American Scientific Research Journal for Engineering, Technology, and Sciences (ASRJETS)
ISSN (Print) 2313-4410, ISSN (Online) 2313-4402

http://asrjetsjournal.org/

An Improved Acoustic Scene Classification Method Using Convolutional Neural Networks (CNNs)

Khalid Hussain[a]*, Mazhar Hussain[b] , Muhammad Gufran Khan[c]

[a]*National University of Computer & Emerging Sciences, Department of Electrical Engineering, Pakistan*
[b]*National University of Computer & Emerging Sciences, Department of Computer Science, Pakistan*
[c]*National University of Computer & Emerging Sciences, Department of Electrical Engineering, Pakistan*
[a]*Email: khalid.hussain@nu.edu.pk*
[b]*Email: mazhar.h@nu.edu.pk*
[c]*Email: m.gufran@nu.edu.pk*

Abstract

Predicting acoustic environment by analyzing and classifying sound recording of the scene is an emerging research area. This paper presents and compares different acoustic scene classification (ASC) methods to differentiate between different acoustic environments. In particular, two deep learning techniques of classification i.e. Deep Neural Network (DNN) and Convolution Neural Network (CNN) have been applied using a combination of Mel-Frequency Cepstral Coefficients (MFCCs) and Log Mel energies as features. DNN and CNN are state-of-the-art techniques which are being used widely in speech recognition, computer vision, and natural language processing applications. These techniques have recently achieved great success in the field of audio classification for various applications. Both techniques have been implemented and tuned by performing a variety of experiments with different hyper parameters, hidden layers and units on public benchmark datasets provided in the DCASE 2017 challenge. The proposed method uses frame level randomization of the combined acoustic features i.e. MFCC and log mel energy, for training of model to achieve higher accuracy with DNN and CNN. It has reported higher accuracy than the previous work done on public benchmark datasets provided in the DCASE 2017 challenge. It is observed that DNN achieved 83.45% and CNN achieved 83.65% accuracy that is higher than the previous work done on public benchmark datasets provided in the DCASE 2017 challenge.

Keywords: Acoustic scene classification; deep neural networks; convolution neural network; mel energy; MFCC.

---

* Corresponding author.

1. Introduction

When we think about the acoustic scene classification, we refer to the abilities of human or artificial system that can perceive the con-text of audio. So, the objective of acoustic scene classification (ASC) is to categorize different audio environments to one of the pre-defined classes such as car, cafe, train, park etc. in which that audio was recorded. Smart devices can use this technology for contextualization and personalization [1] to fulfill the consumer requirements. It offers wide range of applications including context aware services [2], robotic navigation [3], surveillance [4], public place monitoring and assistance to enhance performance of audio event detection tasks [5]. It may also be used for security services and public place monitoring at common places. Another interesting application is the detection of water and gas leakage pipelines and reduces the human efforts of manual detection through a very long pipeline. Overview of the system is shown in figure 1. Although, several techniques have been proposed as a solution for the audio classification based on its different features but still ASC problem is a challenging task for the researchers to dig it out and improve the results. The goal of this paper is to use the state-of-the-art deep learning techniques to tackle the ASC problem. The deep learning techniques outperformed and offered tremendous results in many other applications [6]. So, with the enhanced academic interest and commercial demand of ASC, development of a practical system with high accuracy is required. Our proposed methods use the randomized data at frame level to improve the accuracy with the mel energy features using the DNN and CNN classification techniques. DCASE 2017 [7] dataset has been used that contains recording from several acoustic scenes from different locations. It contains 15 different acoustic scene recordings that need to be classified into the respective environment in which it was recorded [7]. Our system training is based on MFCC and mel energy features that are used as input for the deep learning techniques. These features are the representation of power spectrum of sound signal for very short span of time. The sound signal is broken into tiny frames of fixed length specified by the window which has a length of 40ms with 50% hop size. For feature extraction, librosa [8] a python library was used. The proposed methodology reported significant improvement in the accuracy for the DCASE 2017 challenge task1. The results are better as compared to the existing techniques trained with the several features and classifier. The remaining paper is divided into the sections as follows. In section 2 we illustrate the background work on ASC. In section 3 and 4 we discuss the DNN and CNN architectures. In section 5 we elaborate performance of the proposed solutions and comparison with existing models. Finally, conclusion and future research challenges will be discussed and these can be considered as future research challenges for the researcher's community.

2. Related Work

Acoustic Scene classification is a highly attractive area of research since 2013 from the first challenge of DCASE. Researchers applied different machine learning algorithms to classify the acoustic scenes and tried to improve the accuracy based on different features. Deep learning techniques for ASC used mel energy, mel frequency cepstral coefficients (MFCC) and various other features set to tackle the problem and improve the accuracy. Let's discuss the literature review of DNN first based on different features. Mafra and his colleagues [9] used mel log spectrogram as compact features with DNN, CNN and SVM. Xu and his colleagues [11] worked on hierarchical learning with the DNN by including taxonomy information in the learning environment and proposed two DNN based hierarchical technique to categorize the acoustic scenes. Patiyal and his

colleagues [12] used different mechanisms on different features and concluded that DNN perform better than the other techniques when trained on the same features. Kong and his colleagues [13] applied Gaussian mixture model and DNN on two types of features mel-filter bank with the same bank area and with the same height. It was reported that same height bank performs better as compared to the same area bank. Mun and his colleagues [14] proposed a bottleneck features using deep neural networks to improve results of audio classification. Now, discuss the literature review of CNN for the acoustic scene classification problem with different features. Mun and his colleagues [10] achieved top accuracy of 83.3% on the evaluation dataset for acoustic scene classification challenge 2017 using convolutional neural network based on log mel-energy and spectrogram features. Han and his colleagues [20] got the accuracy of 80.4% using CNN with log-mel energy features. Hertel and his colleagues [7] proposed CNN architecture with single label classification for ASC and and multi label classification on DCASE 2016 challenge for domestic audio tagging. Santoso and his colleagues [15] used MFCC features as input to the network-in-network CNN architecture to classify the audio scenes. Phan and his colleagues [17] presented acoustic classification based on label tree embedding (LTE) features using CNN and achieved promising results as compared with the baseline system for acoustic classification of DCASE dataset. Eghbal-Zadeh and his colleagues [18] proposed 4 techniques for ASC i.e. deep CNN which is based on spectrogram features, binaural I-vectors and late fusion of both CNN and I-vector to improve the overall accuracy of ASC. Kim and his colleagues [21] did the empirical study to ensemble the deep machine to improve performance on ASC. Bae and his colleagues [22] studied the parallel combination of long short-term memory (LSTM) and DNN and enhanced accuracy was reported. Application based on CNN getting more popularity with the passage of time. The related example to the ASC is music analysis [23], speech recognition [24], robust audio event recognition [25] and event detection [26]. In our proposed methodology, we are going to propose convolution neural networks and deep neural networks architectures on randomized training data to achieve more accurate results as com-pared to the other methods to recognize the acoustic scenes on the DCASE 2017 dataset.

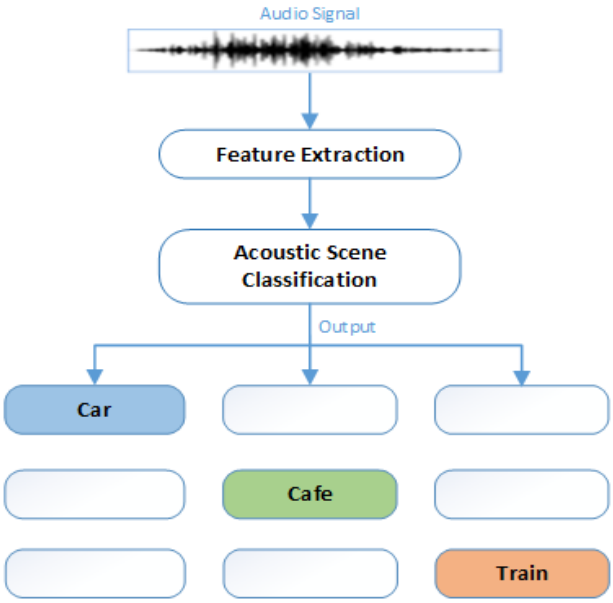

Figure 1: General overview of the acoustic scene classification system

3. DNN Architecture

DNN is a supervised learning feedforward artificial network used in various applications in image and video recognition, automatic speech recognition and it is trained for acoustic scene classification in this paper. It has different layers usually an input layer, several hidden layers to form a deep architecture and an output layer [11].The dataset used to train the network was taken from DCASE 2017 challenge and it consists of re-cording of different audio scenes. The DNN was trained with MFCC features and mel energy features. For the detail analysis, we used different layers with varying number of hidden layers and units in each layer. All results of analysis are mentioned in the next chapter of results. The final architecture of DNN consists of 4 hidden layers and 2 dropout layers. The first two hidden layers contain 512 neurons and last two contain 1024 neurons. All weights are initialized uniformly and optimized with adam optimizer. DNN was trained on 80 log-mel energy and 40 MFCC features for batch size of 256 with training epochs of 100. Softmax activation function was used to classify the different audio signals. For error function, categorical cross entropy was used to calculate the error for multi class prediction. DNN architecture is shown in Figure 2.

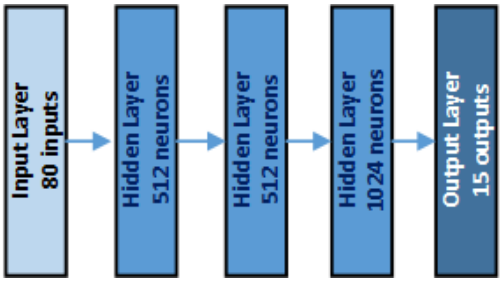

Figure 2: Deep Neural Networks Architecture

4. CNN Architecture

CNN consists of stack of distinct layers to classify the input into the outputs. Commonly used CNN Layers are convolution layer, Max pooling layer and fully connected layer. In convolution layer, filter is convolved with the input features. Max pooling do the job of down-sampling the input and fully connected layer connects all neuron from previous layer with its every neuron. This architecture used six convolution layers, three pooling layer and six regularization layer as shown in Figure 3. First and second Convolution layers have 32 feature maps and 3x3 receptive fields with the input shape of 1x10x12. Third and fourth convolution layers have a kernel size of 64 feature maps with 3x3 receptive fields. The last two convolution layers have a kernel size of 128 feature maps with 3x3 receptive fields. The max pooling layers of 2x2 were used to reduce the feature resolution. Pooling layer also reduce the invariance, dimensionality by down-sampling the input feature. Max pooling layer picks the single maximum value among the block of 2x2. Dropout layer was used as a regularization layer to avoid the overfitting by excluding 20% neurons randomly. Now the flatten layer was used to convert 2D matrix data into vector form. Its output will be processed by the standard fully connected layers. Tow hidden layers, one with 1024 neurons and one with 512 neurons were used to train the network in more sophisticated way with linear rectifier activation function. For output, softmax layer was used that give the probability of occurrence of each class out of 15 at the output. The input data is trained for 100 epochs with

batch size of 256 inputs. The learning rate for the training network was 0.01 and initialized normally. For gradient optimization, the adam optimizer was used.

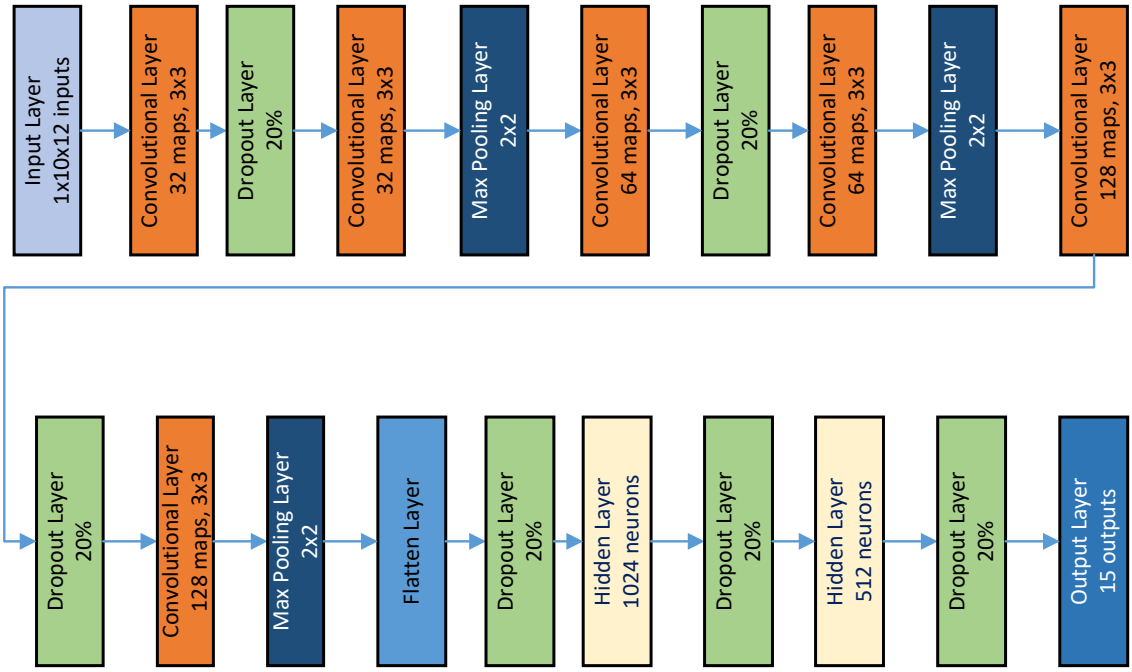

Figure 3: Convolutional Neural Networks Architecture

5. Results and Discussion

In this section, we evaluate the results of proposed architectures on the DCASE 2017 dataset to cope with the ASC problem. There are 4680 audio files in the development dataset. One audio file has 500 frames and MFCC and log-mel energy features are extracted from each frame. So, by randomizing the data on frame level the classifiers learned in a more challenging way. Data randomization with combined acoustic features enhanced the accuracy with DNN and CNN as compared to the existing results on ASC task. The proposed system results are outperformed on each individual class of the benchmark dataset that contains 15 classes of the acoustic scenes. Here, we presented the confusion matrix of the proposed DNN and CNN with the percentage accuracies of each class as shown in Figure 4 and 5 respectively.

Here, we examine the proposed results of DNN and CNN based on MFCC and log mel energy features with the results obtained in the past work based on different features on the DCASE dataset for classification of recorded audio. Different techniques had been proposed with DNN and CNN and achieved promising results are shown in results comparison Table 1 and 2. If we compare and analyze the results on evaluation data set as shown in Table 1 and Table 2 for DNN and CNN then It can be concluded that the results achieved by our DNN and CNN architectures on randomized data are better than the previous techniques as mentioned in the Tables.

	beach	bus	café	car	city_c	forest_p	grocery_s	home	library	metro_s	office	park	resid_area	train	tram
beach	91%	0%	2%	0%	0%	0%	1%	0%	2%	0%	0%	1%	2%	0%	0%
bus	11%	71%	3%	4%	0%	0%	4%	0%	0%	0%	0%	0%	0%	1%	6%
café_restaurant	10%	0%	83%	0%	0%	0%	2%	0%	1%	1%	0%	0%	1%	0%	1%
car	6%	1%	0%	87%	0%	0%	0%	0%	0%	0%	0%	0%	0%	1%	3%
city_center	5%	1%	0%	0%	90%	0%	0%	0%	0%	0%	0%	1%	3%	0%	0%
forest_path	9%	0%	0%	1%	0%	76%	0%	1%	1%	0%	1%	2%	11%	0%	0%
grocery_store	9%	0%	6%	0%	0%	0%	80%	0%	0%	0%	0%	0%	0%	0%	1%
home	7%	0%	0%	0%	0%	0%	0%	82%	1%	0%	8%	1%	0%	0%	0%
library	13%	0%	1%	0%	0%	1%	0%	4%	75%	1%	4%	0%	1%	0%	0%
metro_station	2%	0%	1%	0%	0%	0%	0%	0%	2%	94%	0%	0%	0%	0%	0%
office	7%	0%	0%	0%	0%	1%	0%	3%	2%	0%	86%	0%	0%	0%	0%
park	9%	0%	0%	0%	2%	0%	1%	0%	1%	0%	0%	83%	5%	0%	0%
residential_area	9%	0%	1%	0%	3%	2%	1%	0%	0%	0%	0%	2%	82%	0%	0%
train	4%	2%	0%	2%	0%	0%	0%	0%	0%	0%	0%	0%	0%	90%	1%
tram	9%	5%	1%	1%	0%	0%	1%	0%	0%	0%	0%	0%	0%	1%	82%

Figure 4: Confusion matrix for the proposed DNN with class-wise accuracy.

	beach	bus	café	car	city_c	forest_	grocery_s	home	library	metro_s	office	park	resid_ar	train	tram
beach	90%	0%	1%	0%	1%	0%	1%	1%	1%	0%	0%	2%	3%	1%	0%
bus	10%	77%	1%	3%	0%	0%	1%	0%	0%	0%	0%	0%	0%	1%	6%
café_restaurant	14%	0%	73%	0%	0%	1%	4%	2%	0%	2%	0%	0%	1%	1%	1%
car	6%	1%	0%	77%	0%	0%	0%	0%	0%	0%	0%	0%	0%	2%	12%
city_center	4%	0%	0%	0%	91%	0%	2%	0%	0%	0%	0%	1%	1%	0%	0%
forest_path	3%	0%	1%	1%	1%	89%	0%	0%	1%	1%	1%	2%	1%	0%	0%
grocery_store	9%	0%	2%	0%	0%	1%	85%	2%	0%	0%	0%	0%	0%	0%	1%
home	5%	0%	0%	0%	0%	2%	0%	88%	1%	0%	2%	1%	1%	0%	0%
library	12%	0%	0%	0%	0%	3%	0%	5%	73%	0%	4%	1%	0%	0%	0%
metro_station	2%	1%	1%	0%	1%	0%	0%	0%	1%	89%	0%	2%	0%	2%	0%
office	7%	0%	0%	0%	0%	2%	0%	6%	1%	0%	83%	0%	0%	0%	0%
park	7%	0%	0%	0%	1%	2%	0%	1%	0%	0%	0%	87%	3%	0%	0%
residential_area	10%	0%	0%	0%	4%	5%	0%	1%	0%	0%	0%	4%	75%	0%	0%
train	3%	3%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	91%	3%
tram	7%	2%	1%	0%	0%	0%	1%	0%	0%	0%	0%	0%	0%	2%	87%
									Over All Accuracy					83.65%	

Figure 5: Confusion matrix for the proposed CNN with class-wise accuracy.

Table 1: CNN Results Comparison

Classifier	**Features**	**Accuracy**
Proposed CNN	MFCC, log mel energy	83.65
CNN [20]	log mel energy	83.3%
CNN ensemble [21]	log mel energy	80.4%
CNN [19]	log mel energy	80.3%
CNN [16]	log mel energy	79.9%
CNN [18]	log mel energy	79.6%
CNN [17]	log mel energy	77.7%
CNN [16]	log mel energy	74.8%
CNN [15]	log mel energy	74.1%
CNN [1]	log mel energy	73.8%
CNN [7]	log mel energy	72.6%

Table 2: DNN Results Comparison

Classifier	Features	Accuracy
Proposed DNN	MFCC, log mel energy	83.45%
DNN-GMM [10]	MFFC	85.6%
DNN [14]	various	82.3%
DNN [13]	mel energy	81.0%
DNN [12]	MFCC	78.5%
DNN [11]	mel energy	73.3%
DNN [9]	mel energy	73.1%

6. Conclusions

In this paper, we implemented acoustic scene classification with convolutional neural networks (CNN) and deep neural networks (DNN). Also, it proposed frame level randomization on bench-mark dataset to enhance the accuracy further with DNN and CNN on mel energy features and MFCC. It has subtracted the background noise using median filter before the features are fed into the training network. It is concluded that the proposed DNN and CNN results on acoustic scene classification are better than the baseline system and past work done. The reported accuracy of true classification of all 1620 files with DNN and CNN was 83.45%, 83.65% respectively.

References

[1] Battaglino, Daniele, Ludovick Lepauloux, Nicholas Evans, France Mougins, and France Biot. Acoustic scene classification using convolutional neural networks. DCASE2016 Challenge, Tech. Rep, 2016.

[2] A. J. Eronenet al., “Audio-based context recognition,”IEEE Trans. Audio, Speech, Lang. Process., vol. 14, no. 1, pp. 321–329, Jan. 2006.

[3] S. Chu, S. Narayanan, C.-C. J. Kuo, and M. J. Mataric, “Where am I? Scene recognition for mobile robots using audio features,” inProc. IEEE Int. Conf. Multimedia Expo, 2006, pp. 885–888.

[4] R. Radhakrishnan, A. Divakaran, and P. Smaragdis, “Audio analysis for surveillance applications,” inProc. IEEE Workshop Appl. Signal Process. Audio Acoust.,2005, pp. 158–161.

[5] T. Heittola, A. Mesaros, A. Eronen, and T. Virtanen, “Con-text-dependent sound event detection,” EURASIP J. Audio, Speech, Music Process., vol. 2013, 2013, Art. no. 1.

[6] J. Schmidhuber, “Deep learning in neural networks: An overview,” CoRR, vol. abs/1404.7828, 2014.

[7] Hertel, Lars, Huy Phan, and Alfred Mertins. "Classifying variable-length audio files with all-convolutional networks and masked global pooling." arXiv preprint arXiv:1607.02857 (2016).

[8] B. McFee, C. Raffel, D. Liang, D. P. Ellis, M. McVicar, E. Battenberg, and O. Nieto, “librosa: Audio and music signal analysis in python,” in Proceedings of the 14th Python in Sci-ence Conference, 2015.

[9] Mafra, Gustavo, Ngoc Duong, Alexey Ozerov, and Patrick Pérez. "Acoustic scene classification: An evaluation of an ex-tremely compact feature representation." In Detection and Clas-sification of Acoustic Scenes and Events 2016. 2016.

[10] Takahashi, Gen, Takeshi Yamada, Shoji Makino, and Nobutaka Ono. "Acoustic scene classification using deep neural network and frame-concatenated acoustic feature." Detection and Classification of Acoustic Scenes and Events (2016).

[11] Xu, Yong, Qiang Huang, Wenwu Wang, and Mark D. Plumbley. "Hierarchical learning for DNN-based acoustic scene classification." arXiv preprint arXiv:1607.03682 (2016).

[12] Patiyal, Rohit, and Padmanabhan Rajan. "Acoustic Scene Classification Using Deep Learning."

[13] Kong, Qiuqiang, Iwnoa Sobieraj, Wenwu Wang, and Mark D. Plumbley. "Deep neural network baseline for DCASE chal-lenge 2016." Proceedings of DCASE 2016 (2016).

[14] Mun, Seongkyu, Sangwook Park, Younglo Lee, and Hanseok Ko. Deep Neural Network Bottleneck Feature for Acoustic Scene Classification. DCASE2016 challenge technical report, 2016.

[15] Santoso, Andri, Chien-Yao Wang, and Jia-Ching Wang. "Acoustic Scene Classification Using Network-In-Network Based Convolutional Neural Network."

[16] Lidy, Thomas, and Alexander Schindler. "CQT-based con-volutional neural networks for audio scene classification and domestic audio tagging." IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Events (DCASE 2016), Budapest, Hungary, Tech. Rep (2016).

[17] Phan, Huy, Lars Hertel, Marco Maass, Philipp Koch, and Alfred Mertins. "CNN-LTE: a Class of 1-X Pooling Convolu-tional Neural Networks on Label Tree Embeddings for Audio Scene Recognition." arXiv preprint arXiv:1607.02303 (2016).

[18] Eghbal-Zadeh, Hamid, et al. "CP-JKU submissions for DCASE-2016: A hybrid approach using binaural i-vectors and deep convolutional neural networks." IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Events (DCASE) (2016).

[19] Han, Yoonchang, and Kyogu Lee. Convolutional neural network with multiple-width frequency-delta data augmentation for acoustic scene classification. DCASE2016 Challenge, Tech. Rep, 2016.

[20] Valenti, Michele, Aleksandr Diment, Giambattista Parascandolo, Stefano Squartini, and Tuomas Virtanen. "DCASE 2016 acoustic scene classification using convolutional neural networks." In Proc. Workshop Detection Classif. Acoust. Scenes Events, pp. 95-99. 2016.

[21] Kim, Jaehun, and Kyogu Lee. "Empirical study on ensemble method of deep neural networks for acoustic scene classification." Proc. of IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Events (DCASE) (2016).

[22] Bae, S.H., Choi, I. and Kim, N.S., 2016, September. Acoustic scene classification using parallel combination of LSTM and CNN. In Proceedings of the Detection and Classification of Acoustic Scenes and Events 2016 Workshop (DCASE2016).

[23] J. Schlter and S. Bck, “Improved musical onset detection with convolutional neural networks,” in 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), May 2014, pp. 6979–6983.

[24] O. Abdel-Hamid, A.-r. Mohamed, H. Jiang, and G. Penn, “Applying convolutional neural networks concepts to hybrid NNHMM model for speech recognition,” in 2012 IEEE international conference on Acoustics, speech and signal processing (ICASSP). IEEE, 2012, pp. 4277–4280.

[25] H. Phan, L. Hertel, M. Maass, and A. Mertins, “Robust audio event recognition with 1-max pooling convolutional neural networks,” arXiv preprint arXiv:1604.06338, 2016.

[26] K. J. Piczak, “Environmental sound classification with convolutional neural networks,” in 2015 IEEE 25th International Workshop on Machine Learning for Signal Processing (MLSP), Sept 2015, pp. 1–6

[27] Bisot, Victor, Romain Serizel, Slim Essid, and Gael Richard. "Supervised nonnegative matrix factorization for acoustic scene classification." IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Events (DCASE) (2016).

[28] Park, Sangwook, Seongkyu Mun, Younglo Lee, and Hanseok Ko. "Score fusion of classification systems for acoustic scene classification." IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Events (DCASE) (2016).

[29] Marchi, Erik, Dario Tonelli, Xinzhou Xu, Fabien Ringeval, Jun Deng, and B. Schuller. "The up system for the 2016 DCASE challenge using deep recurrent neural network and multiscale kernel subspace learning." IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Events (DCASE) (2016).