DCASE 2019 TASK 1A: ACOUSTIC SCENE CLASSIFICATION BY SFFCC AND DNN

Technical Report

Chandrasekhar Paseddula and Suryakanth V. Gangashetty

Speech Processing Laboratory
International Institute of Information Technology, Hyderabad - 500032, India
chandrasekhar.p@research.iiit.ac.in, svg@iiit.ac.in

ABSTRACT

In this study, we dealt with the acoustic scene classification (ASC) task in the Detection and Classification of Acoustic Scenes and Events (DCASE)-2019 challenge Task 1A. Single frequency filtering cepstral coefficients (SFFCC) features and Deep Neural networks (DNN) model is proposed for ASC. We have adopted a late fusion mechanism to further improve the performance and finally, to validate the performance of the model and compare it to the baseline system. We used the TAU Urban Acoustic Scenes 2019 development dataset for training and cross-validation, resulting in a 7.9% improvement when compared to the baseline system.

Index Terms— Acoustic scene classification, Log-Mel band energies, Single frequency cepstral coefficients, Deep neural networks.

1. INTRODUCTION

Classification of predefined acoustic scenes from the test audio recordings is known as ASC. ASC is a very interesting research field in the area of computer auditory scene analysis (CASA). Auditory scene recognition has various applications like monitoring sound by smartphones and robots, etc. Recently DCASE challenge organizers are motivating this field by providing public datasets and baseline systems. Due to that, this field has good scientific submissions towards scene representations and various machine learning models. In DCASE 2013, a bag of frames and Gaussian mixture model (GMM) were proposed for ASC [1]. In DCASE 2016, Mel frequency cepstral coefficients (MFCC)s and GMM model were proposed for ASC [2]. In DCASE 2017, Log-Mel band energies and multilayer perceptron model were proposed for ASC [3]. In DCASE 2018, Log-Mel band energies and convolutional neural network (CNN) model was proposed for ASC in [4, 5]. Generative Adversarial Network Based Acoustic Scene Training Set Augmentation and Selection Using SVM Hyper-Plane were proposed for ASC in [6]. Double image features and the CNN model were proposed for ASC in [7]. An ensemble of Spectrograms Based on Adaptive Temporal Divisions base ASC was done in [8]. In this report, we proposed the SFFCC and DNN model for ASC.

The structure of this technical report is as follows. Section 2 describes the feature extraction. Section 3 presents the proposed framework for classification. Section 4 describes the source of experimental data. Section 5 describes the results and analysis. Section 5 describes the conclusion of the technical report.

2. FEATURE EXTRACTION

In this section, features are extracted using single frequency filtering and feature extraction process was found from [10]. Here we extracted 40 dimensions static, 40 dimensions delta, 40 dimensions double delta features for a frame size of 40 ms with 50% hop length from the entire audio signal, which makes a total of 120 dimensions feature vector.

3. PROPOSED FRAME WORK

The proposed framework architecture is depicted in Figure 1:. It consists of feature extraction and classification. In the feature extraction step, SFFCC and Log-Mel band energies are extracted for both training and test data of TAU Urban Acoustic Scenes 2019 development dataset. In the classification step, DNN model is used as a classifier, where it has 1 input layer, 3 hidden layers, and an output layer. Input layer neurons are 120 with linear activation. Each hidden layer has 200 neurons with ReLU activation. An output layer has 10 neurons with softmax activation. ADAM weight optimizer is used [11]. The classification was done based on the max rule for two individual feature sets. Further, to improve performance, we used a weighted summation rule for implementation of late fusion using DNN scores. The experimental systems are given below:
S1: SFFCC with DNN,
S2: Log-Mel with DNN,
S3: DNN scores level fusion of SFFCC and Log-Mel.

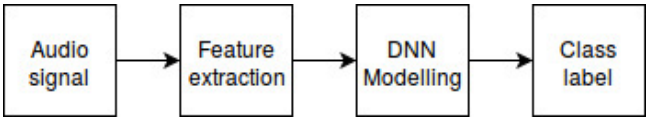

Figure 1: General schematic diagram for proposed system.

4. EXPERIMENTAL DATA

We used the development data of TAU Urban Acoustic Scenes 2019 development dataset for experimentation. Further details can be found from [9, 5]. In this report, system implementation is done by MATLAB.

Table 1: Results for DCASE 2019 task 1A (S1: SFFCC with DNN, S2: Log-Mel with DNN, S3: DNN score level fusion of SFFCC and Log-Mel).

Acoustic Scene(%)	Baseline-2019 [9, 5]	S1 (%)	S2 (%)	S3(%)
Airport	48.4	64.6	47.3	63.4
Bus	62.3	77.8	69.4	77.1
Metro	65.1	69.5	64.2	73.2
Metro_station	54.5	51.0	57.7	57.7
Park	83.1	80.3	81.1	81.9
Public_square	40.7	48.3	49.1	47.3
Shopping_mall	59.4	49.7	73.5	70.5
Street_pedestrian	60.9	65.3	67.6	70.6
Street_traffic	86.7	89.1	91.5	92.0
Tram	64.0	61.2	66.3	70.0
Average	62.5($\pm$ 0.6)	65.7	66.8	70.4

5. RESULTS AND DISCUSSIONS

Table 1 gives the results for DCASE 2019 task 1 subtask A (closed set classification) using the proposed system and baseline system.

From the table, it can be observed that individual Log-Mel band energies perform better than SFFCC. Further, it is observed that the DNN score fusion of SFFCC and Log-Mel band energies (S3) gives a significant improvement in accuracy. Using the proposed features (S3), except Park class, the remaining classes are well classified when compared to DCASE 2019 baseline. From the table, it can also be observed that the proposed system gives an improvement in the average accuracy. The relative improvements of 3.2%, 4.3%, and 7.9% are obtained for S1-S3, respectively, as compared to the DCASE 2019 baseline system.

6. CONCLUSIONS

This report provides a new approach for auditory scene recognition. This approach proposes a new idea in terms of feature extraction. When compared to baseline, 7.9% relative improvement in performance is achieved.

7. REFERENCES

[1] D. Stowell, D. Giannoulis, E. Benetos, M. Lagrange, and M. Plumbley, "Detection and classification of acoustic scenes and events," *Multimedia, IEEE Transactions on*, vol. 17, no. 10, pp. 1733–1746, Oct 2015.

[2] A. Mesaros, T. Heittola, E. Benetos, P. Foster, M. Lagrange, T. Virtanen, and M. D. Plumbley, "Detection and classification of acoustic scenes and events: Outcome of the DCASE 2016 challenge," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 2, pp. 379–393, Feb 2018.

[3] A. Mesaros, A. Diment, B. Elizalde, T. Heittola, E. Vincent, B. Raj, and T. Virtanen, "Sound event detection in the DCASE 2017 challenge," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2019, in press.

[4] http://dcase.community/challenge2018/task-acoustic-scene-classification.

[5] A. Mesaros, T. Heittola, and T. Virtanen, "A multi-device dataset for urban acoustic scene classification," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2018 Workshop (DCASE2018)*, November 2018, pp. 9–13. [Online]. Available: https://arxiv.org/abs/1807.09840

[6] S. Mun, S. Park, D. Han, and H. Ko, "Generative adversarial network based acoustic scene training set augmentation and selection using SVM hyper-plane," DCASE2017 Challenge, Tech. Rep., September 2017.

[7] S. Park, S. Mun, Y. Lee, and H. Ko, "Acoustic scene classification based on convolutional neural network using double image features," DCASE2017 Challenge, Tech. Rep., September 2017.

[8] Y. Sakashita and M. Aono, "Acoustic scene classification by ensemble of spectrograms based on adaptive temporal divisions," DCASE2018 Challenge, Tech. Rep., September 2018.

[9] http://dcase.community/challenge2019/task-acoustic-scene-classification.

[10] K. N. R. K. Alluri, S. Achanta, S. Kadiri, S. Gangashetty, and A. Vuppala, "Detection of replay attacks using single frequency filtering cepstral coefficients," 08 2017.

[11] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. [Online]. Available: http://arxiv.org/abs/1412.6980