Score Fusion of Classification Systems for Acoustic Scene Classification

Sangwook Park[1)]*, Seongkyu Mun*[2)]*, Younglo Lee*[1)]*, and Hanseok Ko,*[1,s)]

Korea University
[1)] School of Electrical Engineering, [2)] Department of Visual Information Processing,
Anam-dong, Seongbuk-gu, Seoul, 136-713, Korea
{swpark, skmoon, yllee}@ispl.korea.ac.kr, and hsko@korea.ac.kr

ABSTRACT

This is a technical report of our study on the acoustic scene classification task of the IEEE AASP Challenge: Detection and Classification of Acoustic Scenes and Events. In order to accomplish the ensuing task, we explore several methods in three aspects; feature extraction, generative/discriminative machine learning, and score fusion for final decision. For finding an appropriate frame-based feature, a new feature is devised after investigating several features. Subsequently, those models based on both generative and discriminative learning are applied for classifying the feature. From these studies, several system designs composed of a combination of the features and classifiers are considered and incorporated. The final result is determined by fusing the individual results. Experimental results are summarized and concluding remarks of this report are presented.

Index Terms— One, two, three, four, five

1. INTRODUCTION

Among acoustic signal analysis tasks, Acoustic Scene Classification (ASC) is one of the most formidable tasks in terms of complexity since it requires sophisticated understanding of individual acoustic events. Recently, the first task of the IEEE AASP Challenge: Detection and Classification of Acoustic Scenes and Events (DCASE2016) considers 15 acoustic 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* [1]. It is a challenging task to classify these situations because several scenes have similar background sounds, and key events giving help with scene classification are sometimes ignored due to very short duration or masked by background sound. Besides, features extracted from a common situation are spread in observation space, because the situation as like *café/restaurant* or *train* includes various spectrums according to individual cases.

Our classification system consists of three steps; cepstral feature extraction, generative and discriminative learning, and score fusion. In the first step, several frame based features are investigated, and a new feature is designed for the ASC task. In the next step, two types of classifying method are considered. One is a Gaussian Mixture Model (GMM) which is directly applied to the features obtained in the previous part. The other one is Nearest Neighbor (NN) which is used in features obtained by applying a discriminative kernel function. In the last step, a final result is determined by fusing these results obtained from the GMM and NN systems for improving classification performance.

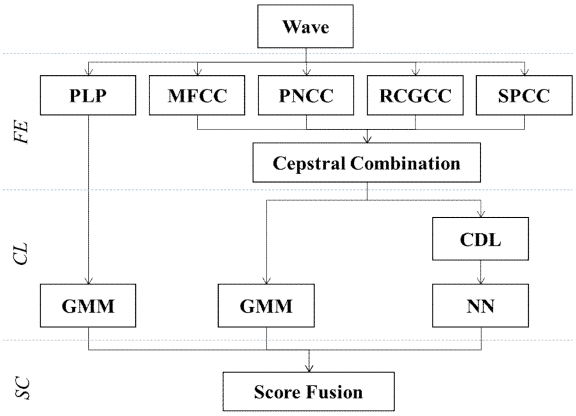


Figure 1: System architecture for acoustic scene classification.

The remainder of this report is organized as follows. Section 2 provides a description on our studies about three steps. Section 3 describes about experiments and discussions. Finally, concluding remarks of this report are presented in Section 4.

2. OUR SYSTEM FOR ACOUSTIC SCENE CLASSIFICATION

Figure 1 shows our system for acoustic scene classification. The system is composed of three parts; feature extraction *(FE)*, classification *(CL)*, and score fusion *(SC)*. As shown in the figure, final result is determined based on the results from three systems; Perceptual Linear Prediction (PLP) - GMM, Cepstral Combination (CepsCom) - GMM, and CepsCom – NN with Covariance Discriminative Learning (CDL). These methods are described in the following.

2.1. Feature Extraction (*FE*)

First of all, several features which are widely used in this field are investigated. As baseline features, Mel-Frequency Cepstral Coefficients (MFCCs) and PLP are considered. These features are already used for recognitions with human voices as well as acoustic sounds, and have shown noticeable performance in many research efforts.

In [2], Power Normalized Cepstral Coefficients (PNCCs) was proposed by studying a human hearing characteristic, and adding power bias subtraction procedure for noise robust speech recognition. Similarly, Robust Compressive Gamma-chirp filterbank Cepstral Coefficients (RCGCCs) was proposed in the same aspect of PNCC [3]. When these features are used for ASC, the

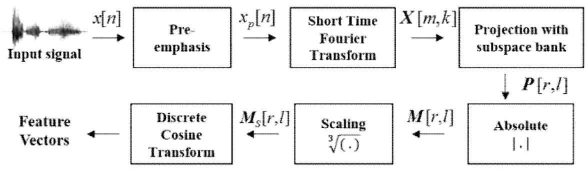

Input signal
$x[n]$
Pre-emphasis
$x_p[n]$
Short Time Fourier Transform
$X[m,k]$
Projection with subspace bank
$P[r,l]$
Absolute $|\cdot|$
$M[r,l]$
Scaling $\sqrt[3]{(\cdot)}$
$M_S[r,l]$
Discrete Cosine Transform
Feature Vectors

Figure 2: Block diagram for extracting SPCC

features capture varying spectrum owing to some acoustic events rather than background sounds. In this regard, the features can be of a substantial help to accomplishing the ASC task.

In contrast, a new feature is designed for capturing non-varying spectrum as like background sound. The new feature named by "Subspace Projection Cepstral Coefficients (SPCCs)" is obtained by projecting onto each target subspace. Figure 2 shows a block diagram for obtaining the new feature. In the figure, the subspace bank includes all target subspaces obtained by applying subspace learning to training data.

In our system, PLP and Cepstral Combination (CepsCom) obtained by concatenating MFCC, PNCC, RCGCC, and SPCC are used for classification.

2.2. Classification (*CL*)

Several classifiers such as GMM, Hidden Markov Model (HMM), Support Vector Machine (SVM), NN, Artificial Neural Network (ANN), and Deep Neural Network (DNN) are applied for ASC during the past decade. Among them, a GMM is firstly applied for classifying PLP and CepsCom features.

Meanwhile, almost all classifiers may be appropriate for classifying feature vectors in Euclidean space. In this ASC task, however, these classifiers face limitation to classifying the feature vectors because the feature vectors extracted from real acoustic scenes lie in a Riemannian manifold.

R. Wang, et al., have derived a kernel function that maps a covariance matrix of feature vector from the Riemannian manifold to a Euclidean space, and applied discriminative learning for improving discriminant characteristics [4]. In order to overcome the limitation, we adopted the third method that use the kernel function proposed by R. Wang, et al. In the third system, after the CDL is applied to CepsCom, then NN classification is performed.

2.3. Score Fusion (*SF*)

In our system, the final result is obtained by fusing scores from three systems; PLP-GMM, CepsCom-GMM, and CepsCom-CDL. Likelihood obtained by using GMM and distance obtained by using NN are used for score. These scores are firstly normalized within interval [0, 1], and summed according to system weight for obtaining final score s_c in (1).

$$s_c = \sum_{i=1}^{N} w_c^i s_c^i \quad (1)$$

$$w_c^i = P(X = C_c \mid Y = C_c) \quad (2)$$

where N is the number of classification system, in this case N=3. s_c^i is the score calculated in the i th system under class C_c, and w_c^i is the i th system weight under class C_c. The weight is a confidence of system result, and it can be estimated in (2). X and Y are random variables that mean a ground truth and a system result, respectively. If other classes are not confused with class C_c, the weigh may be closed to 1. Otherwise, since the result cannot be guaranteed, the weight may be 0. Then, final result can be determined a class that has the largest score.

3. EXPERIMENTS

3.1. Database and Experiment setting

For performance assessment, our system was performed with development dataset provided by DCASE 2016 organizer [5]. For conducting a cross-validation test, the dataset was divided into 8 subsets and then assigned as training and test set with a ratio of 1:3, because an amount of training data is typically limited in comparison with test cases in a real situation [6, 7]. (Note that all subsets, i.e. development dataset, are used for training system when evaluation dataset is evaluated)

For feature extraction, frames were defined as 2048 samples with an overlap with the next frame for 1024 samples. Discrete Fourier Transform (DFT) was conducted with 2048 points hamming window. In case of PLP, 39 coefficients including delta, acceleration and energy coefficients were extracted by using HTK [8]. In the others; MFCC, PNCC, RCGCC, and SPCC, 60 coefficients including delta and acceleration coefficients were extracted. Note that CepsCom composed of MFCC, PNCC, RCGCC, and SPCC was a 240 dimensional vector.

3.2. Experiment results

In the first experiment, classification performances of each feature are evaluated under GMM classifier. The results that show mean average classification rate for 15 classes and all possible combinations for cross validation test are summarized in Figure 3. The result of MFCC is 67.15% when 128-mixture model is

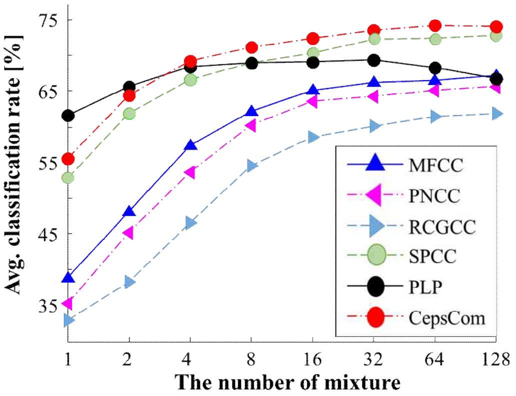
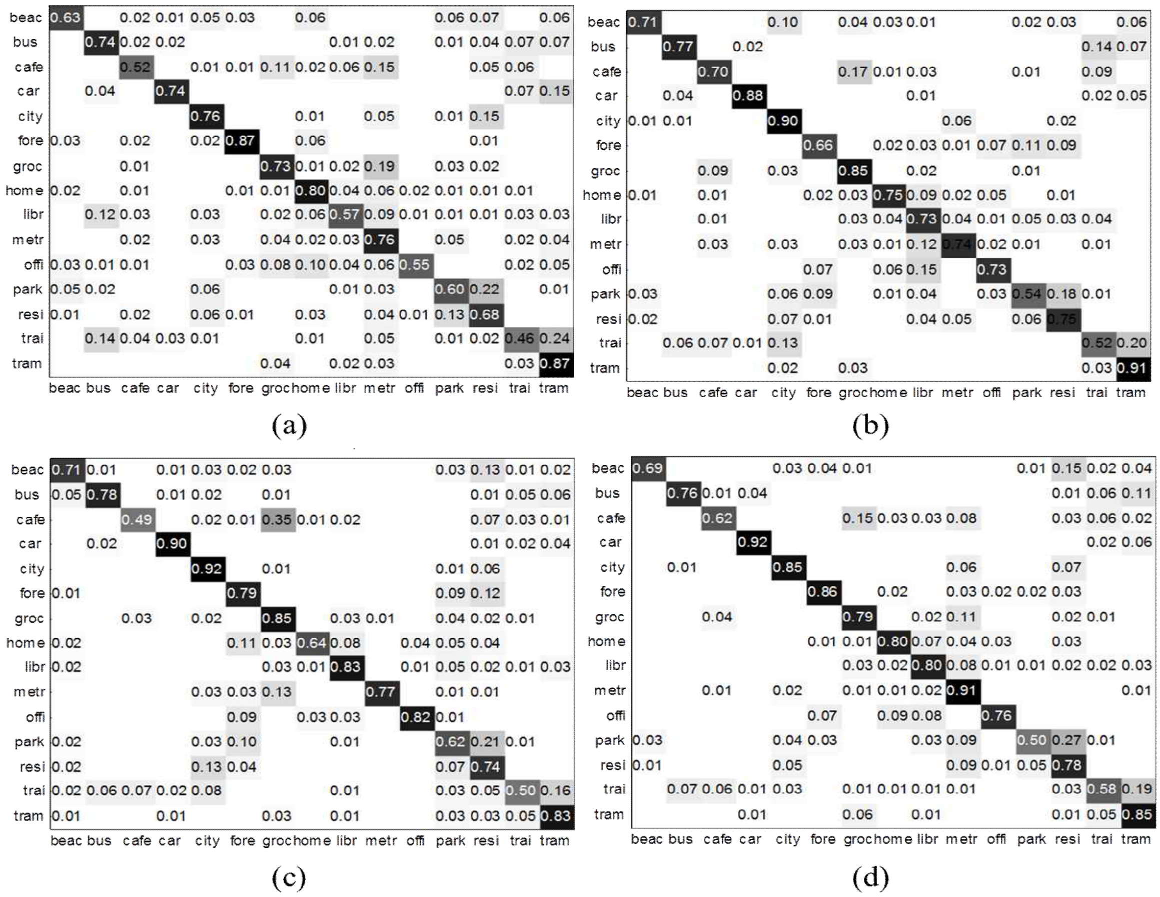


Figure 3: Average classification rate according to frame based features. The experiments were conducted by using the HTK.



(a) (b) (c) (d)

Figure 4: Confusion matrices; (a) PLP-GMM with 4 mixtures (b) CepsCom-GMM with 64 mixtures (c) CepsCom- NN with CDL (d) Score fusion

applied. The results of PNCC and RCGCC are less than MFCC's performance in all cases, whereas SPCC and PLP outperform MFCC. The result of PLP is 68.43% in case of 4 mixture model. When 64 mixture model is used for CepsCom, the feature shows the best performance, 74.15%, among considered features. Since CepsCom has advantages of each feature, the combination feature can outperform than others.

As mentioned previously, feature vectors lied in Riemannian manifold are mapped to Euclidean space by applying a kernel function proposed in [4]. All considered features are considered for this approach. Among them, CepsCom feature shows the best performance as 74.62%.

Figure 4 shows confusion matrices of PLP-GMM, CepsCom-GMM, CepsCom-CDL, and a final result. As shown in the figure, PLP-GMM system shows a superior classification in *forest* and *tram*, and CepsCom-GMM system shows a superior performance in *car*, *city*, *grocery* and *tram*. In case of CepsCom-CDL system, better performances compared to CepsCom-GMM system can be found in *forest*, *library*, and *office*. In this aspect, performance can be improved by fusing these results. As a result of score fusion, average classification rate shows 76.35%, and its confusion matrix is shown in Figure 4 (d). In the final result, average classification rate is improved as 2.25% compared to CepsCom-GMM system.

4. CONCLUSIONS

This report described about the approaches applied in an ASC of the IEEE AASP Challenge: DCASE 2016. For the ASC task, we investigated several cepstral features such as MFCC, PLP, PNCC, and RCGCC, and designed a new feature, SPCC. These features were used for scene classification by means of GMM for performance evaluation. Also, CDL was applied for improving performance, and the feature applied CDL was classified with NN criteria. From the fact that each class has different classification rate according to classification system, final result was determined by fusing individual results, PLP-GMM, CepsCom-GMM, and CepsCom-CDL. Experiment results were summarized with confusion matrix. From these experimental results, our approaches is expected to show good performance in experiment with the evaluation dataset.

5. ACKNOWLEDGMENT

This research was supported in part by the BK2lPIus Program and in part by Korea University Grant

6. REFERENCES

[1] http://www.cs.tut.fi/sgn/arg/dcase2016/.

[2] C. Kim and R. M. Stern, "Feature extraction for robust speech recognition using a power-law nonlinearity and power-bias subtraction," in *Proc. INTERSPEECH*, 2009, pp. 28-31.

[3] M. J. Alam, P. Kenny, and D. O'Shaughnessy, "Robust feature extraction based on an asymmetric level-dependent auditory filterbank and a subband spectrum enhancement technique," *Digital Signal Processing*, vol. 29, pp. 147-157, 2014

[4] R. Wang, H. Guo, L. S. Davis, and Q. Dai, "Covariance Discriminative Learning: A Natural and Efficient Approach to Image Set Classification," in *Proc. IEEE CVPR*, 2012, pp. 2496-2503.

[5] A. Mesaros, T. Heittola, and T. Virtanen, "TUT database for acoustic scene classification and sound event detection," in *Proc. EUSIPCO*, 2016.

[6] S. Park, W. Choi, and H. Ko, "Frequency-cepstral features for bag of words based acoustic context awareness," *The Journal of Acoustical Society of Korea* (in Korean), vol. 33, no. 4, pp. 248-254, 2014.

[7] S. Park, W. Choi, and H. Ko, "Acoustic scene classification using recurrence quantification analysis," *The Journal of Acoustical Society of Korea* (in Korean), vol. 35, no. 1, pp. 42-48, 2016.

[8] *The HTK book Version 3.4*, Cambridge University Engineering Department, (2009).