ACOUSTIC SCENE CLASSIFICATION BASED ON CONVOLUTIONAL NEURAL NETWORK USING DOUBLE IMAGE FEATURES

Sangwook Park
Korea University
School of Electrical Eng., Seoul, 136-713, South Korea
swpark@ispl.korea.ac.kr

Seongkyu Mun
Korea University
Dept. of Visual Information Processing, Seoul, 136-713, South Korea
skmoon@ispl.korea.ac.kr

Younglo Lee
Korea University
School of Electrical Eng., Seoul, 136-713, South Korea
yllee@ispl.korea.ac.kr

Hanseok Ko
Korea University
School of Electrical Eng., Seoul, South Korea
hsko@korea.ac.kr

ABSTRACT

This paper proposes new image features for the acoustic scene classification task of the IEEE AASP Challenge: Detection and Classification of Acoustic Scenes and Events. In classification of acoustic scenes, identical sounds being observed in different places may affect performance. To resolve this issue, a covariance matrix, which represents energy density for each subband, and a double Fourier transform image, which represents energy variation for each subband, were defined as features. To classify the acoustic scenes with these features, Convolutional Neural Network has been applied with several techniques to reduce training time and to resolve initialization and local optimum problems. According to the experiments which were performed with the DCASE2017 challenge development dataset it is claimed that the proposed method outperformed several baseline methods. Specifically, the class average accuracy is shown as 83.6%, which is an improvement of 8.8%, 9.5%, 8.2% compared to MFCC-MLP, MFCC-GMM, and CepsCom-GMM, respectively.

Index Terms— Acoustic scene classification, covariance learning, double FFT, convolutional neural network

1. INTRODUCTION

Audio signals ranging from speech and general sounds (non-linguistic sound) to background sounds may be quite informative in characterizing context such as presence of humans, objects, their activities, or the environment. Among these contexts, location information is not only vital in multimedia analysis but also widely applicable to many tasks in scene understanding [1-2]. Location information is also useful as prior information for enhancing performance of speech/acoustic event recognition [3-4]. Thus, Acoustic Scene Classification (ASC) which focuses on identifying where the audio signal has been obtained has drawn considerable attention. IEEE AASP Challenge: Detection and Classification of Acoustic Scenes and Events (DCASE) 2017 also included this task as the ASC challenge.

Typically, an ASC system consists of feature extraction and classification. In feature extraction, Mel Frequency Cepstral Coefficients (MFCCs) and Perceptual Linear Prediction (PLP) have been applied as the early stage of the proposed ASC system. Low-level spectral features such as zero-crossing rate, spectral statistics, and timbre were employed for ASC by combining them with MFCCs [5]. A Bag-Of-Frames (BOF) method, which considers an acoustic scene as a set of bags of various sounds, was applied to ASC [6-8]. The BOF approach has used statistical distribution (e.g. histogram) as features, which represents the occurrence count of cepstral features, quantized by a codebook like dictionary. However, the BOF approach is too sensitive to training data due to the requirement of training phases in both feature extraction and classification.

Recently, approaches based on i-vector which is widely being used for speaker recognition has also been applied for ASC [9]. Since i-vector is extracted from hyper-dimensional vector space by applying factor analysis, potential discriminable characters can be revealed with the feature. In the last challenge, numerous approaches based on deep learning were introduced. Mun et al. proposed a classification framework based on bottleneck feature extraction with Deep Neural Networks (DNN) [10]. Takahashi et al. investigated DNN-Gaussian Mixture Model (GMM) framework for classifying MFCCs [11]. Similarly, Convolutional Neural Network (CNN) was applied for classifying log-mel spectrograms [12]. In [13], an ensemble method which is composed of hundreds of CNNs was proposed for stochastic feature extraction. Recurrent Neural Network (RNN) based approaches were also proposed [14-15]. In [16], combined CNN and Long Short-Term Memory (LSTM) model has been proposed for ASC. For applying deep learning methods, sufficient training data is required to avoid local optimum problems. Thus, manipulation of development datasets is also considered as one of the major issues for training DNNs.

Although many approaches have attempted ASC applications, they still suffer from realistic environment problems. Even in the same environment, audio signals may vary depending on presence of people, objects, and their behaviors. For example, in a café, a microphone may collect differing occurrences of sounds such as cleaning, coffee grinding, or people talking. Therefore, the feature vectors obtained in cafés will likely be widely scattered in a feature space, although these vectors have originated from the same place. Meanwhile, features representing conversations will always be observed in all environments where there are people. As illustrated by this example, performing location classification for ASC is a challenging task in realistic environments.

This paper describes an ASC method which is applied for the DCASE 2017 challenge under this practical issue. According

to a hypothesis that sound frequency over the time may differ according to places, two types of image features, covariance matrix and double Fast Fourier Transform (FFT) image, are proposed. These features represent temporal energy density and energy variations for each frequency, respectively. Also, they are insensitive to training data, because they can be obtained without any trained data. To perform scene classification with these features, Convolutional Neural Network (CNN) is considered. Additionally, the appropriate CNN structure is also investigated to improve performance. In experiment, the proposed method is demonstrated by using the DCASE 2017 challenge database [17].

The remainder of this paper is organized as follows. Section 2 explains proposed image features with its motivation. Section 3 introduces CNN approaches for classifying the proposed features. After a discussion on the experimental results, conclusions are drawn in the final section.

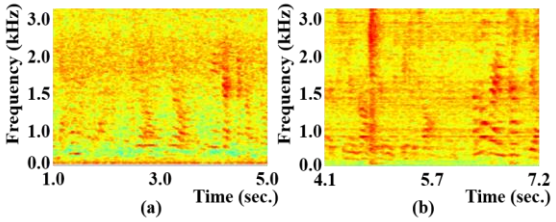

Figure 1: Human voice spectrograms included in development dataset for DCASE2017 challenge; (a) in library, a part of "b027_190_200.wav" (b) in home, a part of "a031_100_110.wav"

2. PROPOSED IMAGE FEATURE

2.1. Motivation

Although the spectrograms are obtained in distinct places, human voices are heard in both places, *library* and *home* as shown in Figure 1. In this case, many conventional features extracted from spectrum may encounter confusion due to not only human voice but also other sounds heard anywhere, because their spectrums look very similar to each other. To overcome this problem, it is necessary to search for difference by the combination of spectrums during a finite interval. One method to observe this combination of these spectral features is via histogram with BOF. However, BOF is too sensitive to training data, because it requires training phases for feature extraction as well as classification.

In this paper, to representing combination of spectrums during a finite interval, covariance matrix and frequency analysis of frequency bins are considered. Figure 2 shows covariance matrices and images obtained by performing FFT on spectrogram in each frequency bin. As shown in Figure 2, *library* and *home* can be distinguished by using these images. Based on this fact, two image features, covariance matrix of spectrums (COV) and Double FFT Image (DFI), are proposed.

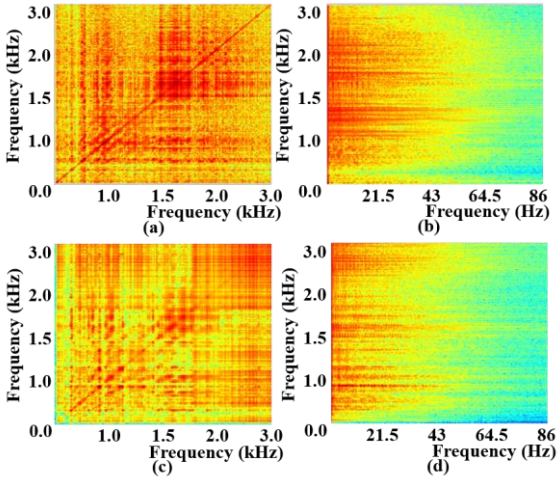

Figure 2: Two image features corresponding to Figure 1; (a) covariance matrix in library (b) frequency analysis in library (c) covariance matrix in home (d) frequency analysis in home

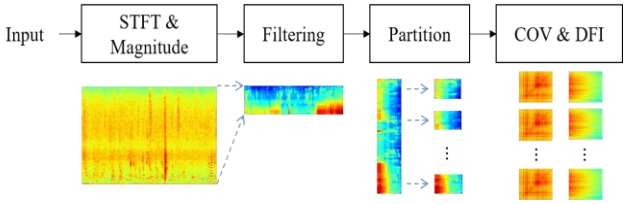

Figure 3: Procedure for obtaining proposed image feature

2.2. Image Feature Extraction

Figure 3 shows the procedure of obtaining the two image features. A 1-dimensional wave is transformed to spectrogram after pre-emphasis. In *Filtering*, a compressive Gammachirp filterbank is applied to all spectrums for dimension reduction [18]. The result of *Filtering* is partitioned into several blocks that are composed of consecutive filter responses. For each block, COV is calculated by performing expectation as

$$\boldsymbol{C}^i = E\left[\left(X[:,m]-\overline{X}^i\right)\left(X[:,m]-\overline{X}^i\right)^T\right],\quad X[:,m]\in B^i \qquad (1)$$

where B^i is a set whose elements are filter responses included in the i^{th} block. $\boldsymbol{C}^i$ is a covariance matrix of the i^{th} block. X and X^i are filter responses in each frame and frame average, respectively, and m is frame index.

To obtain DFI, FFT is performed on each subband of filter responses as

$$\boldsymbol{D}^i = F_{k=1:K}\left\{X[k,:]\right\},\quad X[k,:]\in B^i \qquad (2)$$

where $\boldsymbol{D}^i$ is a DFI of the i^{th} block, F is a function for FFT. K and k is the number of frequency bin and frequency index, respectively. Finally, min-max normalization is performed on both $\boldsymbol{C}^i$ and $\boldsymbol{D}^i$ for representing gray-scale image.

3. CONVOLUTIONAL NEURAL NETWORK

Many deep learning methods suffer from several problems such as over-fitting, local optimum, and training time. In CNN, the number of parameters is reduced by sharing weights in convolutions to avoid over-fitting problems. Also, CNN is well known to be appropriate for classifying image. From these reasons, CNN is applied for classification of proposed image feature. As shown in Figure 4, the CNN structure consists of three parts; input, convolution, and fully connected.

	Output 1024×15 BN-Relu
Fully connected	Fully connected 512×1024 BN-Relu
	Fully connected 2048×512 BN-Relu
	Fully connected 2048×2048 BN-Relu
Convolution	2×2 max-pooling (padding, stride-2)
	3×3 conv.(padding, stride-1) 32 BN-Relu
	3×3 conv.(padding, stride-1) 32 BN-Relu
	3×3 conv.(padding, stride-1) 32 BN-Relu
	2×2 max-pooling (padding, stride-2)
	5×5 conv.(padding, stride-1) 16 BN-Relu
	5×5 conv.(padding, stride-1) 16 BN-Relu
	5×5 conv.(padding, stride-1) 16 BN-Relu
	2×2 max-pooling (padding, stride-2)
	7×7 conv.(padding, stride-1) 8 BN-Relu
	7×7 conv.(padding, stride-1) 8 BN-Relu
	7×7 conv.(padding, stride-1) 8 BN-Relu
	Input: 64×64×4 [left & right: COV, DFI,]

Figure 4: A CNN structure applied for classification of proposed image features

3.1. CNN Input

Similar to human perception, stereo sound contains additional information such as direction of sound and spatial characters. Since these are also helpful for location recognition, CNN input is composed of four images which are COVs and DFIs obtained from each channel. These inputs are converted to gray scale image whose pixel value is integer within 0 to 255. Thus, all pixels in the proposed feature are normalized with zero-mean and unit-variance.

In feature extraction, the number of filters included in the considered filterbank is set to 64, and the number of double FFT points is set to 128.

3.2. Convolution

In training of the deep network, obtaining initialization parameters is very important to avoid local optimum problem. Since CNN structures are empirically determined, training time is also another issue. To alleviate these problems, batch normalization has been applied in every layer [19]. Thus, three substeps, convolution, batch normalization, and pooling, are conducted in this step.

In convolution, if a large filter is used, microscopic features can be obtained, but the number of training parameters is also increased. To avoid this constraint, convolutions are iteratively performed with a small filter. Based on this concept, filter sizes applied to final structure are depicted in Figure 4. After convolution, Batch Normalization (BN) is also performed.

In pooling, the image size is diminished in half by applying max pooling. Note that the number of tensors is increased after pooling. This is also iteratively performed after every convolution iteration.

3.3. Fully connected

After all iterations, the result of *Convolution* is reshaped to vector (8x8x32 dimension) for application to the fully connected network. In this time, Rectified Linear Unit (Relu) function is used for activation function. The number of nodes in each layer is depicted in Figure 4.

4. EXPERIMENT

4.1. Experimental Setting

For performance assessment, DCASE2017 dataset that consists of 15 scenes, *bus, café/restaurant, car, city center, forest path, grocery store, home, lakeside beach, library, metro station, office, residential area, train, tram,* and *urban park*, was used. By using four fold lists provided by DCASE2017 committee, cross-validation tests were conducted.

For performance comparison, several baselines were considered. Firstly, the results of MFCC-MLP and MFCC-GMM were provided by DCASE2017 challenge committee [20]. Secondly, CepsCom that is a 240-dimensional vector composed by concatenating four cepstral features was evaluated by using 128 mixture GMM [21]. Finally, CepsCom based i-vector framework was considered [22]. Based on 128 mixture GMM, a 400 dimensional i-vector was extracted in this experiment. After applying multi-class Linear Discriminative Analysis (LDA) to 400 dimensional i-vector, classification was performed by using a minimum Cosine Distance Score (CDS).

In proposed method, the length of block was empirically set to 1 second. Thus, CNN was trained with approximately 33,000 inputs in each fold test.

4.2. Development Results

The accuracies according to classes are summarized in Table 1. In baseline systems except i-vector-CDS, the performance is shown to be about 75%. Although logMel-MLP and MFCC-GMM shows similar performance (i.e. class averaging accuracy), CepsCom-GMM shows the best averaging accuracy, which is an improvement of 0.6%. In logMel-MLP and MFCC-GMM, accuracies above 90% can be obtained in *car, city center* and *office*, and accuracy below 60% is shown in *train*. On the other hand, MFCC-GMM outperforms other methods in *café* and

Table 1. *Experiment results for four baseline systems and proposed system in development*

[%]	Avg.	beach	bus	cafe	car	city	forest	groc.	home	lib.	metro	office	park	resid.	train	tram
logMel-MLP	74.8	75.3	71.8	57.7	97.1	90.7	79.5	58.7	68.6	57.1	**91.7**	**99.7**	70.2	64.1	58.0	81.7
MFCC-GMM	74.1	75.0	84.3	**81.7**	91.0	91.0	73.4	67.9	71.4	63.5	81.4	97.1	39.1	**74.7**	41.0	79.2
CepsCom-GMM	75.4	78.2	82.9	75.0	91.0	**91.7**	61.1	81.4	70.8	58.0	78.1	96.8	53.9	74.7	54.6	**83.5**
i-vector-CDS	61.9	79.5	63.5	56.7	32.1	86.9	68.4	77.9	54.8	56.7	69.4	66.0	55.1	48.7	51.5	64.0
Proposed	**83.6**	**88.5**	**93.8**	73.1	**93.2**	86.3	**97.1**	**84.6**	**82.5**	**77.8**	89.2	91.0	**73.1**	70.5	**70.1**	82.7

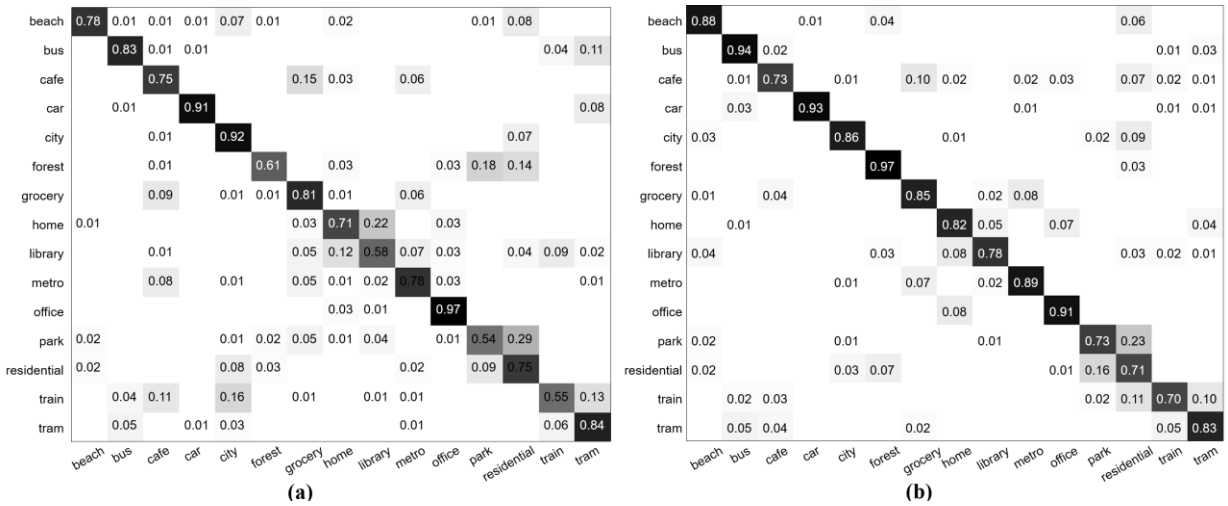

Figure 5: Confusion matrices for CepsCom-GMM and the proposed method (a) CepsCom-GMM (b) Proposed

Table 2. *Experiment result for proposed system in evaluation*

[%]	Avg.	beach	bus	cafe	car	city	forest	groc.	home	lib.	metro	office	park	resid.	train	tram
Proposed	**72.6**	54.6	59.3	71.3	79.6	91.7	85.2	75.0	98.1	44.4	98.1	84.3	23.1	76.9	82.4	64.8

residential area while logMel-MLP outperforms others in *metro-station* and *office*. This difference may come from that different classifier, MLP or GMM, which has been applied to each method. In addition, the features is also different. Since 40 filters for Mel filterbank are used for feature extraction, logMel is a 40-dimensional vector while MFCC is a 12-dimensional vector by conducting Discrete Cosine Transform (DCT). In case of CepsCom-GMM, this method outperforms others in *city center* and *tram*.

In i-vector-CDS, the performance is about 62%. An UBM is very important for extracting i-vector, and a huge volume of database including a lot of scenes is required for training UBM. Despite this fact, development dataset consisted of 15 scenes is only used for training UBM in this experiment. To obtain reliable results using i-vector, a larger database is required to successfully training UBM.

According to the results, the proposed method outperforms other methods. The class average accuracy was observed as 83.6%, which is an improvement of 8.8%, 9.5%, 8.2% compared to MFCC-MLP, MFCC-GMM, and CepsCom-GMM, respectively, which implies that the best class accuracies were obtained for most classes. Additionally, confusion matrices for CespCom-GMM and the proposed method are shown in Figure 5. As mentioned previously, spectrum based features such as MFCC and CepsCom confuse scenes where common sound may be heard. Although class accuracies have been lower than baseline in *café* and *office*, the proposed method resolves this problem as shown in *bus*, *library*, *home* and *train*. To additionally improve performance of the proposed method, the confusion between *park* and *residential area* has to be resolved.

4.3. Evaluation Results

For this experiment, 90% of development data were used for training CNN described in Figure 4. And the remainder was considered as validation data for determining the number of epoch. After training, evaluation dataset is tested in the CNN, and the results are summarized in Table 2. Accuracies above 90% can be obtained in *city center*, *home*, and *metro-station*, on the other hand accuracies in *library* and *park* are below 50%. Also, average accuracy of evaluation is degraded as much as about 11 % compared to development accuracy. From these fact, generalization issue still remains in the proposed method, and the issue will be investigated in future.

5. CONCLUSIONS

This paper proposed new image features, COV and DFI, for resolving an issue that common sounds can be heard anywhere. The COV is a covariance matrix of spectrums which represents energy densities for each frequency subband. The other feature, DFI, represents variation of energy in each subband, which can be obtained by performing FFT. These features can be easily obtained without training data. To perform classifying using these features, CNN is applied with several techniques for reducing training time and resolving problems about initialization and local optimization. Efficiency of proposed method is demonstrated in experiment with development dataset provided for DCASE2017 challenge. From the results, proposed method outperforms other methods by means of class average accuracy with 83.6%, which is an improvement of 8.8%, 9.5%, 8.2% compared to MFCC-MLP, MFCC-GMM, and CepsCom-GMM, respectively. In future, a method for resolving generalization issue will be investigated.

6. ACKNOWLEDGMENT

This subject is supported by Korea Ministry of Environment (MOE) as "Advanced Technology Program for Environmental Industry".

7. REFERENCES

[1] W. Choi, S. Kim, M. Keum, D. K. Han, and H. Ko, "Acoustic and visual signal based context awareness system for mobile application," *IEEE Trans. Consum. Electron.*, vol. 57, no. 2, pp. 738-746, 2011.

[2] K. Yamano and K. Itou, "Browsing audio life-log data using acoustic and location information," in *IEEE Int. Conf. Mobile Ubiquitous Computing, Systems, Services and Technologies*, pp. 96-101, 2009.

[3] T. Nishiura, S. Nakamura, K. Miki, and K. Shikano, "Environmental sound source identification based on hidden markov model for robust speech recognition," in *EUROSPEECH* 2003, pp. 2157-2160, 2003.

[4] T. Heittola, A. Mesaros, A. Eronen, and T. Virtanen, "Context-dependent sound event detection," *EURASIP Journal Audio, Speech, and Music Processing*, pp. 1-13, 2013.

[5] A. J. Eronen, V. T. Peltonen, J. T. Tuomi, A. P. Klapuri, S. Fagerlund, T. Sorsa, G. Lorho, and J. Huopaniemi, "Audio-based context recognition," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 14, no. 1, pp. 321-329, 2006.

[6] V. Carletti, P. Foggia, G. Percannella, A. Saggese, N. Strisciuglio, and M. Vento, "Audio surveillance using a bag of aural words classifier," in *IEEE Int. Conf. on Advanced Video and Signal Based Surveillance*, pp. 81-86, 2013.

[7] J.-J. Aucouturier, B. Defreville, and F. Pachet, "The bag-of-frames approach to audio pattern recognition: A sufficient model for urban soundscapes but not for polyphonic music," *The Journal of the Acoustical Society of America*, vol. 122, no. 2, pp. 881-891, 2007.

[8] S. Pancoast and M. Akbacak, "Bag-of-Audio-Words approach for multimedia event classification," in *INTERSPEECH 2012*, pp. 2105-2108, 2012.

[9] H. E. Zadeh, B. Lehner, M. Dorfer, and G. Widmer, "A hybrid approach with multi-channel i-vectors and convolutional neural networks for acoustic scene classification," arXiv preprint arXiv:1706.06525, 2017.

[10] S. Mun, S. Park, Y. Lee, and H. Ko, "Deep neural network bottleneck feature for acoustic scene classification," *DCASE2016 challenge technical report*, 2016

[11] G. Takahashi, T. Yamada, S. Makino, and N. Ono, "Acoustic scene classification using deep neural network and frame-concatenated acoustic feature," *DCASE2016 challenge technical report*, 2016.

[12] M. Valenti, A. Diment, G. Parascandolo, S. Squartini, and T. Virtanen, "DCASE2016 acoustic scene classification using convolutional neural networks," *DCASE2016 challenge technical report*, 2016.

[13] J. Kim and K. Lee, "Empirical study on ensemble method of deep neural networks for acoustic scene classification," *DCASE2016 challenge technical report*, 2016

[14] T. H. Vu and J. C. Wang, "Acoustic scene and event recognition using recurrent neural networks," *DCASE2016 challenge technical report*, 2016.

[15] M. Zohrer and F. Pernkopf, "Gated recurrent networks applied to acoustic scene classification and acoustic event detection," *DCASE2016 challenge technical report*, 2016.

[16] S. H. Bae, I. Choi, and N. S. Kim, "Acoustic scene classification using parallel combination of LSTM and CNN," *DCASE2016 challenge technical report*, 2016.

[17] A. Mesaros, T. Heittola, A. Diment, B. Elizalde, A. Shah, E. Vincent, B. Raj, and T. Virtanen, "DCASE 2017 challenge setup: tasks, datasets and baseline system," *in Proc. Detection and Classification of Acoustic Scenes and Events 2017 Workshop (DCASE2017)*. Nov. 2017. submitted

[18] T. Irino, "A compressive gammachirp auditory filter for both physiological and psychophysical data," *J. Acoust. Soc. Am.*, vol. 109, no. 5, pp. 2008-2022, May 2001.

[19] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. Int. conf. Machine Learning*, vol. 37, pp. 448-456, 2015.

[20] http://www.cs.tut.fi/sgn/arg/dcase2017/.

[21] S. Park, S. Mun, Y. Lee, and H. Ko, "Score fusion of classification systems for acoustic scene classification," *DCASE2016 challenge technical report*, 2016.

[22] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Trans. Aud. Spee. Lang. Proc.*, vol 19, no. 4, pp. 788-798, May 2011.