ACOUSTIC SCENE CLASSIFICATION USING MULTI-CHANNEL AUDIO FEATURE WITH CONVOLUTIONAL NEURAL NETWORKS AND SUBTRACT FILTER AUGMENTATION

Jaehun Kim[*]

AI Research Lab, IVS Inc, Seoul, South Korea
kjh21212@gmail.com

ABSTRACT

This paper presents a multi-channel audio feature using imagenet model based on convolutional neural networks for DCASE 2020 Task1-A Acoustic scene classification with multiple devices. We use the TAU Urban Acoustic Scenes 2020 Mobile Dataset. It consists of 10 seconds of audio clips about 10 scenes. We proposed a multi-channel audio feature to use imagenet pre-trained model weight. also, we proposed filtered augmentation for other devices' recorded audio. the multi-channel feature consists of raw and harmonic, percussive (HPSS) data's Log-Mel-Spectrogram. Also, we use EfficientNet pre-trained model weight.

Index Terms— Audio databases, Secene classification, Sound classification, Convolutional neural networks

1. INTRODUCTION

This challenge is The Detection and Classification of Acoustic Scenes and Events (DCASE) [1], This paper based on DCASE's Task1-A (Figure 1). This task's primary goal is to classify 10 scenes (Airport, Indoor shopping mall, Metro station, Pedestrian street, Public square, Street with a medium level of traffic, Travel-ling by a Tram, Travelling by a Bus, Travelling by an underground Metro, Urban Park) in 10 seconds of other device recorded audio files. The development dataset [2] is composed of 13965 record-ings in the training dataset and 2970 recordings in the validate dataset. this dataset is recorded by other devices each. That device's names are A to C, S1 to S6. S4 to S6 is just test data. Recently, image classification research has greatly improved by google. The network name is EfficientNet [3]. This network surpassed the performance of existing networks. we propose how to apply this network to this task for the best performance. also, we proposed filter augmentation for recorded audio files by other devices.

2. FEATURE EXTRACTION

To use the pre-trained weight of EfficientNet, the audio features were processed like images. JPEG Image data consists of width, height, channel (RGB), this data shape is (width, height, channel). So, we made the audio data shape like image data shape. We used the 10s audio data of 44.1Khz sample and used the librosa library [4] for feature extraction.

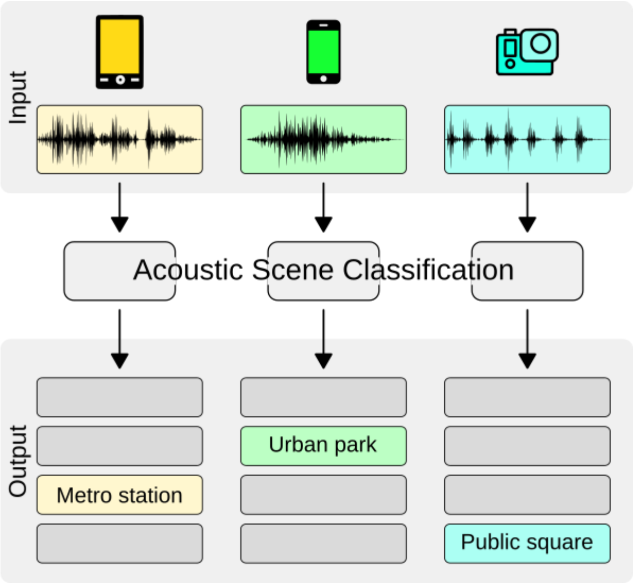

Figure 1: Overview of acoustic scene classification system.

2.1. HPSS

We used the median-filtering harmonic percussive source separation (HPSS) [5] to construct a multi-channel feature. This process is split the raw audio (R) to harmonic (H) and percussive (P) components. then, It can earn a total of three components raw, harmonic, percussive data (RHP) similar to the RGB channel.

2.2. Normalization

Digital raw audio min, max range of float32 is -1 to 1. we worked the normalization of each channel data consisted of RHP (figure 2).

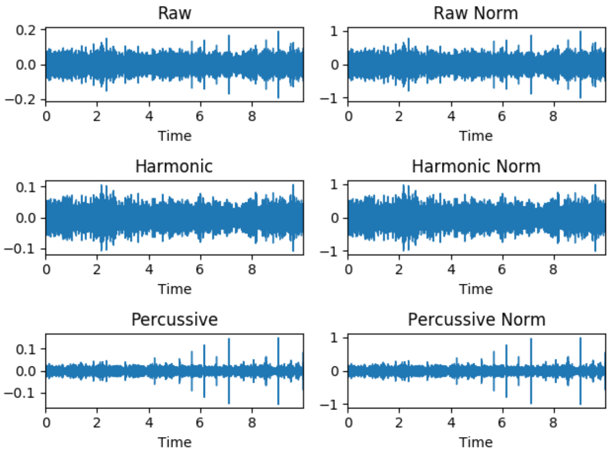

Figure 2: Raw, Harmonic, Percussive channels normalization

2.3. Log-Mel-Spectrogram

Recently many audio researchers use the Log-Mel-Spectrogram. because it compresses the raw audio data to dB and frequency according to time. It shows the best performance in the audio environment sound classification task [6]. We applied this function to normalized channel data every each (figure 3). The used parameters depend on Efficientnet model.

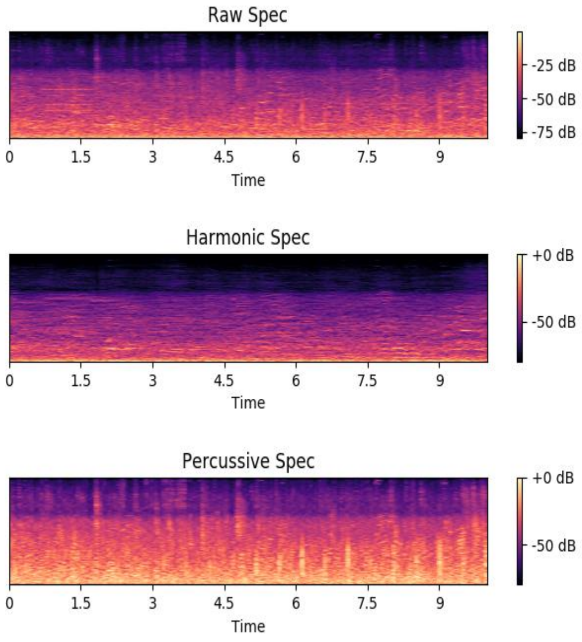

Figure 3: RHP channels log-mel-spectrogram

Finally, applied all channel entirety normalization. If used Efficientnet B5 Model result shape is (456, 456, 3) in 10 seconds audio. This processing is important. Final normalization reduces the distance of each data. Figure 4, 5 shows the location of the data by Principal Component Analysis (PCA) [7].

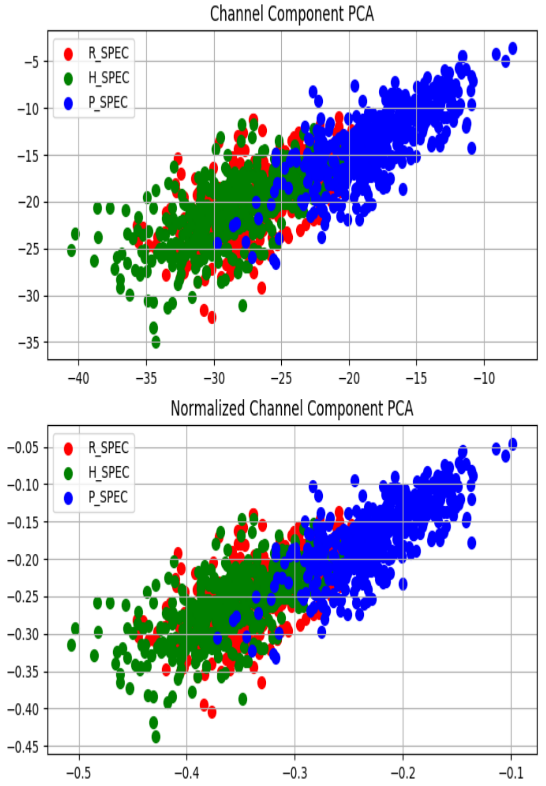

Figure 4: Each channels PCA graph
(Unnormalized, Normalized)

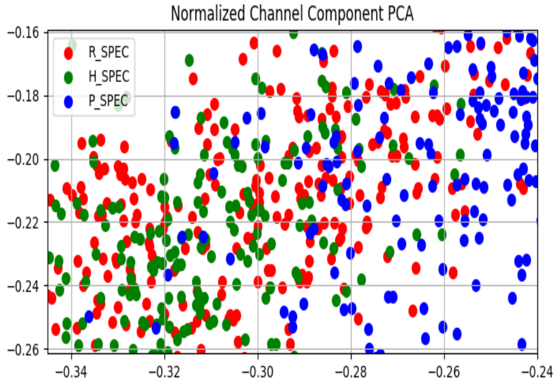

Figure 5: Final normalized channels PCA graph zoom in

3. AUGMENTATION

We propose to subtract filter augmentation (SFA) for the recorded audio file by other devices. How to make the subtract filter is just the average of subtracts the main device's Mel spectrogram from other device's Mel spectrogram. This method must have the same environment's same sound source and the sound source is recorded by different devices. we made each device's filter (b, c, and s1 to s6). And we just added each filter value to the train

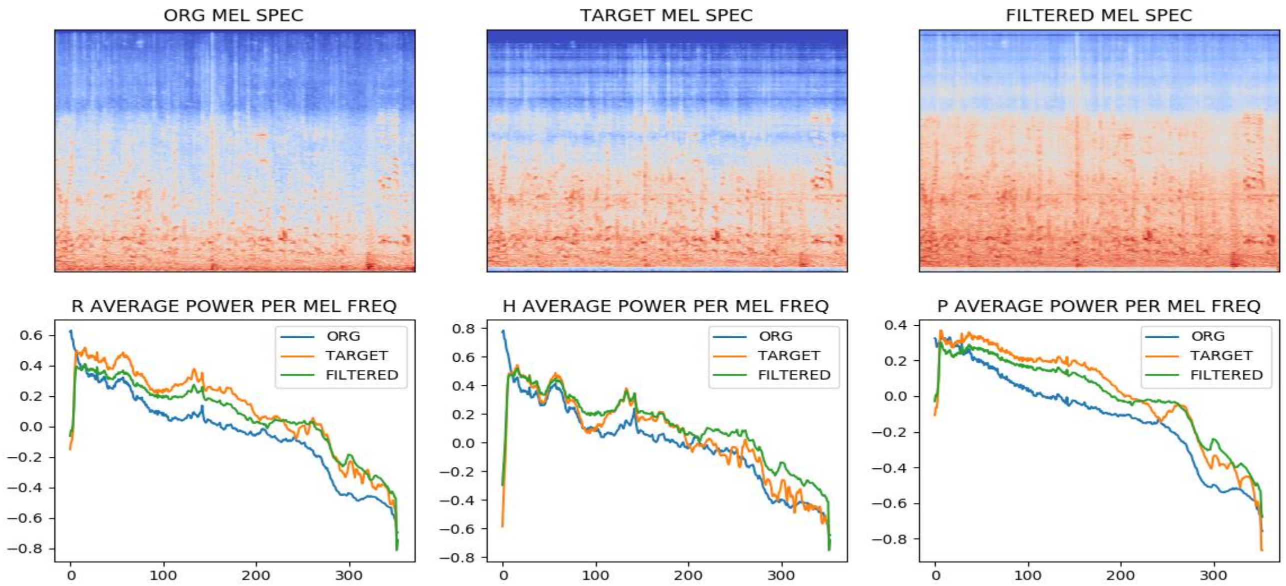

Figure 6: Mel spectrogram and, average power per mel-frequency about SFA

data's Mel spectrogram of the main device. figure 6 shows mel-spectrogram about augmentation, the graph is average power per mel-frequency of the Filtered Mel spec, original Mel spec, target Mel spec. Target Mel spec is another device recorded audio.

4. NETWORK ARCHITECTURE

We use 10-second audio data for feature extraction. So, the input is two split audio features. Also, we use EfficientNet architecture. figure 7 is a proposed net-work architecture.

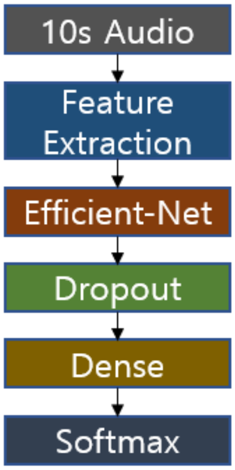

Figure 7: Proposed network architecture

5. EXPERIMENT

The EfficientNet has B0 to B7 model. we used B0 and B5 noisy-student pre-trained model weight and not included top, used the average pooling. When training we used this parameter setting. using the early stopper, batch size 16, adam optimizer [8] 0.0001, L2 regularizer [9] 0.0001, dropout [10] size depends on EfficientNet model and use masked loss function. we evaluated each epoch model in architecture for the best performance about coarse, fine level. And an additional point is this model's batch size required to more than 8 per GPU. Otherwise, model performance is low about 2~3% (Table 1).

Table 1: Accuracy according to batch size about proposed model in Task1A validate dataset
(Accuracy, %)

Model	Batch per gpu	Acc
B0	4	66.07
	8	67.18
B3	4	66.84
	8	68.82

6. RESULT

We used the Dcase evaluation metrics [11] for this task. Table 2 shows the architectures network result in each used EfficientNet model B0, B3, B5. It also shows augmentation's usefulness. SFA show about 2% accuracy increase. Our GPU server performance couldn't use B6 to B7. but we achieved the best performance in this task. it showed accuracy increase according to model version increase. consequently, the best architecture was EfficientNet-B5 is an applied network. If we have a larger GPU server, we can make a better performance model.

Table 2: Evaluation result in Dcase Task1A validate dataset (Accuracy, %)

Model	Acc			
Baseline	54.10			
	Non-Aug		Aug	
	Batch per gpu	Acc	Batch per gpu	Acc
B0	8	65.01	8	67.18
B3	8	66.12	8	68.82
B5	**4**	**68.21**	**4**	**70.11**

7. CONCLUSION

This paper shows us audio data use like images and, how to use the imagenet pre-trained model in the audio tasks. And, SFA is useful for other device recorded audio. Consequently, it is possible to determine whether or not the weights of existing trained models can be used according to the processing of data. This just shows that the data is not limited to the task.

8. ACKNOWLEDGMENT

This research was supported by a grant (19PQWO-B153369-02) from Smart road lighting platform development and empirical study on test-bed by Ministry of Land, Infrastructure and Transport of Korean Government.

9. REFERENCES

[1] http://dcase.community/workshop2020/.

[2] http://dcase.community/challenge2020/task-acoustic-scene-classification

[3] TAN, Mingxing; LE, Quoc V. Efficientnet: Rethinking model scaling for convolutional neural networks. arXiv preprint arXiv:1905.11946, 2019.

[4] https://librosa.github.io/librosa/

[5] FITZGERALD, Derry. Harmonic/percussive separation using median filtering. In: Proc. of DAFX. 2010.

[6] HUZAIFAH, Muhammad. Comparison of time-frequency representations for environmental sound classification using convolutional neural networks. arXiv preprint arXiv:1706.07156, 2017.

[7] WOLD, Svante; ESBENSEN, Kim; GELADI, Paul. Principal component analysis. Chemometrics and intelligent laboratory systems, 1987, 2.1-3: 37-52.

[8] KINGMA, Diederik P.; BA, Jimmy. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980, 2014.

[9] CORTES, Corinna; MOHRI, Mehryar; ROSTAMIZADEH, Afshin. L2 regularization for learning kernels. arXiv preprint arXiv:1205.2653, 2012.

[10] SRIVASTAVA, Nitish, et al. Dropout: a simple way to prevent neural networks from overfitting. The journal of machine learning research, 2014, 15.1: 1929-1958.

[11] http://dcase.community/challenge2020/task-acoustic-scene-classification#evaluation