ACOUSTIC SCENE CLASSIFICATION USING DEEP NEURAL NETWORK

P. Chandrasekhar[*]

Speech Processing Lab
International Institute of Information Technology
Hyderabad, India
chandrasekhar.p@research.iiit.ac.in

Suryakanth V. Gangashetty[†]

Speech Processing Lab
International Institute of Information Technology
Hyderabad, India
svg@iiit.ac.in

ABSTRACT

In this paper, deep neural networks (DNN) are applied for acoustic scene classification task provided by DCASE2017 challenge. We perform experiment on a dataset consisting of 15 types of acoustic scenes with a given total development data and evolution data of task1. We propose an DNN architecture for utterance level classification. Evaluation of models were performed on given evolution data of task1 for 4 folds using development data. In this approach MFCC and IMFCC feature vectors are used to train DNN model and their DNN scores were combined to test the system. On the official development data set of the task1 challenge, an accuracy of 81.28% is achieved.

Index Terms— Mel-Frequency Cepstral Coefficients (MFCC) , Inverse Mel-Frequency Cepstral Coefficients (IMFCC) , Multi Layer Perceptron (MLP) model, Deep Neural Network (DNN) and Acoustic Scene Classification (ASC) .

1. INTRODUCTION

The aim of acoustic scene classification is to classify a test recording into one of predefined classes that characterizes the environment in which it was recorded for example "park", "street", "office" and etc. The acoustic data will include recordings from 15 contexts, approximately one hour of data from each context [1], understanding the perceptual processes driving the human ability to categorize and recognize sounds and soundscapes [5]. The dataset consists of recordings from various acoustic scenes, all having distinct recording locations. For each recording location, 3-5 minute long audio recording was captured. The original recordings were then split into segments with a length of 10 seconds. These audio segments are provided in individual files. This contribution describes our investigated method for ASC. From the recordings,Mel-Frequency Cepstral Coefficients (MFCC) and Inverse Mel-Frequency Cepstral Coefficients (IMFCC) features are extracted and these feature are feed to DNN to train and test the system. Mel-Frequency Cepstral Coefficients (MFCC) presents a way to convert a physically measured spectrum of speech in to a perceptually meaningful subjective spectrum based on the human auditory system.Here we propose Inverse Mel-Frequency Cepstral Coefficients (IMFCC) defined by a competing filter bank structure which is indicative of a hypothetical auditory system which has followed a diametrically opposite path of evolution than the human auditory system.The idea is to capture those information which otherwise could have been missed by original Mel-Frequency Cepstral Coefficients (MFCC),due to the inverted filter bank structure ,the Inverse Mel-Frequency Cepstral Coefficients (IMFCC) will be able to represent the high frequency range more finely [6].

2. SYSTEM DESCRIPTION

This system is built on the top of python DCASE2017 of task1 baseline distributed by the organizers and described [2] .Feature extraction is done with librosa package and neural network modelling on Keras library, In this approach the feature extraction and DNN modelling done by MATLAB.Then the remainder of this section describes the core parts of system.

2.1. Recording and annotation procedure

For all acoustic scenes, the recordings were captured each in a different location: different streets, different parks, different homes. Recordings were made using a Soundman OKM II Klassik/studio A3, electret binaural microphone and a Roland Edirol R-09 wave recorder using 44.1 kHz sampling rate and 24 bit resolution. The microphones are specifically made to look like headphones, being worn in the ears. As an effect of this, the recorded audio is very similar to the sound that reaches the human auditory system of the person wearing the equipment [1].

2.2. Feature extraction

The baseline mel band energy acoustic features are replaced by MFCC and IMFCC are extracted for 40 filter banks with 40 ms frames with 20 ms overlap. Even for stacked context of features are used Mel-Frequency Cepstral Coefficients (MFCC)(13) and delta Mel-Frequency Cepstral Coefficients (MFCC)(13) and double delta Mel-Frequency Cepstral Coefficients (MFCC)(13) a total of 39 dimension feature vectors are extracted per frame and similarly for Inverse Mel-Frequency Cepstral Coefficients (IMFCC) [6] feature extraction and these delta and double delta features are captures dynamic acoustic scene information or contextual acoustic scene information.

3. NEURAL NETWORK ARCHITECTURE

The DNN used in this experiment is a fully connected feed forward neural network. It consists of a input layer, 3 hidden layers and output layer. Input layer has 39 neurons with linear activation function,each hidden layer has 200 neurons with RelU activation function and output layer has 15 neurons with softmax activation.

[*]
[†]

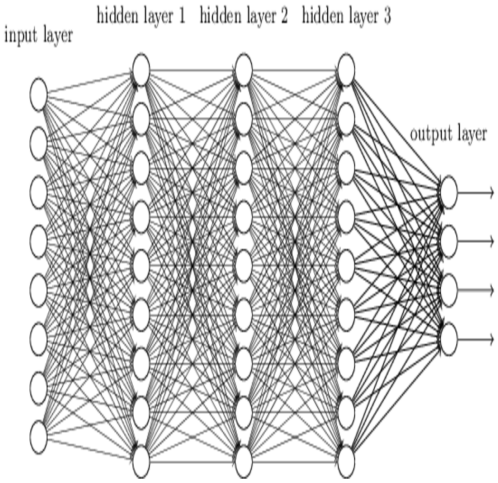

Figure 1: DNN used for Task1

ADAM [7] optimizer is used for weight optimization since it is generally faster than SGDCM [8] and the DNN structure is shown in Figure. 1.

4. EXPERIMENTAL RESULTS

In this section we evaluate the performance fusion of MFCC and IMFCC using DNN on DCASE 2017 challenge Task1 on ASC.We use 39(13 static MFCC+13 delta MFCC +13 double delta MFCC) and similarly for IMFCC feature vector per frame with frame duration 40 ms with 20 ms overlap. Then we apply these 39 dimension feature vectors are feed to the DNN to train the system as shown in Figure1 ,utterance by utterance level [10],after training DNN scores were generaed for two models (MFCC and IMFCC) , a weighted sum rule [11] adopted to fuse DNN scores, where lambda=0.7 is the weight that we have chosen for combining the DNN scores during test phase to get enhanced performence and these systems are implemented in MATLAB 2017a.

4.1. Task1:Acoustic Scene Classification

DCASE2017 dataset of task1 is used in this experiment. The data set consists of recordings from various acoustic scene , all having different recording locations, there are 15 classes with 4 fold cross validation,for training DNN[3] the batch size is set to 1000.ADAM learning rate is set to 0.001. The maximum number of epochs is set to 10. Then the results are shown on Table. 1.

5. CONCLUSION

In this paper, we have applied the DNN structure for task1 of DCASE2017 challenge.In summery fusion of Mel-Frequency Cepstral Coefficients (MFCC) and Inverse Mel-Frequency Cepstral Coefficients (IMFCC) with DNN accuracy 81.28% is better than baseline MLP based system accuracy 75% on fold1 using development data and average accuracy is also improved.

Table 1: Accuracy of Task1

	baseline	MFCC	IMFCC	MFCC+IMFCC
fold1	75.4	80.5	67.2	81.28
fold2	76.6	78.1	61.5	78.97
fold3	75.2	76.5	64.8	76.82
fold4	72.9	73.5	56.2	74.70
average(%)	75.0	77.5	62,4	77.94

6. ACKNOWLEDGEMENT

Thanks to the members of Speech Processing Lab(International Institute of Information Technology(IIIT),Hyderabad,India) for their valuable suggestions,proof-reading and wish to acknowledge IIIT Hyderabad for providing computational resources.
.'

7. REFERENCES

[1] http://www.cs.tut.fi/sgn/arg/dcase2017/challenge/task-acoustic-scene-classification.

[2] Annamaria Mesaros, Toni Heittola, Tuomas Virtanen,"TUT database for acoustic scene classification and sound event detection," in Proc. 24th European Signal Processing Conference 2016 (EUSIPCO, 2016),pp.1128-1132.

[3] Qiuqiang Kong, Iwona Sobieraj, Wenwu Wang and Mark Plumbley , "Deep Neural Network Baseline for DCASE Challenge 2016," In: Detection and Classification of Acoustic Scenes and Events 2016.

[4] Erik Marchi, Dario Tonelli, Xinzhou Xu, Fabien Ringeval, Jun Deng, Stefano Squartini and Bjoern Schull,"Pairwise Decomposition with Deep Neural Networks and Multiscale Kernel Subspace Learning for Acoustic Scene Classification," in: Detection and Classification of Acoustic Scenes and Events 2016.

[5] Daniele Barchiesi, Dimitrios Giannoulis, Dan Stowell, Mark D. Plumbley,"Acoustic Scene Classification: Classifying environments from the sounds they produce" in: IEEE Signal Processing Magazine Year: 2015, Volume: 32, Issue: 3 Pages: 16 - 34, DOI: 10.1109/MSP.2014.2326181.

[6] Sandipan Chakroborty, Anindya Roy, Goutam Saha , "Fusion of a Complementary Feature Set withMFCC for Improved Closed Set Text-Independent Speaker Identification," in:IEEE Conference Year: 2006 Pages: 387 - 390, DOI: 10.1109/ICIT.2006.372388.

[7] Adam: Diederik P. Kingma, Jimmy Ba,"A Method for Stochastic Optimization," in:Published as a conference paper at the 3rd International Conference for Learning Representations, San Diego, 2015.

[8] Ilya Sutskever, James Martens, George Dahl, and Geoffrey Hinton, On the importance of initialization and momentum in deep learning, in Proc. of ICML, 2013, pp. 113 9 1147.

[9] Inkyu Choi, Kisoo Kwon, Soo Bae and Nam Kim , "DNN-Based Sound Event Detection with Exemplar-Based Approach

for Noise Reduction," in:Detection and Classification of Acoustic Scenes and Events 2016.

[10] Mounika K V, Sivanand Achanta, Lakshmi H R, Suryakanth V Gangashetty, and Anil Kumar Vuppala,"An Investigation of Deep Neural Network Architectures for Language Recognition in Indian Languages," in:Proc.INTERSPEECH,2016,pp,2930-2933.

[11] Mashao, D. J. / Skosan, M., " Combining classifier decisions for robust speaker identification," in: PATTERN RECOGNITION; 39, 1; 147-155; PATTERN RECOGNITION von Elsevier Science B.V., Amsterdam. ; 2006.