Hierarchic ConvNets Framework for Rare Sound Event Detection

Fabio Vesperini, Diego Droghini, Emanuele Principi, Leonardo Gabrielli, Stefano Squartini
Department of Information Engineering
Università Politecnica delle Marche, Ancona, Italy
f.vesperini@pm.univpm.it

***Abstract*—In this paper, we propose a system for rare sound event detection using a hierarchical and multi-scaled approach based on Convolutional Neural Networks (CNN). The task consists on detection of event onsets from artificially generated mixtures. Spectral features are extracted from frames of the acoustic signals, then a first event detection stage operates as binary classifier at frame-rate and it proposes to the second stage contiguous blocks of frames which are assumed to contain a sound event. The second stage refines the event detection of the prior network, discarding blocks that contain background sounds wrongly classified by the first stage. Finally, the effective onset time of the active event is obtained. The performance of the algorithm has been assessed with the material provided for the second task of the IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Events (DCASE) 2017. The achieved overall error rate and F-measure, resulting respectively equal to 0.22 and 88.50% on the evaluation dataset, significantly outperforms the challenge baseline and the system guarantees improved generalization performance with a reduced number of free network parameters w.r.t. other competitive algorithms.**

***Index Terms*—Convolutional Neural Network, Sound Event Detection, DCASE2017, Linear Prediction, Discrete Wavelet Transform**

I. Introduction

Nowadays, one of the most important tasks in the field of computational auditory scene analysis (CASA) is the automatic sound event detection (SED), which can be exploited in various application areas, ranging from acoustic surveillance [1], [2] and multimedia event detection [3] to smart home devices [4]–[6]. In particular, SED is defined as the task of analysing a continuous audio stream in order to extract a description of the sound events occurring in it. This description is commonly expressed as a label that marks the start, the ending, and the nature of the occurred sound (e.g., children crying, cutlery, glass jingling).

The "Detection of rare sound events" task of the 2017 Detection and Classification of Acoustic Scenes and Events (DCASE) challenge [7] consisted in determining the presence and the precise onset time of three types of sounds, "baby cry", "glass break" and "gun shot" in artificially generated audio sequences. The task takes into account real-world issues that introduce additional complexity to the problem, such as the acoustic variability of the sounds belonging to each event class, the presence of environmental noise and its variability, etc. The rules of the challenge allow to know *a priori* the event typology possibly present in the audio sequence under examination, thus it is possible to have a separate binary classifier for each class.

We gratefully acknowledge the support of NVIDIA Corporation with the donation of the GPU used for this research.

A. Related Works

In the recent era of the "Deep Learning" different approaches to SED have been proposed marking use of the capabilities of deep neural networks (DNNs) to learn the relation between time-frequency features of the raw audio signal and a target vector representing sound events. Although the DNNs based systems are more computationally intensive with respect to widely used statistical modelling methods such as hidden Markov models (HMMs) or Gaussian mixture models (GMMs) [8], [9], a comparative study [10] has highlighted that they are able to achieve top performance in the sound recognition problem.

A well-fitting example of such performance is given in [11], where different DNNs are trained on three datasets recorded in real life environments in order to detect abnormal events or hazardous situations exploiting only the information carried by the acoustic signal. The experimental results show that Deep Recurrent Neural Networks (DRNNs) outperform the probabilistic approaches over the three databases.

In occasion of the DCASE 2017 challenge, many novel systems featuring deep neural networks have been proposed, in particular involving hybrid architectures making use of Convolutional Neural Networks (CNN) and DRNNs. In detail, both the first two classified algorithms make use of mel spectrogram coefficients as spectral representation of the audio signal which is processed by a CNN with 1D filters in the case of the first ranked [12] or by a 2D CNN with frequency pooling in the case of the second classified [13]. The architectures are, then, combined with recurrent layers to process the features obtained by the convolutional blocks. In [14] the authors propose a hierarchical structure based on CNNs and DNNs trained with multi-task loss functions. Specifically, in the first stage the networks are trained for background noise rejection, using a weighted loss function to penalize the false positive errors. In the second stage the multi-task loss enables the networks to simultaneously perform the event classification task and the onset time estimation. This approach obtained the third place in the final ranking. All of the aforementioned systems largely outperform the baseline system based on

ISBN 978-90-827970-1-5

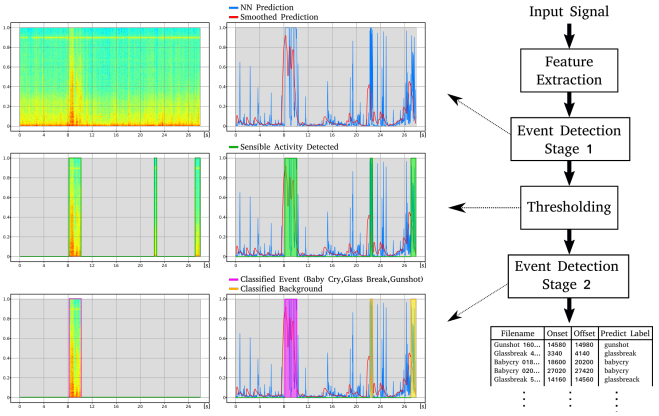

Fig. 1. Flow chart of the proposed method for rare sound event detection. Each event class implements such a scheme. In the first column are shown the spectrograms of the input signal and of the detected events. In the second column the network outputs at each stage of the algorithm.

a Multi Layer Perceptron architecture (MLP) and Logmel energies as features.

II. Proposed Method

The proposed system is a hierarchical algorithm composed of five stages: the acoustic features extraction (II-A), the event detection stage 1 (II-B) which produces an output at frame-rate and a dedicated smoothing procedure of this signal (II-C). Then, a refinement of the previous decision stage (II-D) is performed by a 2D CNN which discards possible false positives detected by the stage 1. The final decision procedure (II-E) annotates the effective onset time of the active event. In Fig.1 the phases of the algorithm are depicted. This is an extended and improved method with respect to our contribution to the DCASE 2017 [15].

A. Features Extraction

The feature extraction stage operates on mono audio signals sampled at 44.1 kHz. Following the results obtained at the DCASE2017 challenge by [13], we use the log mel energy coefficients (Logmel) as an efficient representation of the audio signal. In addition, we explored the combination of the Logmel with features based on wavelet coefficients and forward prediction errors (WC-LPE) [16]. A brief description of the features extraction procedures is given below.

1) Logmel coefficients: The audio signal is split into frames of 40 ms and a frame step of 20 ms, then the Logmel coefficients are obtained by filtering the power spectrogram of the frame by means of a mel filter-bank, then applying a logarithmic transformation to each sub-band energy in order to match the human perception of loudness. We used a filter bank with 40 mel scaled channels, obtaining 40 coefficients/frame.

2) WC-LPE Feature: The Wavelet Coefficient (WC) and Linear Prediction Error (LPE) feature set relies on non-stationary signal components and it has been successfully exploited for musical note onset detection [16]. WC-LPE extraction is done by first processing the input signal with a Discrete Wavelet Transform (DWT) dyadic tree. Then, each DWT sub-band is filtered by a linear prediction error filter (LPEF), obtaining Forward Prediction Errors (FPE). All LPEF outputs and DWT sub-bands are resampled to an intermediate sampling rate and rectified. The feature set is, finally, created from the DWT sub-bands, their first order time derivatives, the FPE and their first order time derivatives.

For both feature sets the range values of each coefficient is normalized independently according to the mean and the standard deviation computed on the training sets of the neural networks.

B. Event Detection Stage 1

The Event detection (ED) stage 1 has the goal to discard frames containing only background sounds, reducing as much as possible the false negative decisions. We evaluated two DNN architectures as binary classifiers: the MLP and the CNN with 2D kernels and frequency pooling. In both cases, the output layer is formed by two units with the *softmax* non-linear function. Thus, the networks outputs represent the probabilities that an input feature vector $\mathbf{x}[t]$ at the frame index t belongs to the background or the event class. In our analysis, we evaluated as network input the Logmel coefficients and the combination of the latter with the WC-LPE features.

1) Multi Layer Perceptron Neural Network: The artificial neuron is the main element of the MLP. It consists of an activation function applied to the sum of the weighted inputs [17]. Neurons are then arranged in layers, with feed-forward connections from one layer to the next. The supervised learning of the network makes use of the stochastic gradient descent with error back-propagation algorithm. The network is designed to consider a temporal context C, thus the network input feature vector $\hat{\mathbf{x}}[t]$ is obtained concatenating $\mathbf{x}[t]$ with the previous $\mathbf{x}[t-c]$, with $c = 1, \ldots, C$. During the training procedure, additive zero-centered Gaussian noise with $\sigma = 0.1$ was applied to $\hat{\mathbf{x}}[t]$ as a form of data augmentation, improving the generalization capabilities of the DNN and avoiding overfitting [11].

2) Convolutional Neural Network: CNNs are feed-forward neural networks [18] composed of three types of layers: convolutional layers, pooling layers and densely connected layers of neurons. The convolutional layers perform the mathematical operation of convolution between a multi-dimensional input and kernels of fixed size. The kernels are generally smaller compared to the input, allowing CNNs to process large inputs with a modest number of parameters to learn. CNNs are often used in audio tasks, where they exploit one input dimension to keep track of the temporal evolution of a signal [19]. In our case the convolutional layer input is a matrix $\mathbf{X} \in \mathbb{R}^{F \times T}$, where F and T represent respectively and the number of Logmel channels and the number of frames of the acoustic signal. When we combine the two aforementioned feature sets, we process them with two separate sets of convolutional layers, gathering two feature maps that are concatenated along the feature axis. Before concatenation, batch normalization [20] is applied to each feature map and a leaky rectified linear unit

ISBN 978-90-827970-1-5 © EURASIP 2018

activation function (LeakyReLU) with $\alpha = 0.3$, followed by a feature domain max-pooling layer. Finally, fully connected layers are stacked, applying the same weights and biases to each frame element. The output layer for each of the binary classifier neural networks has two neurons corresponding to the probability of the background or the event onset. We can discard, thus, one of the two neurons without loss of information, and we will consider the output of the neuron corresponding to the event activation $u[t] = y_{t,2}$, as the output of the network at frame t.

The neural networks training was accomplished by the AdaDelta stochastic gradient-based optimisation algorithm [21] for a maximum of 500 epochs on the binary cross entropy loss function. The optimizer hyperparameters were set according to [21] (i.e., initial learning rate $lr = 1.0$, $\rho = 0.95$, $\epsilon = 10^{-6}$). An early stopping strategy monitoring the validation loss was employed in order to reduce the computational burden. Thus if the validation loss does not decrease for 20 consecutive epochs, the training is stopped and the last saved model is selected as the final model. In addition, dropout is used as regularization technique [22] with rate 0.5.

C. Post Processing

In the post processing stage, each network output is convolved with an exponential decay window of length M defined as:

$$\mathbf{w}[t] = e^{\frac{t}{\tau}} \quad \text{with}\, \tau = \frac{-(M-1)}{log_e(0.01)} \tag{1}$$

The result is processed with a sliding median filter with local window-size k. Finally, a decision threshold θ is applied.

D. Event Detection Stage 2

The aim of the event detection stage 2 is to eliminate false positives, by removing the events wrongly detected at the previous stage. This is done by feeding a binary-classifier CNN with chunks of features in correspondence to the detected events (colored region in the bottom right spectrogram of Figure 1). At this stage only Logmel coefficients are used as input features, in order to reduce the computational burden of the model. Non-overlapping feature matrices $\mathbf{X}$ of size $F \times 20$ are used during training, while 95%-overlapping feature matrices are employed during testing (1-frame shift). A chunk size of 20 corresponds to 0.4 seconds of audio, i.e. half the minimum possible length of the occurring events, leading to an analysis of the audio event at different time and frequency resolutions with respect to previous stages. The ED Stage 2 NN is trained for 100 epochs on the binary cross entropy loss function with the AdaDelta gradient descent algorithm.

E. Final Decision

For each audio sequence, we perform a classification on contiguous blocks of frames detected as event by the ED stage 1. Among contiguous frame chunks classified as "event" by the CNN, the first frame with highest network output is indicated as event onset.

III. Dataset

The DCASE2017 challenge dataset [23] has been used to develop and evaluate the algorithm. The dataset consists of 30-second long sequences of background acoustic scenes recorded in different public or domestic spaces (park, home, street, cafe, train etc.) [24], some of which have been added with isolated recordings from at most one of the three different target sound event classes: baby crying, glass breaking and gun shot. The presence probability of a sound event in each mixed sequence of the original Development set was 0.5, thus we kept only sequences containing a sound event of the original training set and we generated additional mixtures assigned to the training and the validation sets. For the development set a total number of sequences respectively equal to 2750 for training, 300 for validation and 1496 for test have been employed. This change increases the percentage of the frames including a target event in the training data, which helps to ease the problem of data imbalance. In addition, due to the fast decay of the "gun shot" sound, we generated more sequences containing this event class compared to the others, in order to maintain approximately the same percentage between frames containing event samples and backgrounds.

In the evaluation set, the training and test sequences of the development set are combined into a single training set, while the validation set is the same used in the Development dataset. The system is evaluated against an "unseen" set of 1500 samples (500 for each target class) with a sound event presence probability for each class equal to 0.5.

IV. Experimental Set-Up

According to the DCASE 2017 guidelines, the performance of the proposed algorithm has been assessed by using the development dataset for training and validation of the system. Furthermore, a blind test on the provided evaluation dataset has been performed. The performance metric of the DCASE 2017 challenge is the event-based error rate (ER) calculated using onset-only condition with a collar of 500 ms. Detailed information on metrics calculation are available in [25]. The algorithm has been implemented in the Python language using Keras [26] as deep learning library. All the experiments were performed on a computer equipped with a 6-core Intel i7, 32 GB of RAM and two Nvidia Titan X graphic cards.

A. First Event Detection Stage

To assess the performance of the MLP employed in the event detection stage 1 we resorted to a random search

TABLE I
HYPER-PARAMETERS OPTIMIZED IN THE RANDOM-SEARCH PHASE FOR THE MLP ED STAGE 1, AND THEIR RANGE.

Parameter	Range	Distribution
MLP layers Nr.	[2 - 7]	uniform
MLP layers dim.	[20 - 4048]	log-unifom
MLP Context	[1 - 7]	uniform
Activation	[tanh - relu]	uniform

ISBN 978-90-827970-1-5 © EURASIP 2018

TABLE II
RESULTS IN TERMS OF ER SCORE FOR ALL THE EVALUATED COMBINATION OF PROPOSED ANNS AND FEATURES USED IN EVENT DETECTION STAGE 1.

	Development Dataset				Evaluation Dataset			
Features	Babycry	Glassbreak	Gunshot	**Average**	Babycry	Glassbreak	Gunshot	**Average**
MLP ED Stage 1								
Logmel	0.19	0.12	0.16	**0.16**	0.64	0.54	0.58	0.59
Logmel + WC-LPE	0.23	0.10	0.19	0.17	0.76	0.55	0.55	0.62
CNN ED Stage 1								
Logmel	0.23	0.13	0.18	0.18	0.48	0.23	0.44	0.38
Logmel + WC-LPE	0.25	0.09	0.16	0.17	0.46	0.10	0.36	**0.31**
MLP ED Stage 1 + CNN ED Stage 2								
Logmel	0.14	0.08	0.16	**0.13**	0.31	0.25	0.44	0.33
Logmel + WC-LPE	0.20	0.09	0.19	0.16	0.37	0.27	0.40	0.35
CNN ED Stage 1 + CNN ED Stage 2								
Logmel	0.19	0.10	0.16	0.15	0.31	0.17	0.39	0.29
Logmel + WC-LPE	0.18	0.08	0.17	0.14	0.25	0.10	0.31	**0.22**

strategy [27]. Table I shows the parameters explored in the random search, as well as the prior distribution and ranges. We evaluated 300 sets of layout parameters (100 for each event class) repeated for the two input features combination.

Regarding the CNNs, we explored the hyper-parameters space by means of a grid search for a total of 75 experiments (25 for each event class) covering the number of convolutional filters per layer $\{16, 32, 64\}$, the kernels shape $\{3 \times 3, 5 \times 5\}$, the number of MLP layers $\{1, 2, 3\}$ and their respective number of units $\{16, 32, 64, 128\}$. The feature max-pool sizes after each convolutional layer were $\{5, 4, 2\}$ for all the explored layouts. Also in this case the experiments were repeated for both the input features combination.

A successive grid search was performed for each network configuration evaluated, in order to find the post-processing parameters that yielded the minimum error rate. Investigated parameters in the grid search were: exponential window length w in the range $\{10, 20, \ldots, 90\}$, median filter kernel k in the range $\{9, 11, \ldots, 31\}$ and threshold θ in the range $\{0, 0.05, \ldots, 0.5\}$.

Once the best models on the Development dataset were found, a fine tuning of the post processing parameters was done during the validation stage, in order to assess the performance of the whole system. In fact, the hierarchical architecture of the algorithm allows to set a lower threshold in the first decision stage in order to reduce the deletions at the expenses of some insertions. These will be removed by the ED stage 2.

B. Refinement Stage

1) Training set for CNN based ED Stage 2 : To compose the dataset for training and evaluation of the CNNs dedicated to each target audio event we proceeded as follows: the samples of each event class were selected between the audio sections respectively labelled as "baby cry", "glass break" and "gun shot" from the mixtures of the DCASE 2017 development dataset, in addition with the isolated events source signals. To obtain the background samples, we processed with the first stage of our algorithm sequences from the same dataset which do not contain events. Thus, the frames detected as event in this case represent the "false positive" or "insertions" of the stage 1. We used those frames as background samples in the CNN training phase to improve its event classification abilities and balancing the dataset.

To design the best refinement CNN model for our purposes, we generated a shuffle stratified validation split from the dataset composed as described above. We left out the 30% of the samples as validation set for the CNN model and we selected the layout parameters of the neural network based on the F-measure score obtained on this data sub-set. The best performing model was the same for all the target audio events and was composed as follows: three convolutional layers with $\{32, 32, 32\}$ filters, respectively, of size 5×5. The convolutional layers were followed by a feature max pooling layer with kernels of size $\{5, 4, 2\}$, respectively. Three dense layers composed of 32 neurons with $tanh$ activation functions were applied before the network output layer, for a total number of network parameters equal to 35K.

V. RESULTS

Results reported in Table II are obtained as follows: we selected the models with lowest ER for each combination of DNN architecture and input features operating in the ED stage 1 and we evaluated the systems separately for each target class before the ED stage 2 on the Evaluation set, keeping ED stage 1 post processing parameters fixed. Then, with the same settings we obtained the performance of the whole system both on Development and Evaluation datasets. The architecture composed of a first stage with 2D CNN fed by Logmel and WC-LPE features resulted the best performing on the Evaluation dataset, obtaining an average ER equal to 0.17. Details of these architectures are reported in Table III.

TABLE III
DETAILS OF MODELS FOR CNN BASED ED STAGE 1 WITH THE LOWEST ER ON DEVELOPMENT SET. ALL OF THEM USE A COMBINATION OF LOG MEL ENERGIES AND WC-LPE AS INPUT FEATURES.

Hyper-parameters	Babycry	Glassbreak	Gunshot
Conv. Kernels	5×5, 3×3, 3×3	3×3, 3×3, 3×3	3×3, 3×3, 3×3
Kernel shape	32, 16, 16	64, 64, 64	32, 16, 16
MLP Layers size	32, 32	128, 128	32, 32
# Parameters	18K	185K	17K

ISBN 978-90-827970-1-5 © EURASIP 2018

TABLE IV
COMPARISON BETWEEN THE OBTAINED ER SCORES AND THE NUMBER OF PARAMETERS WITH THE FIRST THREE RANKED APPROACHES AT THE DCASE2017 CHALLENGE.

Approach	Evaluation ER	# Parameters
Lim et al. [12]	**0.13**	6200K
Cakir et al. [13]	0.17	756K
Proposed system	0.22	**108K**
Phan et al. [14]	0.27	2100K

The experimental results show how this combination improves generalization properties of the algorithm. In fact, the MLP based stage 1 with only Logmel features obtains the best overall ER equal to 0.13 on the Development dataset, but the performance decreases significantly on the Evaluation set. In addition, the number of free parameters of the best performing MLP models was always an order of magnitude greater w.r.t. the CNN models. Regarding the stage 2, its beneficial effect is supported especially with the Evaluation dataset: in this case, the improvement in terms of ER given by the joint detection procedure is evident and it gives additional robustness to the system in terms of generalization.

In Table IV the overall results between best ranked systems of the DCASE 2017 Challenge are compared. It can be observed that the best two scores have been obtained with ensemble methods, involving the additional computational cost of running several architectures in parallel, while the table reports the number of parameters per architecture. Although the proposed system does not outperform the first two methods, the average number of network parameters is significantly lower. This provides greater scalability in real-world applications.

VI. CONCLUSION

In this paper, a framework that makes use of hierarchical CNN classifiers fed with Logmel and WC-LPE features has been proposed for rare SED, providing significantly improved performance over the baseline system for every target sound event class in DCASE 2017 challenge dataset. The system also provides a significant reduction of the network parameters w.r.t. other competitive algorithms. The multi-scaled approach inherent to the two different CNN architectures results to be effective.

For future work, strategies to customize the loss function embedding the evaluation metric into the training procedure can be considered. Specifically, this task is particularly affected by the dataset unbalancing: to counteract this problem an alternative to the data augmentation is to design tailored loss functions which enhance the detection of the rare events.

REFERENCES

[1] C. Clavel, T. Ehrette, and G. Richard, "Events detection for an audio-based surveillance system," in *Proc. of ICME*, 2005, pp. 1306–1309.

[2] P. Foggia, N. Petkov, A. Saggese, N. Strisciuglio, and M. Vento, "Reliable detection of audio events in highly noisy environments," *Pattern Recognition Letters*, vol. 65, pp. 22–28, 2015.

[3] Y. Wang, L. Neves, and F. Metze, "Audio-based multimedia event detection using deep recurrent neural networks," in *Proc. of ICASSP*, 2016, pp. 2742–2746.

[4] S. Krstulović, "Audio event recognition in the smart home," in *Computational Analysis of Sound Scenes and Events*. Springer, 2018, pp. 335–371.

[5] D. Droghini, D. Ferretti, E. Principi, S. Squartini, and F. Piazza, "An end-to-end unsupervised approach employing convolutional neural network autoencoders for human fall detection," in *Proc. of WIRN*, Vietri sul Mare, Italy, June, 14-16 2017.

[6] E. Principi, D. Droghini, S. Squartini, P. Olivetti, and F. Piazza, "Acoustic cues from the floor: a new approach for fall classification," *Expert Systems with Applications*, vol. 60, pp. 51–61, 2016.

[7] T. Virtanen, A. Mesaros, T. Heittola, A. Diment, E. Vincent, E. Benetos, and B. M. Elizalde, *Proc. of DCASE*, 2017.

[8] T. Heittola, A. Mesaros, A. Eronen, and T. Virtanen, "Audio context recognition using audio event histograms," in *Proc. of EUSIPCO*, 2010, pp. 1272–1276.

[9] Y.-T. Peng, C.-Y. Lin, M.-T. Sun, and K.-C. Tsai, "Healthcare audio event classification using hidden Markov models and hierarchical hidden Markov models," in *Proc. of ICME*, 2009.

[10] S. Sigtia, A. M. Stark, S. Krstulović, and M. D. Plumbley, "Automatic environmental sound recognition: Performance versus computational cost," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, pp. 2096–2107, 2016.

[11] 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.

[12] H. Lim, J. Park, and Y. Han, "Rare sound event detection using 1D convolutional recurrent neural networks," in *Proc. of DCASE*, 2017, pp. 80–84.

[13] E. Cakir and T. Virtanen, "Convolutional recurrent neural networks for rare sound event detection," in *Proc. of DCASE*, 2017, pp. 27–31.

[14] H. Phan, M. Krawczyk-Becker, T. Gerkmann, and A. Mertins, "DNN and CNN with weighted and multi-task loss functions for audio event detection," *arXiv preprint arXiv:1708.03211*, 2017.

[15] F. Vesperini, D. Droghini, D. Ferretti, E. Principi, L. Gabrielli, S. Squartini, and F. Piazza, "A hierarchic multi-scaled approach for rare sound event detection," Department of Information Engineering, Università Politecnica delle Marche, Ancona, Italy, Tech. Rep., 2017, copyright-free.

[16] 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 *Proc. of ICASSP*, 2014, pp. 2183–2187.

[17] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, pp. 533–536, Oct. 1986.

[18] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[19] S. Thomas, S. Ganapathy, G. Saon, and H. Soltau, "Analyzing convolutional neural networks for speech activity detection in mismatched acoustic conditions," in *Proc. of ICASSP*, 2014, pp. 2519–2523.

[20] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. of ICMC*, 2015, pp. 448–456.

[21] M. D. Zeiler, "AdaDelta: an adaptive learning rate method," *arXiv preprint arXiv:1212.5701*, 2012.

[22] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[23] A. Mesaros, T. Heittola, A. Diment, B. Elizalde, A. Shah, E. Vincent, B. Raj, and T. Virtanen, "DCASE 2017 challenge setup: Tasks, datasets and baseline system," in *Proc. of DCASE*, 2017.

[24] A. Mesaros, T. Heittola, and T. Virtanen, "TUT database for acoustic scene classification and sound event detection," in *Proc. of EUSIPCO*, 2016, pp. 1128–1132.

[25] ——, "Metrics for polyphonic sound event detection," *Applied Sciences*, vol. 6, no. 6, p. 162, 2016.

[26] F. Chollet *et al.*, "Keras," https://github.com/keras-team/keras, 2015.

[27] J. Bergstra and Y. Bengio, "Random search for hyper-parameter optimization," *Journal of Machine Learning Research*, vol. 13, no. Feb, pp. 281–305, 2012.

ISBN 978-90-827970-1-5 © EURASIP 2018