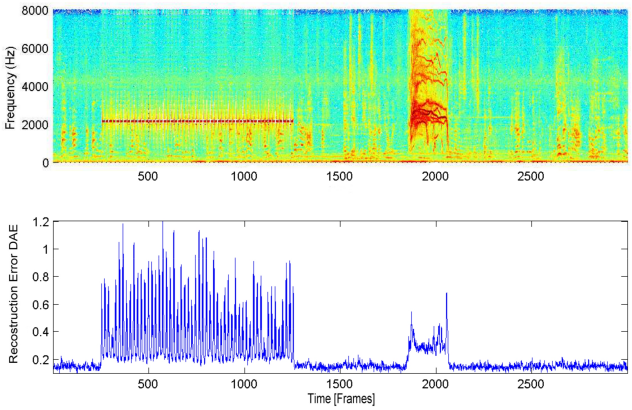

Fig. 3: *Top*: Spectrogram of a 30 seconds sequence containing two novel events, such as a siren and a scream. *Bottom*: Reconstruction error signal of the related sequence.

4. EXPERIMENTS AND RESULTS

This section contains the data set used for our evaluation (Section 4.1), the experiments' setup (Section 4.2), and a description of the performances obtained with the proposed approach (Section 4.3).

4.1. Evaluation Data Set

Our evaluation dataset is composed of around three hours of recordings of a home environment, taken from the PASCAL CHiME speech separation and recognition challenge dataset [25]. It consists of a typical in-home scenario (a living room), recorded during different days and times, while the inhabitants (two adults and two children) perform common actions, such as talking, watching television, playing, eating. We used randomly chosen sequences to compose 100 minutes of background for training set, and around 70 minutes for testing set. The testing set was generated adding digitally and randomly different kinds of sounds[1], such as screams, alarms, falls, and fractures (cf. Table 1). The original dataset was recorded in 2 channels (with a binaural microphone) and a sample-rate of 16 kHz.

4.2. Experimental Setup

Several experiments were conducted, to find the the most suitable setup. The networks were trained with the gradient steepest descent algorithm on the SSE with a fixed momentum of 0.9, at different constant values of learning rate $l = \{1e^{-4}, 1e^{-5}, ..., 1e^{-8}\}$, and

[1] taken from www.freesound.org

Table 1: Novel acoustic events in the test set. Shown are the number of different events, average durations and the total duration in seconds per event type.

Type	# Events	Avg. Duration (s)	Total Duration (s)
Alarm	76	6.0	435.8
Scream	111	1.9	214.6
Falls	48	1.8	89.5
Fracture	32	2.2	70.4
Total	267	2.4	810.3

different noise sigma values $\sigma = \{0.1, 0.25, 0.5\}$. In the case of compression BLSTM and LSTM, no additive Gaussian noise was applied. The autoencoders were trained using our open-source CUDA RecurREnt Neural Network Toolkit (CURRENNT) [26]. As evaluation metrics we used Precision, Recall, and F-measure. We evaluated several topologies for the denoising autoencoder networks ranging from 54-128-54 to 270-370-270, and from 54-20-54 to 54-54-54 in the case of compression/basic autoencoder. Every network topology was evaluated for each 100 epochs of training. In order to compare our results with the state of the art methods, we employed further two typical approaches based on GMM and HMM. In the case of GMMs, models were trained at different numbers of Gaussian components 2^n with $n = \{1, 2, ..., 8\}$, whereas left-right HMMs were trained with different numbers of states $s = \{3, 4, 5\}$ and 2^n Gaussian components with $n = \{1, 2, ..., 7\}$. GMMs and HMMs were trained using the *Torch* [27] toolkit. The log-likelihood signal produced as output of the probabilistic models was post-processed with a similar thresholding algorithm (cf. Section 3) in order to fairly compare the performances among the different methods. For all the experiments and settings we maintained the same feature set.

4.3. Results

Table 2 reports performances for three different σ of the additive Gaussian noise using a BLSTM-DAE and a LSTM-DAE, respectively. We evaluated several layouts (cf. Section 4.2) per network type, but we show only the best two. Setting an input noise standard deviation of 0.25 shows best performances of up to 93.4 % F-Measure in the BLSTM network, whereas for LSTM we observe better performances with an input noise standard deviation of 0.1. The valuable behaviour of the BLSTM-DAE might be due to the ability of BLSTM to access future frames, which LSTM cannot.

As an overall evaluation on the test set, Table 3 shows the comparison between state-of-the-art methods and our proposed approach in terms of F-Measure, Precision, and Recall. We observe that the proposed BLSTM-DAE method provided the best performance in terms of Precision, Recall, and F-Measure of up to 94.7 %, 92.0 % and 93.4 %, respectively. A significant absolute improvement (one-tailed z-test [28], p<0.01) of 2.0 % F-Measure is observed against the HMM-based approach, while an absolute improvement of 3.0 % F-measure is exhibited with respect to the GMM-based method. It is not surprising that HMMs show better performances than GMMs since HMMs consider the temporal evolution of the signal which is relevant in modelling and decoding in this task.

The DAE approach outperforms also CAE on both network types – in fact, a maximum of 3.3 % absolute improvement is observed between LSTM-CAE and LSTM-DAE. CAE was evaluated under different layouts by increasing the hidden layer units from 20 to 40.

Table 2: Best F-Measure (%) per different topology and network types obtained by varying the amount of noise σ used in the training phase.

F-Measure (%)			
	Input noise σ		
Network (layout)	0.1	0.25	0.5
LSTM-DAE (*156-256-156*)	**92.9**	92.0	92.4
LSTM-DAE (*216-216-216*)	92.2	92.1	92.0
BLSTM-DAE (*156-256-156*)	92.5	93.2	92.8
BLSTM-DAE (*216-216-216*)	92.3	**93.4**	92.6

In addition, we investigated if the gain by DAE might be dependent on the dimension of the layers. We evaluated several layouts ranging from 54-20-54 to 54-54-54 obtaining results in the range of 91.3 %, which are clearly not comparable with the DAE performance. This leads to the conclusion that the proposed learning approach brings valuable information compared to other approaches.

The strength of a DAE is its ability of encoding the input by preserving the information about the input itself and simultaneously undoing the effect of a corruption process applied to the input of the auto-encoder. The combination of these two learning processes seems to be effective in our task. On the other hand, CAE applies only a single compression learning process which apparently is not sufficient to encode effectively information about the input.

Table 3: Comparison with existing methods percentage of Precision, Recall, and F-Measure. Reported approaches are: GMMs, HMMs, Compression Autoencoder with BLSTM (BLSTM-CAE) or LSTM (LSTM-CAE), denoising autoencoder with BLSTM (BLSTM-DAE) or LSTM (LSTM-DAE).

Method	**Precision** (%)	**Recall** (%)	F**-measure** (%)
GMM	91.1	87.8	89.4
HMM	94.1	88.9	91.4
LSTM-CAE	91.7	86.6	89.1
BLSTM-CAE	93.6	89.2	91.3
LSTM-DAE	94.2	90.6	92.4
BLSTM-DAE	**94.7**	**92.0**	**93.4**

5. CONCLUSIONS AND OUTLOOK

We have presented a novel, purely unsupervised approach to acoustic novelty detection. It relies on auditory spectral features and denoising autoencoders with bidirectional Long Short-Term Memory acting as a one-class classifier. Our approach exploits the reconstruction error of the denoising autoencoder when trying to denoise a novel sound which the network has never seen in the training phase. The strength of the BLSTM-DAE is owed to the combination of two learning processes: encoding the input by preserving the information about the input, and simultaneously removing the corruption process applied to the input. Additionally applying LSTM and BLSTM enables the system to use and learn more context which clearly helps in our task. We compare results with state-of-the-art methods and we conclude that our novel approach significantly outperforms existing methods by achieving up to 93.4 % F-Measure with an absolute improvement of 2 % over a HMM system.

Future works are intended to use different type of features, likely more suitable to deal with non-stationary events, as already considered by some of the authors in the musical onset case study [29]. Moreover, further efforts will be oriented to evaluate the effectiveness of the approach with real-life databases.

6. ACKNOWLEDGEMENT

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement No. 289021 (ASC-Inclusion) and No. 338164 (ERC Starting Grant iHEARu).

7. REFERENCES

[1] L. Tarassenko, P. Hayton, N. Cerneaz, and M. Brady, "Novelty detection for the identification of masses in mammograms," in *Proceedings Fourth International Conference on Artificial Neural Networks, 1995.* IET, 1995, pp. 442–447.

[2] K. Worden, G. Manson, and D. Allman, "Experimental validation of a structural health monitoring methodology: Part i. novelty detection on a laboratory structure," *Journal of Sound and Vibration*, vol. 259, no. 2, pp. 323–343, 2003.

[3] L. Clifton, H. Yin, and Y. Zhang, "Support vector machine in novelty detection for multi-channel combustion data," in *Advances in Neural Networks - ISNN 2006*, J. Wang, Z. Yi, J. Zurada, B.L. Lu, and H. Yin, Eds., vol. 3973 of *Lecture Notes in Computer Science*, pp. 836–843. Springer Berlin Heidelberg, 2006.

[4] J. Foote, "Automatic audio segmentation using a measure of audio novelty," in *Proceedings IEEE International Conference on Multimedia and Expo, ICME 2000.* IEEE, 2000, vol. 1, pp. 452–455.

[5] C.M. Bishop, "Novelty detection and neural network validation," *IEEE Vision, Image and Signal Processing*, vol. 141, no. 4, pp. 217–222, Aug 1994.

[6] N. Japkowicz, C. Myers, M. Gluck, et al., "A novelty detection approach to classification," in *Proceedings International Joint Conference on Artificial Intelligence, IJCAI 1995*, 1995, pp. 518–523.

[7] B.B. Thompson, R.J. Marks, J.J. Choi, M.A. El-Sharkawi, M.Y. Huang, and C. Bunje, "Implicit learning in autoencoder novelty assessment," in *Neural Networks, 2002. IJCNN'02. Proceedings of the 2002 International Joint Conference on*. IEEE, 2002, vol. 3, pp. 2878–2883.

[8] S. Hawkins, H. He, G. Williams, and R. Baxter, "Outlier detection using replicator neural networks," in *Data Warehousing and Knowledge Discovery*, pp. 170–180. Springer, 2002.

[9] N. Japkowicz, "Supervised versus unsupervised binary-learning by feedforward neural networks," *Machine Learning*, vol. 42, no. 1-2, pp. 97–122, 2001.

[10] L. Manevitz and M. Yousef, "One-class document classification via neural networks," *Neurocomputing*, vol. 70, no. 79, pp. 1466 – 1481, 2007.

[11] G. Williams, R. Baxter, H. He, S. Hawkins, and L. Gu, "A comparative study of rnn for outlier detection in data mining," in *Proceedings IEEE 13th International Conference on Data Mining, 2002*. IEEE Computer Society, 2002, pp. 709–709.

[12] H. Sohn, K. Worden, and C.R. Farrar, "Statistical damage classification under changing environmental and operational conditions," *Journal of Intelligent Material Systems and Structures*, vol. 13, no. 9, pp. 561–574, 2002.

[13] P. Atrey, M. Maddage, and M. Kankanhalli, "Audio based event detection for multimedia surveillance," in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2006.*, May 2006, vol. 5, pp. V–V.

[14] A. Harma, M.F. McKinney, and J. Skowronek, "Automatic surveillance of the acoustic activity in our living environment," in *IEEE International Conference on Multimedia and Expo, ICME 2005*. IEEE, 2005, p. 4.

[15] C. Clavel, T. Ehrette, and G. Richard, "Events detection for an audio-based surveillance system," in *Multimedia and Expo, 2005. ICME 2005. IEEE International Conference on*, July 2005, pp. 1306–1309.

[16] 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.

[17] S. Ntalampiras, I. Potamitis, and N. Fakotakis, "On acoustic surveillance of hazardous situations," in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2009.* IEEE, 2009, pp. 165–168.

[18] I. Goodfellow, H. Lee, Q.V. Le, A. Saxe, and A.Y. Ng, "Measuring invariances in deep networks," in *Advances in Neural Information Processing Systems 22*, Y. Bengio, D. Schuurmans, J. Lafferty, C. Williams, and A. Culotta, Eds., pp. 646–654. Curran Associates, Inc., 2009.

[19] Y. Bengio, P. Lamblin, D. Popovici, and H. Larochelle, "Greedy layer-wise training of deep networks," in *Advances in Neural Information Processing Systems 19*, B. Schölkopf, J. Platt, and T. Hoffman, Eds., pp. 153–160. MIT Press, 2007.

[20] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P.A. Manzagol, "Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion," *Journal of Machine Learning Research*, vol. 11, pp. 3371–3408, Dec. 2010.

[21] F. Eyben, S. Böck, B. Schuller, and A. Graves, "Universal onset detection with bidirectional long short-term memory neural networks.," in *ISMIR*, 2010, pp. 589–594.

[22] F. Eyben, M. Wöllmer, and B. Schuller, "openSMILE – The Munich Versatile and Fast Open-Source Audio Feature Extractor," in *Proceedings of the 18th ACM International Conference on Multimedia, MM 2010*, Florence, Italy, October 2010, ACM, pp. 1459–1462, ACM.

[23] M. Schuster and K.K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Transactions on Signal Processing*, vol. 45, no. 11, pp. 2673–2681, Nov 1997.

[24] A. Graves and J. Schmidhuber, "Framewise phoneme classification with bidirectional {LSTM} and other neural network architectures," *Neural Networks*, vol. 18, no. 56, pp. 602 – 610, 2005, {IJCNN} 2005.

[25] J. Barker, E. Vincent, N. Ma, H. Christensen, and P. Green, "The pascal chime speech separation and recognition challenge," *Computer Speech & Language*, vol. 27, no. 3, pp. 621–633, 2013.

[26] F. Weninger, J. Bergmann, and B. Schuller, "Introducing CURRENNT, the Munich open-source CUDA RecurREnt Neural Network Toolkit," *Journal of Machine Learning Research*, 2014, in press.

[27] R. Collobert, K. Kavukcuoglu, and C. Farabet, "Torch7: A matlab-like environment for machine learning," in *BigLearn, NIPS Workshop*, 2011, number EPFL-CONF-192376.

[28] M. Smucker, J. Allan, and B. Carterette, "A comparison of statistical significance tests for information retrieval evaluation," in *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*. ACM, 2007, pp. 623–632.

[29] E. Marchi, G. Ferroni, F. Eyben, L. Gabrielli, S. Squartini, and B. Schuller, "Multi-resolution linear prediction based features for audio onset detection with bidirectional lstm neural networks," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, May 2014, pp. 2164–2168.