Acoustic Scene Classification using Single Frequency Filtering Cepstral Coefficients and DNN

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*—Various representations have been developed for acoustic scene classifications (ASC) task using spectral information. However, there is a wide gap in dealing with acoustic scene representations. In this paper, we propose to use a single frequency filtering (SFF) approach, which provides good temporal and spectral resolution at each instant. Single-frequency filtering cepstral coefficients (SFFCC) with deep neural network (DNN) model as the classifier is used for the experimentation on DCASE 2019 and DCASE 2018 Task 1, development data of subtasks A and B. From the conducted experiments on the development datasets, the usage of the SFFCC features significantly improved ASC performance. This approach has got 35^{th} team rank out of 46 submissions to the corresponding DCASE 2019 Task 1A challenge with a 52.6% classification accuracy on the evaluation dataset. Also, the effect of raw waveforms taken as features for ASC using DNNs was observed.**

***Index Terms*—Log Mel band energies, Single Frequency Filtering Cepstral Coefficients (SFFCC), Acoustic Scene Classification (ASC), Deep Neural Network (DNN).**

I. Introduction

Classification of predefined acoustic scenes from the test audio recordings is known as acoustic scene classification (ASC) (eg., Park, Metro, etc,.). ASC is a very interesting research field nowadays as it has various applications like monitoring sound by smartphones and robots, sound monitoring by artificial intelligence (AI), etc [1], [2]. Detection and Classification of Acoustic Scenes and Events (DCASE) challenge organizers have motivated this field by providing public datasets and baseline systems from the past few years. Due to that, this field has good scientific submissions towards scene representations. In DCASE 2013 baseline, a bag of frames was used for ASC representations and the Gaussian mixture model (GMM) model for classification [1]. In DCASE 2016 baseline, Mel frequency cepstral coefficients (MFCC)s were used for acoustic scene representations and GMM model for classification [3], [4]. Acoustic event detection in real-life recordings using MFCC and hidden Markov model (HMM) were proposed for ASC [5]. In DCASE 2017, log-Mel band energies and multilayer perceptron models were proposed for ASC in [6]. In DCASE 2018, log-Mel band energies and convolutional neural network (CNN) models were proposed for ASC in [7]. Generative Adversarial Network (GAN) based acoustic scene training set augmentation and selection using support vector machine (SVM) hyper-plane were proposed for ASC in [8]. Double image features and the CNN model were proposed for ASC in [9]. An ensemble of spectrograms based on adaptive temporal divisions based ASC was done in [10]. Wavelet transform-based Mel-scaled features for ASC is presented in [11]. DNN based multi-level feature ensemble for ASC was presented in [12]. Audio feature space analysis for ASC was presented in [13]. CNNs for ASC were investigated in [14], [15]. A multi-level attention model for weakly supervised audio classification was proposed in [16]. The significance of phase in single frequency filtering outputs of speech signals was described in [17]. In our approach, SFFCCs were used to represent the acoustic scenes and DNN model is used for classifying the acoustic scenes. The main motivation behind SFFCC handicraft representation of an acoustic scene is that these features capture spectro-temporal information at each instant in mismatched conditions robustly. Acoustic scene detection is possible as it can capture instantaneous spectral variations with high temporal and spectral resolution in the low-frequency regions [18]. Also in [18], SFF envelopes were used for speech and non speech detection as it captures spectral differences between speech and non speech even in the presence of low signal to noise regions highly. Due to that, SFFCC are useful to capture spectral differences more finely between acoustic scenes even in mismatched recording conditions. Motivated by this, we proposed to investigate acoustic scene representations using SFFCC and DNN modelling for ASC and also investigated the effect of raw waveforms with our proposed DNN architecture on DCASE 2018 task1 subtask A.

The remainder of the paper is organized as follows. In Section II, SFFCC features extraction is presented. Section III describes the experimental settings and the database used. In Section IV, Section V, and Section VI, results and discussion are presented. Finally, Section VII provides conclusions.

II. SFFCC features extraction

In this section, features are extracted using single frequency filtering and feature extraction process was found from [18]–[20]. The aim of Single Frequency Filtering (SFF) is to capture the amplitude envelope of the signal as a function of time. Using these SFF envelopes, we can observe the spectral difference between clean and mismatched recordings more clearly [18]. The spectro-temporal resolution can be adjusted by varying the r parameter used in single pole filter transfer function. The SFF method steps are as follows [18]–[20].

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

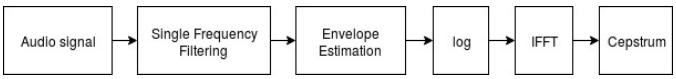

Fig. 1. Block diagram of SFFCC feature extraction.

- The input audio signal $x[n]$ is pre-emphasized to enhance the signal.

$$s[n] = x[n] - \alpha * x[n-1],\ here,\ \alpha = 0.97\ used. \quad (1)$$

- $s[n]$ is multiplied with a complex exponential $e^{j\bar{w}_k n}$, where $\bar{w}_k = \pi - w_k = \pi - 2\pi f_k/f_s$. Then the frequency shifted signal is denoted by

$$s[n,k] = s[n]e^{j\bar{w}_k n}, \quad (2)$$

  where k lies between 0...M, and M refers to the total number of components extracted from speech which is equal to f_s/(2*frequency-hop). Here f_s refers to sampling frequency, and a frequency hop of 50 Hz is used in this study.
- The frequency shifted signal is fed through a single-pole filter $H(z)$, where

$$H(z) = 1/1 + rz^{-1} \quad (3)$$

  here, $r = 0.995$ is considered.
- The output signal $y[n,k]$ is represented by

$$y[n,k] = -ry[n-1,k] + s[n,k] \quad (4)$$

  the amplitude envelope of the signal is given by

$$v[n,k] = \sqrt{(y_r[n,k])^2 + (y_j[n,k])^2} \quad (5)$$

  where y_r, y_j represents the real and imaginary parts respectively. The term $v[n,k]$ corresponds to the SFF envelope of the signal at frequency f_k . The magnitude spectrum can be obtained for each instant of n.
- Cepstrum $c[n,k]$ is computed from $v[n,k]$, and is given by

$$c[n,k] = IFFT(log(v[n,k])) \quad (6)$$

From $c[n,k]$, first 40 cepstral coefficients are considered and they are named as single frequency filtering cepstral coefficients (SFFCCs). The SFFCCs can be obtained at each sampling instant. In this study, using the low SNR instants within 20 ms segment, the cepstral coefficients were extracted. In each segment, the low SNR instant is represented by l. Here l is given by

$$l_k = arg\min_i E_k[i] \quad (7)$$

and

$$E[n] = \sum_{k=0}^{K} v[n,k] \quad (8)$$

Where $E[n]$ is the instantaneous energy. Then the schematic block diagram of SFFCCs extraction is shown in Figure 1.

Here, we considered 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, totalling to a 120 dimensional feature vector.

III. EXPERIMENTAL SETTINGS

The proposed system consists of feature extraction and classification (DNN Modelling) as depicted in Figure 2. 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 and 2018 development dataset of sub-tasks A and B. 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 rectified linear unit (ReLU) activation. An output layer has 10 neurons with softmax activation. ADAM weight optimizer is used at a learning rate of 0.005 to regulate the overfitting [21]. This feed forward neural network architecture and hyperparameters are used for getting the optimal performance. Being frame level supervised training, the standard score normalized data is fed to DNN. The classification was carried based on the majority vote by audio track level. DNNs learn handcrafted features well than CNNs as CNNs learn its own features from the spectrograms. Based on this intuition, we considered standard DNN as a classifier. 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 band energies with DNN,
S3: DNN scores level fusion of SFFCC and log-Mel band energies.

A. Experimental data

For experimentation, we used the development data (DD) of TAU Urban Acoustic Scenes 2019, TAU Urban Acoustic Scenes 2019 Mobile, TAU Urban Acoustic Scenes 2018 and TAU Urban Acoustic Scenes 2018 Mobile. Further, DCASE 2019, 2018 database, baseline system details and subtasks A and B description can be found from [7], [22], [23]. In this paper, feature extraction and acoustic scene classification modelling is implemented using MATLAB. Feature extraction, training and testing is done using Nvidia GeForce GTX 1080 Ti GPU.

B. Decision Strategy

For individual feature sets, computation of DNN score fusion of any two different features is performed as follows: Let us consider the fusion of SFFCC and log-Mel band energies. If X^i_{SFFCC} and $X^i_{log-Mel\,band\,energies}$ are the DNN scores generated by two models for the i^{th} acoustic scene, then a combined score is given by

$$X^i_{combined} = \alpha X^i_{SFFCC} + (1-\alpha) X^i_{log-Mel\,band\,energies}. \quad (9)$$

We observed an improvement in performance for all experiments at $\alpha = 0.4$.

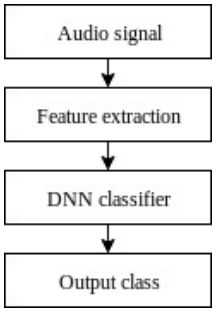

Fig. 2. Block diagram for ASC task implementation using DNN classifier.

IV. Results and Discussions

A. DCASE 2019 task 1 subtask A results analysis

This subtask is considered as basic ASC task. Table I presents the results for DCASE 2019 task 1 subtask A 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) give a considerable improvement in classification accuracy. Using 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.

B. DCASE 2019 task 1 subtask B results analysis

The results of DCASE 2019 task 1 subtask B are presented in Tables II and III. This subtask is concerned with the situation in which an application will be tested with a few different types of audio recording devices (device A, device B-Samsung Galaxy S7 and device C-iPhone SE), not the same device like the one used to record the development data. The results of the subtask B are given in Table II for the DCASE 2019 baseline and the proposed system which uses the DNN score fusion of SFFCC and log-Mel band energies (S3).

From Table III, it can be noted that individual SFFCC features perform better than log-Mel band energies. It can be observed that the DNN score fusion of SFFCC and log-Mel band energies in the proposed system (S3) has given significant improvement compared to the DCASE 2019 baseline system. Overall, 6.3% relative improvement is achieved with the proposed system. From Table II, using proposed features (S3), except Metro, Shopping_mall and Street_pedestrian classes, remaining all classes are well classified for subtask B. Further, we have also experimented with the other proposed systems (S1, S2, and S3). The average (B, C) performance obtained with proposed feature sets was shown in Table III. From the Table, it can be observed that all the proposed systems have given an improvement in the average (B, C) accuracy. The relative improvements of 4.4%, 3.9%, and 6.3% are obtained for S1-S3 respectively, as compared to the DCASE 2019 baseline system.

C. DCASE 2018 task 1 subtask A results analysis

Table IV gives the results for DCASE 2018 task 1 subtask A using the proposed study 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 considerable improvement in classification accuracy and indicates complementary information, which can be concluded based on 12.0% of relative improvement. Using proposed features (S3), except Airport class, the remaining classes are well classified when compared to DCASE 2018 baseline. From the table, it can also be observed that the proposed system gives an improvement in the average accuracy. The relative improvements of 6%, 8.5%, and 12.0% are obtained for S1-S3, respectively, as compared to the DCASE 2018 baseline system.

D. DCASE 2018 task 1 subtask B results analysis

The results of DCASE 2018 task 1 subtask B are presented in Tables V and VI. The results of the subtask B are given in Table V for the DCASE 2018 baseline and the proposed system (S3). The average (B, C) performance obtained with proposed feature sets was shown in Table VI. From the table, it can be observed that all the proposed systems give an improvement in the average (B, C) accuracy. The relative improvements of 12.7%, 8.3%, and 15.0% are obtained for S1-S3, respectively, as compared to the DCASE 2018 baseline system. From this Table VI, individual log-Mel band energies perform better than SFFCC features. For two features combined, the proposed system S3 gave better performance than S1 and S2 with a relative performance improvement of 15.0% compared to DCASE 2018 baseline system. It can be observed that the DNN score fusion of SFFCC and log-Mel band energies in the proposed system (S3) gives significant improvement compared to the DCASE 2018 baseline system. From Table V, using proposed features (S3), except the Airport class, the remaining classes are well classified when compared to DCASE 2018 subtask B baseline.

TABLE I
RESULTS FOR DCASE 2019 TASK1 SUBTASK A (S1: SFFCC WITH DNN, S2: LOG-MEL BAND ENERGIES WITH DNN, S3: DNN SCORE LEVEL FUSION OF SFFCC AND LOG-MEL BAND ENERGIES).

Class Name	Baseline-2019 [22] (%)	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(± 0.6)	65.7	66.8	**70.4**

TABLE II
RESULTS FOR DCASE 2019 TASK1 SUBTASK B FOR THE PROPOSED SYSTEM (S3). A: RECORDING DEVICE (ZOOM F8 AUDIO RECORDER) , B: RECORDING MOBILE DEVICE (SAMSUNG GALAXY S7) AND C: RECORDING MOBILE DEVICE (IPHONE SE).

Class Name	Baseline-2019 [22]				S3			
	A (%)	B (%)	C (%)	Average (B,C) (%)	A (%)	B (%)	C (%)	Average(B,C) (%)
Airport	51.2	18.3	24.1	21.2	60.6	14.8	27.8	21.3
Bus	68.0	40.4	70.0	55.2	78.3	61.1	79.6	70.4
Metro	62.4	50.7	36.1	43.4	68.4	51.9	33.3	42.6
Metro_Station	54.4	28.7	36.1	30.0	54.0	37.0	35.2	36.1
Park	80.4	45.2	57.0	51.1	79.8	92.6	90.7	91.7
Public_square	35.4	22.8	11.3	17.0	48.1	38.9	13.0	26.0
Shopping_mall	64.4	63.5	64.8	64.2	69.4	46.3	66.7	56.5
Street_pedestrian	63.3	37.0	37.6	37.3	69.5	24.1	35.2	29.7
Street_traffic	85.8	77.0	86.5	81.8	92.5	85.2	90.7	88.0
Tram	52.2	12.0	12.6	12.3	68.3	16.7	13.0	14.9
Average	61.9(±0.8)	39.6(±2.7)	43.1(±2.2)	41.4(±1.7)	68.9	46.9	48.5	**47.7**

TABLE III
AVERAGE (B,C) ACCURACIES FOR DCASE 2019 SUBTASK B.

Accuracy	Baseline-2019 [22] (%)	S1 (%)	S2 (%)	S3 (%)
Average	41.4 (± 1.7)	45.8	45.3	47.7

V. DCASE 2019 TASK 1A CHALLENGE RESULTS

This approach gave 52.6% classification accuracy on evolution data. Evolution dataset accuracies corresponding to Seen cities and Unseen cities are shown in Table VII [24].

VI. COMPARISION BETWEEN VARIOUS SYSTEMS

Table VIII shows the comparison between systems of DCASE 2018, DCASE 2019, baseline systems, top rank state of art systems of DCASE 2018, DCASE 2019 and proposed system. From the table, our proposed system outperforms the baseline systems of DCASE 2018 and DCASE 2019 subtask A and subtask B respectively. Also, our proposed system performed nearer to top rank performance of subtask A and subtask B of DCASE 2018 challenge, whereas in the case of DCASE 2019 challenge, our proposed system outperforms the baseline systems of DCASE 2018 and DCASE 2019 subtask A and subtask B respectively. The top rank systems of DCASE 2018 and DCASE 2019 have been suffering more computations and memory than our proposed system due to the ensembling of multi neural networks using CNN models with various spectrograms for getting improved performance [10].

The raw waveform with 40 ms frame length and 20ms overlap vectors and our proposed DNN architecture performance of ASC task using DCASE 2018 task1 subtask A development data is shown in Table IX. The mean accuracy on DCASE 2018 task 1 subtask A dataset on the raw waveforms using our proposed DNN approach got 31.2% accuracy. The low performance is achieved due to unstructured data present in the raw waveforms of acoustic scenes.

VII. CONCLUSIONS

This scientific approach proposes a different idea in terms of feature extraction for ASC. We conclude that from the proposed system (S3) performance in DCASE 2019, 2018 task1 subtask A and B log-Mel band energies performed better on subtask A but SFFCCs better performed on subtask B. This study concludes that log-Mel band energies are useful in a clean environment, whereas SFFCCs perform better in mismatched environments. The reason behind SFFCCs better

TABLE IV
RESULTS FOR DCASE 2018 TASK1 SUBTASK A.

Class Name	Baseline-2018 [23] (%)	S1 (%)	S2 (%)	S3 (%)
Airport	72.9	71.7	50.9	74.0
Bus	62.9	67.4	74.0	78.9
Metro	51.2	71.3	75.9	78.5
Metro_station	55.4	57.1	62.2	63.3
Park	79.1	77.3	86.0	84.3
Public_square	40.4	38.4	45.8	45.4
Shopping_mall	49.6	65.6	89.6	78.5
Street_pedestrian	50.0	63.2	48.2	59.1
Street_traffic	80.5	86.6	90.2	90.7
Tram	55.0	59.0	59.4	64.0
Average	59.7(± 0.7)	65.7	68.2	**71.7**

TABLE V
RESULTS FOR TASK1 DCASE 2018 SUBTASK B FOR THE PROPOSED SYSTEM (S3).

Acoustic Scene	Baseline-2018 [23]				S3			
	A (%)	B (%)	C (%)	$\overline{(B,C)}(\%)$	A (%)	B (%)	C (%)	$\overline{(B,C)}(\%)$
Airport	73.4	68.9	76.1	72.5	61.9	33.3	38.9	36.1
Bus	56.7	70.6	86.1	78.3	71.5	77.8	88.9	83.3
Metro	46.6	23.9	17.2	20.6	62.1	61.1	50.0	55.6
Metro_Station	52.9	33.9	31.7	32.8	64.9	55.6	38.9	47.2
Park	80.8	67.2	51.1	59.2	84.7	83.3	88.9	86.1
Public_square	37.9	22.8	26.7	24.7	50.5	61.1	44.4	52.8
Shopping_mall	46.4	58.3	63.9	61.1	72.4	77.8	88.9	83.3
Street_pedestrian	55.5	16.7	25.0	20.8	55.9	33.3	50.0	41.7
Street_traffic	82.5	69.4	63.3	66.4	89.0	72.2	83.3	77.8
Tram	56.5	18.9	20.6	19.7	75.9	44.4	38.9	41.7
Average	58.9(±0.8)	45.1(±3.6)	46.2(±4.2)	45.6(±3.6)	68.9	60.0	61.1	**60.6**

TABLE VI
AVERAGE (B, C) ACCURACIES FOR DCASE 2018 TASK1B.

Accuracy	Baseline-2018 [23] (%)	S1 (%)	S2 (%)	S3 (%)
	45.6(± 3.6)	58.3	53.9	**60.6**

TABLE VII
ACCURACIES FOR DCASE 2019 TASK1 SUBTASK A ON EVOLUTION DATASET (ED), SEEN CITIES (SC)EVOLUTION DATASET AND UNSEEN CITIES (UC) EVOLUTION DATASET.

Accuracy	ED [24] (%)	DD (%)	SC (ED) (%)	UC (ED) (%)
	52.6	70.4	54.8	41.3

performance is due to good capture of spectral variations of acoustic scenes in mismatched environment recordings than clean environment recordings. The relative improvement of 7.9% and 6.3% in performance was achieved when compared to DCASE 2019 task1 subtask A and B baselines respectively. The relative improvement of 12.0% and 15.0% in performance was achieved when compared to DCASE 2018 task1 subtask A and B baselines respectively. By observing the performance of raw waveforms on DCASE 2018 task1 subtask A development data, we can concluded that research on feature representation and fusion of different features is needed for better performance of ASC task. Then the future work could be devoted to exploring different features for ASC.

TABLE VIII
RESULTS COMPARISIONS ON DCASE 2018, 2019 TASK1 SUBTASK A, SUBTASK B AND PROPOSED SYSTEM

Systems	subtask A (%)	subtask B (%)
DCASE 2018 Baseline	59.7 (± 0.7)	45.6 (± 3.6)
DCASE 2018 Rank 1	76.9 [10]	63.6 [25]
Proposed System (S3)	71.7	60.6
DCASE 2019 Baseline	62.5 (± 0.6)	41.4 (± 1.7)
DCASE 2019 Rank1	85.0 [26]	70.0 [27]
Proposed System (S3)	70.4	47.7

TABLE IX
THE PERFORMANCE OF PROPOSED SYSTEM TAKEN RAW WAVEFORM AS FEATURES USING DCASE 2018 TASK1 SUBTASK A DEVELOPMENT DATASET

Database	Accuracy (%)
DCASE 2018 task1 subtask A development dataset	31.2

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] S. Aziz, M. Awais, T. Akram, U. Khan, M. Alhussein, and K. Aurangzeb, "Automatic scene recognition through acoustic classification for behavioral robotics," *Electronics*, vol. 8, no. 5, p. 483, 2019.

[3] 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 challenge," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 2, pp. 379–393, Feb 2018.

[4] C. Paseddula and S. V. Gangashetty, "Dnn based acoustic scene classification using score fusion of mfcc and inverse mfcc," in *2018 IEEE 13th International Conference on Industrial and Information Systems (ICIIS)*, 2018, pp. 18–21.

[5] A. Mesaros, T. Heittola, A. Eronen, and T. Virtanen, "Acoustic event detection in real-life recordings," in *18th European Signal Processing Conference*, 07 2014.

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

[7] 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 Workshop (DCASE)*, November 2018, pp. 9–13. [Online]. Available: https://arxiv.org/abs/1807.09840

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

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

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

[11] S. Waldekar and G. Saha, "Wavelet transform based mel-scaled features for acoustic scene classification," in *Proc. Interspeech*, 2018, pp. 3323–3327. [Online]. Available: http://dx.doi.org/10.21437/Interspeech

[12] J.-w. Jung, H.-s. Heo, H.-j. Shim, and H.-j. Yu, "DNN based multi-level feature ensemble for acoustic scene classification," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events Workshop (DCASE)*, November 2018, pp. 113–117.

[13] T. Maka, "Audio feature space analysis for acoustic scene classification," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events Workshop (DCASE)*, November 2018, pp. 113–117.

[14] L. Pham, I. McLoughlin, H. Phan, and R. Palaniappan, "A robust framework for acoustic scene classification," *Proc. Interspeech 2019*, pp. 3634–3638, 2019.

[15] H. Zeinali, L. Burget, and J. H. Cernocky, "Convolutional neural networks and x-vector embedding for DCASE acoustic scene classification challenge," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events Workshop (DCASE)*, November 2018, pp. 202–206.

[16] C. Yu, K. S. Barsim, Q. Kong, and B. Yang, "Multi-level attention model for weakly supervised audio classification," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events Workshop (DCASE)*, November 2018, pp. 188–192.

[17] N. Chennupati, S. R. Kadiri, and Y. B., "Significance of phase in single frequency filtering outputs of speech signals," *Speech Communication*, vol. 97, p. 66–72, 2018. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167639317301747

[18] G. Aneeja and B. Yegnanarayana, "Single frequency filtering approach for discriminating speech and nonspeech," *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, vol. 23, no. 4, pp. 705–717, Apr. 2015. [Online]. Available: http://dx.doi.org/10.1109/TASLP.2015.2404035

[19] K. R. Alluri, S. Achanta, S. R. Kadiri, S. V. Gangashetty, and A. K. Vuppala, "Detection of replay attacks using single frequency filtering cepstral coefficients," in *Proc. Interspeech*, 2017, pp. 2596–2600. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.

[20] S. R. Kadiri and B. Yegnanarayana, "Analysis and detection of phonation modes in singing voice using excitation source features and single frequency filtering cepstral coefficients (sffcc)," in *Proc. Interspeech*, 2018, pp. 441–445. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.

[21] 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

[22] *DCASE 2019*, (accessed March 31, 2020), http://dcase.community/challenge2019/task-acoustic-scene-classification.

[23] *DCASE 2018*, (accessed March 31, 2020), http://dcase.community/challenge2018/task-acoustic-scene-classification.

[24] *DCASE 2019 Results*, (accessed March 31, 2020), http://dcase.community/challenge2019/task-acoustic-scene-classification-results-a.

[25] T. Nguyen and F. Pernkopf, "Acoustic scene classification using a convolutional neural network ensemble and nearest neighbor filters," DCASE2018 Challenge, Tech. Rep., September 2018.

[26] H. Chen, Z. Liu, Z. Liu, P. Zhang, and Y. Yan, "Integrating the data augmentation scheme with various classifiers for acoustic scene modeling," DCASE2019 Challenge, Tech. Rep., June 2019.

[27] M. Kośmider, "Calibrating neural networks for secondary recording devices," DCASE2019 Challenge, Tech. Rep., June 2019.