Deep Neural Network Bottleneck Feature for Acoustic Scene Classification

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

Korea University
[1] Department of Visual Information Processing, [2] School of Electrical Engineering
Anam-dong 5-ga, Seongbuk-gu, Seoul, 136-713, Korea
{skmoon, swpark, yllee}@ispl.korea.ac.kr, and hsko@korea.ac.kr

ABSTRACT

Bottleneck features have been shown to be effective in improving the accuracy of speaker recognition, language identification and automatic speech recognition. However, few works have focused on bottleneck features for acoustic scene classification. This report proposes a novel acoustic scene feature extraction using bottleneck features derived from a Deep Neural Network (DNN). On the official development set with our settings, a feature set that includes bottleneck features and Perceptual Linear Prediction (PLP) feature shows a best accuracy rate.

Index Terms— Deep neural network, Bottleneck feature, Dimension reduction, feature extraction, deep belief network

1. INTRODUCTION

Among acoustic signal analysis tasks, Acoustic Scene Classification (ASC) is a challenging task since several real life acoustic scenes have similar background sounds. It has been a key issue for research efforts to find the discriminative features from similar background sounds [1-3].

In order to extract the discriminative acoustic characteristics from these acoustic scenes, this report proposes to use BottleNeck (BN) features. Bottleneck features are widely used for low information loss nonlinear feature transformation and dimensionality reduction method in speech recognition, speaker recognition and language identification [4-6]. However, few works have focused on bottleneck features for acoustic scene classification [7]. Hence, this report proposes an acoustic scenes classification system using a bottleneck feature framework for extracting discriminative features from various features.

On the official development set with our settings, the highest accuracy rate was achieved by the BN feature set using hierarchical recognition system consists of full covariance matrix Gaussian Mixture Model (GMM) and Supported Vector Machine (SVM).

2. PROPOSED SYSTEM FOR ACOUSTIC SCENE CLASSIFICATION

The process of proposed system based on bottleneck features is depicted in Figure 1. Before extracting bottleneck features, MFCC and PLP features with delta, acceleration and third differential coefficients are combined. The bottleneck features are extracted using DBN (Deep Belief Network)-DNN with input features (Figure 1-A). Figure 1-B shows an acoustic scene classifier. The bottleneck features extracted from DNN are then input alone or concatenated with the MFCC or PLP features to train the final acoustic classifier.

2.1. Bottleneck features using DBN-DNN

Bottleneck features were extracted from a DNN [7] in which one of the internal layers has a small number of hidden units relative to the size of the other layers. The DNN to extract the bottleneck features is shown in Figure 1. The DNN configuration 256-256-14(BN)-128-15 was used. MFCC and PLP features (C0 included) with delta-acceleration-third diff. (13 x 4 x 2 dim.) were extracted from a frame (total 104 dimensional vector) and was used as input layer of bottleneck feature extractor. In this report, the number of hidden layers (including the bottleneck layer) is set to 4. The number of hidden units in the innermost layer is smaller than those in the other layers. This layer is called the bottleneck layer.

In the pre-training step, we trained each layer of the RBM (Restricted Boltzmann Machine) (Gaussian-Bernoulli RBM for first layer) to construct a DBN using the common DBN training [8]. With the pre-training step, the DBN achieved better initial values of the neural network. This structured bottleneck layer could be treated as a nonlinear mapping of input features.

After the pre-training step, we used the acoustic scene labels as the target signal. DNN's can be trained by back propagating derivatives of a cost function that measures the cross entropy between the target outputs and the actual outputs produced for each training case. After supervised training, last two layers and activation function of bottleneck layer were removed. Finally, the bottleneck features extracted from the bottleneck layer were used to train the acoustic scene classifier.

2.2. Classification

Several classifiers such as GMM, SVM and DNN are applied for ASC during the past decade [1-3]. Among them, on the official development set with our settings, the highest accuracy rate was achieved by using a hierarchical recognition system composed of full covariance matrix GMM and SVM. The GMM was used as main classifier and SVM was used for classifying confusing class pairs, such as 'Home vs. Library' and 'Park vs. Residential area'. Two SVMs are trained for each class pair and used for post-processing of GMM results. SVMs are only used for situations when 2-best of GMM outputs are SVM correspond classes, such as 'Home and Library' or 'Park and Residential area'.

3. EXPERIMENTS

3.1. Database and Experiment setting

For performance assessment, our system was performed with development dataset provided by DCASE 2016 organizer. The dataset was down-sampled to 16 KHz and provided stereo wave files are divided into 3 files, left data only, right data only and 50% mixed.

For conducting a cross-validation test, the dataset was assigned as training and test set with a ratio of 1:3 (4-folds), since the amount of training data is typically limited in comparison with test cases in real situation [6, 7]. Note that all subsets, i.e. development dataset, are used for training system when evaluation results are processed from the provided evaluation dataset.

For feature extraction, frame length was defined as 0.025 [sec] with 50% overlap. PLP/MFCC features with 52 coefficients including delta, acceleration and third differential coefficients were extracted by using HTK [8]. PCA or LDA was applied to the BN feature for feature de-correlation.

3.2. Experiment results

Among the various results from parameter tuning, each approach's best performance cases are shown in Table 1. In the feature comparison experiments, an input feature set includes BN feature and PLP feature shows best performance. The full covariance GMM with SVM post-classifier shows best performance in the classifier performance comparison experiments.

4. CONCLUSIONS

This report described about the approaches applied to the ASC task of the IEEE AASP Challenge: DCASE 2016. For the ASC task, BN features derived from DNN were proposed and applied to the hierarchical recognition system of full covariance matrix GMM and SVM. Additional work will investigate improved methods for finding effective DNN bottleneck structure and optimizing learning algorithm. These are the key issues for applying the proposed framework in large scale acoustic scene classification applications.

REFERENCES

[1] W. Choi, S. Park, D. K. Han and H. Ko, "Acoustic event recognition using dominant spectral basis vectors", INTERSPEECH 2015, Proceedings, pp. 2002–2006, 2015.

[2] C. Ludena and A. Gallardo "Acoustic event classification using spectral band selection and non-negative matrix factorization-based features" Expert Systems with Applications, vol. 46, pp. 77-86, 2016.

[3] J. Dennis, et al., "Temporal coding of local spectrogram features for robust sound recognition", IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 803-807, 2013.

[4] D. Yu and L. Michael, "Improved bottleneck features using pretrained deep neural networks", INTERSPEECH 2011 – 12th Annual Conference of the International Speech Communication Association, Proceedings, pp.240-243, 2011.

[5] S. Yaman, J. Pelecanos and R. Sarikaya, "Bottleneck features for speaker recognition", IEEE Odyssey'12, pp. 105–108, June 25-28, Singapore, 2012.

[6] P. Matejka, et al., "Neural network bottleneck features for language identification", IEEE Odyssey'14, pp. 299-304, Joensuu, Finland, 2014

[7] S. Mun, S. Shon, W. Kim and H. Ko, "Deep neural network bottleneck features for acoustic event recognition", INTERSPEECH 2016 (accepted).

[8] 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.

[9] 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.

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

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

Table 1. Average classification rate [%] in features and classifier comparison experiments

Input feature set	# of dim.	[%]
GMM-SVM classifier (64 mixtures with Full-covariance matrix GMM)		
MFCC-Energy-Delta-Acc.-Third.	52	70.8
PLP-Energy-Delta-Acc.-Third.	52	71.2
BN (PCA)	14	71.4
BN (PCA) + MFCC	66	72.0
BN (PCA) + PLP	**66**	**72.7**
BN (LDA)	14	71.6
BN (LDA) + MFCC	66	72.0
BN (LDA) + PLP	66	72.3
DNN-SVM classifier (Hidden layer nodes : 256-256-256-256)		
MFCC-Energy-Delta-Acc.-Third.	52	71.5
PLP-Energy-Delta-Acc.-Third.	52	72.1
BN (PCA)	14	71.6
BN (PCA) + MFCC	66	71.9
BN (PCA) + PLP	66	72.2
BN (LDA)	14	71.6
BN (LDA) + MFCC	66	72.1
BN (LDA) + PLP	66	72.3

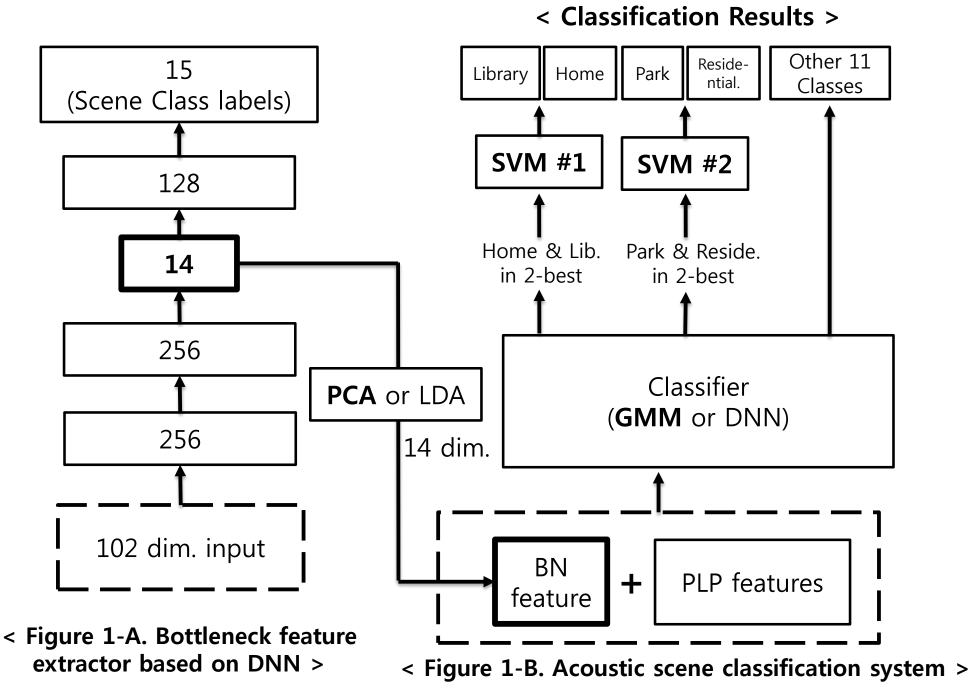

Figure 1: Proposed acoustic scene classification framework based on bottleneck features.