Improved Acoustic Scene Classification with DNN and CNN

Khalid Hussain[1], Mazhar Hussain[2] and Muhammad Gufran Khan[1]

[1]National University of Computer & Emerging Sciences, Department of Electrical Engineering, Pakistan

[2]National University of Computer & Emerging Sciences, Department of Computer Science, Pakistan
khalid.hussain@nu.edu.pk, mazhar.h@nu.edu.pk, m.gufran@nu.edu.pk

ABSTRACT

This paper presents the acoustic scene classification (ASC) to differentiate between different acoustic environments corresponding to the DCASE 2017 challenge task1. In this contribution we have applied two techniques of classification i.e. Deep Neural Network (DNN) and Convolution Neural Network (CNN). DNN and CNN are widely used in speech recognition, computer vision, and natural language processing applications. These techniques have recently achieved great success in the field of audio classification for the various applications. We achieved higher accuracy than the previous work done on benchmark datasets provided in the DCASE 2016 challenge. We used frame level randomization of the training dataset and log mel energy features to achieve higher accuracy with DNN and CNN. It is observed that DNN achieved 90.41%, 90.03% and CNN achieved 90.71%, 88.86% accuracy on randomized data based on 80 and 60 mel energy features, respectively.

Index Terms—Acoustic scene classification, DCASE 2017, deep neural networks, convolution neural network, mel energy

1. INTRODUCTION

The objective of acoustic scene classification (ASC) is to categorize different audio environments to one of the pre-defined classes in which it was recorded. Examples include car, cafe, train, urban park etc. 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 assist to enhance performance of audio event detection tasks [5]. Overview of the system is shown in the 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 potential challenge for the researchers to dig it out and improve the results.

The goal of this paper is to use the well know deep learning techniques to tackle the ASC problem. The deep learning techniques [6] outperformed and offered tremendous results in many other applications. So, with the enhanced academic and commercial demand of ASC and deep learning, everyone eager to use applications based on it. Our proposed methods use the randomized data to enhance the results on the benchmark dataset based on the mel energy features using the DNN and CNN classification techniques.

DCASE 2017 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 is based on mel energy features that are used as input for the deep learning techniques. Log mel-band energy 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% hope 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 consider as future research challenges for the researcher's community.

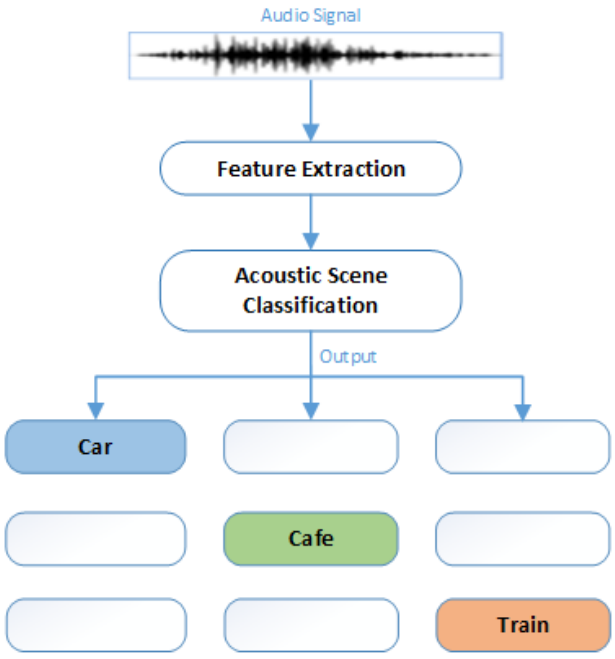

Figure 1: General overview of the acoustic scene classification system

2. RELATED WORK

Researchers applied different machine learning algorithms to classify the DCASE dataset and tried their best to improve the accuracy based on different features. In the literature review, deep learning techniques to ASC utilized mel energy, mel frequency cepstral coefficients (MFCC) and various other features set to tackle the problem. Let's discuss the literature review of DNN first based on different features. Mafra et al. [9] used mel log spectrogram as compact features with DNN, CNN and SVM. Takahashi et al. [10] achieved 85.6% accuracy on the evaluation dataset with deep neural network Gaussian mixture model (DNN-GMM) using MFFC features vector as input. Xu et al. [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 et al. [12] used different mechanisms on different features and concluded that DNN perform better than the other techniques when trained on the same features. Kong et al. [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 et al. [14] proposed a bottleneck features using deep neural networks to improve results of audio classification. The promising accuracy about 82.3% on the evaluation dataset for acoustic scene classification was observed based on various features set. Now, discuss the literature review of CNN for the acoustic scene classification problem with different features. Hertel et al. [7] proposed CNN architecture with single label classification for ASC and and multi label classification on DCASE 2016 challenge for domestic audio tagging. Santoso et al. [15] used MFCC features as input to the network-in-network CNN architecture to classify the audio scenes. Schindler et al. [16] enhanced the results by using constant-Q-transformed (CQT) features as input to the CNN and achived 81.8% accuracy on the evaluation dataset. They worked on both domestic audio tagging and ASC. Phan et al. [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 et al. [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. Lee et al. [19] used multiple width frequency-delta data augmentation and showed the accuracy of 84.6% on the evaluation dataset. Valenti et al. [20] work exhibited 86.2% accuracy on the DCASE 2016 evaluation dataset using CNN based on log mel spectrogram. Kim et al. [21] did the empirical study to ensemble the deep machine to improve performance on ASC. Bae et al. [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 are 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 data to achieve more accurate results as compared to the other methods to recognize the acoustic scenes on the DCASE 2017 dataset.

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.

We implemented two deep neural network architectures of DNN that were trained with 80 and 60 mel energy features. For the training of DNN, we used 3 hidden layers with rectifier linear activation. First two layers have 512 neurons while third layer has 1024 number of neurons. All weights are initialized uniformly and optimized with adam optimizer. DNN was trained on 80 log-mel energy features for batch size of 256 with training epochs of 200. 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.

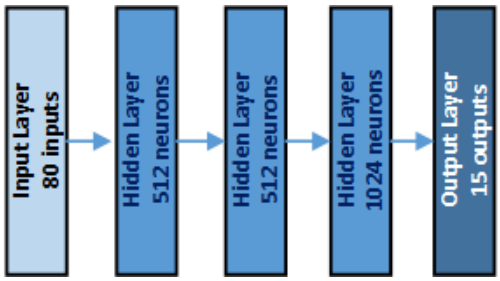

Figure 2: Deep Neural Networks Architecture for 80 mel energy features.

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.

We proposed two different architectures for CNN that are based on different number of features. CNN was trained on mel energy features. We used two convolution layers following with pooling and regularization layer as shown in figure 3. First, Convolution layer has 64 feature maps and 3x3 receptive fields with the input shape of 1x10x8. Second convolution layer has a kernel size of 128 feature map with 3x3 receptive fields. After the convolution layers max pooling layer of 2x2 was applied 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 22% neurons randomly. After the regularization layer, flatten layer was used to convert 2D matrix data into

vector form. Its output will be processed by the standard fully connected layers. Three hidden layers of 1000 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 50 epochs with batch size of 128 inputs. The learning rate for the training network was 0.001 and initialized normally. For gradient optimization, the adam optimizer was used.

CNN for 60 mel energy features has input shape of 1x10x6 and developed using 2 successive convolution layer with 32 and 128 kernel size and 3x3 and 2x2 filter size respectively. The hidden units in each dense layer were 1000 and rest of the system parameters were same as was in 80 mel energy features

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 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 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.

	be	bu	ca	ca	ci	fo	gr	ho	li	me	of	pa	re	tr	tr
beach	95%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	1%	1%	0%	1%
bus	3%	88%	1%	1%	0%	0%	0%	0%	0%	0%	0%	0%	0%	2%	4%
cafe_restaurant	4%	0%	88%	0%	1%	0%	3%	0%	0%	1%	0%	0%	0%	1%	1%
car	1%	0%	0%	96%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	3%
city_center	2%	0%	1%	0%	91%	0%	0%	0%	0%	0%	0%	0%	4%	0%	0%
forest_path	2%	0%	0%	0%	0%	94%	0%	1%	0%	0%	1%	1%	1%	0%	0%
grocery_store	3%	0%	3%	0%	0%	0%	90%	0%	0%	2%	0%	0%	0%	0%	0%
home	4%	0%	1%	0%	0%	1%	1%	86%	4%	0%	2%	0%	0%	0%	0%
library	2%	0%	0%	0%	0%	0%	0%	4%	89%	0%	1%	0%	1%	0%	0%
metro_station	0%	0%	0%	0%	0%	0%	1%	0%	0%	98%	0%	0%	0%	0%	0%
office	1%	0%	0%	0%	0%	0%	0%	1%	1%	0%	95%	1%	0%	0%	0%
park	3%	0%	1%	0%	0%	1%	0%	0%	0%	0%	1%	90%	3%	0%	0%
residential_area	3%	0%	1%	0%	2%	1%	0%	0%	0%	0%	0%	3%	88%	0%	0%
train	5%	2%	2%	0%	1%	0%	0%	0%	0%	0%	0%	0%	1%	84%	5%
tram	5%	3%	1%	5%	0%	0%	0%	0%	0%	0%	0%	1%	0%	3%	82%

Overall Accuracy **90.41%**

Figure 4: Confusion matrix for the proposed DNN (80 features) with class-wise accuracy.

	be	bu	ca	ca	ci	fo	gr	ho	li	me	of	pa	re	tr	tr
beach	96%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	2%	0%	0%
bus	9%	84%	1%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	3%	2%
cafe_restaurant	4%	0%	89%	0%	1%	0%	3%	0%	0%	0%	0%	0%	0%	1%	0%
car	3%	1%	0%	90%	0%	0%	0%	0%	0%	0%	0%	0%	0%	1%	5%
city_center	3%	0%	1%	0%	89%	0%	0%	0%	0%	0%	0%	0%	6%	0%	0%
forest_path	2%	0%	0%	0%	0%	95%	0%	0%	0%	0%	0%	1%	1%	0%	0%
grocery_store	4%	0%	3%	0%	0%	0%	91%	0%	0%	0%	0%	0%	0%	0%	0%
home	6%	0%	1%	0%	0%	2%	1%	84%	4%	0%	1%	0%	0%	0%	0%
library	4%	0%	0%	0%	0%	1%	1%	4%	89%	0%	1%	0%	1%	0%	0%
metro_station	2%	0%	1%	0%	0%	0%	1%	0%	0%	95%	0%	0%	0%	0%	0%
office	4%	0%	0%	0%	0%	1%	0%	1%	2%	0%	91%	0%	0%	0%	0%
park	4%	0%	0%	0%	0%	1%	0%	0%	0%	0%	1%	89%	4%	0%	0%
residential_area	4%	0%	0%	0%	2%	1%	0%	0%	0%	0%	0%	3%	89%	0%	0%
train	5%	1%	2%	0%	0%	0%	1%	0%	0%	0%	0%	0%	1%	88%	2%
tram	9%	2%	1%	2%	0%	0%	1%	0%	0%	0%	0%	1%	0%	5%	78%

Overall Accuracy **90.71%**

Figure 5: Confusion matrix for the proposed CNN (80 features) with class-wise accuracy.

Here, we examine the proposed results of DNN and CNN based on 80 and 60 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 as 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 below.

Table 1
CNN Results Comparison Table

Classifier	***Features***	***Accuracy***
Proposed CNN	80 mel energy	90.71%
Proposed CNN	60 mel energy	88.86
CNN [20]	mel energy	86.2%
CNN ensemble [21]	Unsupervised	85.4%
CNN [19]	mel energy	84.6%
CNN [16]	CQT	83.3%
CNN [18]	spectrogram	83.3%
CNN [17]	label tree embeding	83.3%
CNN [16]	CQT	81.8%
CNN [15]	MFCC	80.8%
CNN [1]	mel energy	80.0%
CNN [7]	spectrogram	79.5%

Table 2
DNN Results Comparison Table

Classifier	*Features*	*Accuracy*
Proposed DNN	80 mel energy	90.41%
Proposed DNN	60 mel energy	90.03%
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%

Now, we compare our proposed DNN results with different classifiers results as shown in Table3. Eghbal-Zadeh [15] achieved highest accuracy on Evaluation DCASE dataset by using late-fusion classifier based on MFFC and spectrogram features till 2016. The results show that our CNN achieved the 1^{st} , 3^{rd} and DNN achieved the 2^{nd} , 4^{th} place among all the techniques and improved the overall as well as class-wise performance on DCASE 2017 challenge task1 of acoustic scenes classification.

Table 3
Different Classifiers Results Comparison Table

Classifier	*Features*	*Accuracy*
Proposed CNN	80 mel energy	90.71%
Proposed DNN	80 mel energy	90.41%
Proposed DNN	60 mel energy	90.03%
Fusion [15]	MFFC + Spectrogram	89.7%
Proposed CNN	60 mel energy	88.86%
I-vector [15]	MFFC	88.7%
NMF [27]	spectrogram	87.7%
Fusion [28]	various	87.2%
Fusion [29]	various	86.4%

6. CONCLUSION

In this paper, we illustrated acoustic scene classification with convolutional neural networks (CNN) and deep neural networks (DNN). Also, we proposed frame level randomization on benchmark dataset to enhance the accuracy further with DNN and CNN on mel energy features. It was concluded that the proposed DNN and CNN results on acoustic scene classification are outperformed the baseline system and past work done. This is the first time competitive results are reported on benchmark datasets provided from the detection and classification of acoustic scenes and events (DCASE) challenge 2017. We improved the state-of-the-art results of deep neural networks and convolution neural networks. We obtained 90.41%, 90.03% accuracy with DNN and 90.71%, 88.86% accuracy with CNN based on 80 and 60 mel energy features respectively for ASC. In future, we can extend our research to enhance the accuracy of DCASE Challenges based on spectrogram, MFCC, CQT features using different classification mechanisms to categorize the different audio environments.

7. 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, “Context-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 convolutional 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).