An Investigation of High-Resolution Modeling Units of Deep Neural Networks for Acoustic Scene Classification

Xiao Bao, Tian Gao, Jun Du and Li-Rong Dai
National Engineering Laboratory for Speech and Language Information Processing,
University of Science and Technology of China, Hefei, Anhui, China
Email: baox@mail.ustc.edu.cn, gtian09@mail.ustc.edu.cn, jundu@ustc.edu.cn, lrdai@ustc.edu.cn

***Abstract*—In this paper, we investigate high-resolution modeling units of deep neural networks (DNNs) from concrete to abstract for acoustic scene classification based on Gaussian mixture model (GMM) and ergodic hidden Markov model (HMM). A direct modeling strategy for DNN to classify acoustic scenes is to map each frame feature of an audio to one scene category. However, all frames tagged with the same label may not be the best choice because the representative pattern of an audio is sparse. GMM is also often employed to model each acoustic scene directly as a generative model. Because the multiple Gaussians in a GMM model have different levels of contribution, and each Gaussian can be seen as a subclass of the scene category, so we can utilize the subclass of GMM as a bit abstract modeling unit to adopt DNN-GMM system. When single scene category is subdivided into various subclasses, prior scores for each subclass calculated from training set are stored as one part of model to response the sparseness of representative pattern. Ergodic HMM should be more appropriate to model the acoustic scenes than GMM due to the uncertain structure of scene audio. Using HMM states as modeling units, we build DNN-HMM hybrid system. By comparison, we find high-resolution modeling units are more effective than direct modeling. The final system is obtained by performing system combination to take advantage of the complementarity of different-level modeling units. Experiments on acoustic scene classification task of DCASE2016 challenge show that our final system yields 25.9% relative error rate reduction compared with a GMM baseline on evaluation set.**

I. Introduction

Sounds carry a large amount of information about our everyday environment and physical events that take place in it. Humans can perceive the sound scene we are within (busy street, office, etc.), and recognize individual sound sources (car passing by, footsteps, etc.). This process is called auditory scene analysis [1]. The research field studying this process is called computational auditory scene analysis (CASA) [2]. The computational algorithms attempt to automatically make sense of the environment through the analysis of sounds using signal processing and machine-learning methods. The task is the so-called acoustic scene classification [3], and the goal is to classify a test recording into one of predefined classes that characterizes the environment in which it was recorded – for example ”bus”, ”office”, ”home” as shown in Fig. 1.

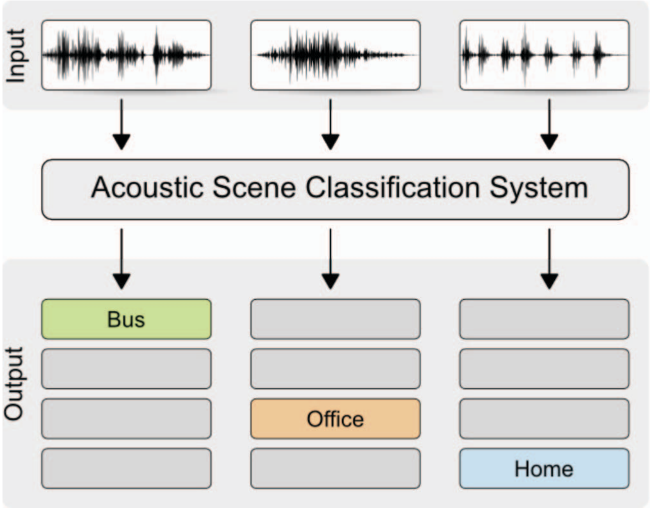


Fig. 1. Overview of acoustic scene classification system [4].

Over the last few years, acoustic scene classification has been gradually receiving attention in the field of audio signal processing and machine learning. Substantial progress has been made by several important challenges, such as DCASE2013 (Detection and Classification of Acoustic Scenes and Events) [5], [6] and DCASE2016 [4]. Many techniques have been widely investigated, including the aspects of features, statistical models, decision criteria and meta-algorithms. Several categories of audio features have been employed in acoustic scene classification systems, such as low-level time-based and frequency-based audio descriptors [7], [8], frequency-band energy features (energy/frequency) [7], auditory filter banks (Gammatone, Mel filters), cepstral features (MFCC), spatial features (ITD: interaural time difference, ILD: interaural level difference) [9], voicing features (fundamental frequency f_0) [10] and i-vector [11]. The features described here can be further processed to derive new quantities that are used either in place or as an addition to the original features, like PCA and time derivatives [12].

Once the features are extracted from the audio samples, the next stage is learning statistical models of the distribution of the features. Statistical models can be divided into generative and discriminative methods. When working with generative

978-1-5090-6182-2/17/$31.00 ©2017 IEEE

models, feature vectors are interpreted as being generated from one of a set of underlying statistical distributions. During the training stage, the parameters of the distributions are optimized based on the statistics of the training data. In the test phase, a decision criterion is defined to determine the most likely model that generated a particular observed example. One classical generative model for acoustic scene classification is the Gaussian mixture models (GMM) [13] where features are interpreted as being generated by a sum of Gaussian distributions. MFCC features and maximum likelihood criterion are used for GMM training and testing. Hidden Markov models (HMM) [14] is also used in several systems [7], [15] to account for the temporal unfolding of events within complex soundscapes.

As for discriminative models, SVM is a popular discriminative classifiers for acoustic scene classification [9], [10]. The model output from an SVM determines a set of hyperplanes that optimally separate features associated to different classes in the training set (according to a maximum-margin criterion). An SVM can only discriminate between two classes. When the classification problem includes more than two categories, multiple SVMs can be combined to determine a decision criterion that allows for discrimination between multiple classes.

The recent breakthrough of deep learning [16], [17], [18], and the applications of deep neural networks (DNNs) in classification tasks [19], [20], [21], creates a new direction of acoustic scene classification. A straightforward way of deep learning to classify acoustic scenes is to map each frame feature of an audio to one scene category directly. Followed with the direct modeling, post-processing is employed to decide which scene category the audio belonged to. In [22], [23], the deploy of DNN was investigated for acoustic scene classification. The input of DNN is acoustic features like MFCC or Mel-filterbank. The learning target of each frame feature is pre-defined scene category. Recently, convolutional neural networks (CNN) [24] was also employed to classify acoustic scenes directly.

In this paper, we focus on the exploration of modeling units of DNN. We investigate high-resolution modeling units from concrete to abstract for acoustic scene classification based on GMM and ergodic HMM. Through the review of previous work, we know a direct modeling unit of DNN to classify acoustic scenes is scene category. For the direct modeling, all frame features of an audio have the same label. However, all frames tagged with one label may not be the best choice because the representative pattern of an audio is sparse. We know GMM is also often used to model each acoustic scene directly as a generative model. The multiple Gaussians in GMM have different levels of contribution. If we locate each frame feature to a single Gaussian distribution with maximum likelihood among all Gaussians in the GMM, we can use the single Gaussian as a bit abstract modeling unit to adopt DNN-GMM system. Each Gaussian in the GMM model can be seen as a subclass of the scene category. When single scene category is subdivided into various subclasses, the distribution of labels is no longer uniform. Prior scores for each subclass calculated from training set are stored as one part of model to response the sparseness of representative pattern. Due to the representative pattern of an audio for scene classification is also unordered, ergodic HMM should be more appropriate to model acoustic scenes than GMM. In paper [15], ergodic HMM was proved to have better modeling capacity than GMM for indefinite duration classes. The HMM states can be explained as the results of automatic clustering for each scene. Using states as modeling units, we build DNN-HMM hybrid system, where DNN models the scaled observation likelihood of all HMM states, and the ergodic HMM models the sequential property of the observation. Finally, a system combination method is performed to obtain the final system from the multiple systems to take advantage of the complementarity of different-level modeling units. Experiments on acoustic scene classification task of DCASE2016 challenge indicate that the accuracy on evaluation set was improved from 77.2% for the GMM baseline system to 83.1% for our final system, and to 83.3% for the best single system DNN-GMM.

The rest of the paper is organized as follows. In Section II, we first give an overview of our proposed system. In Section III, direct modeling of DNN and GMM are described in detail. In Section IV and Section V, DNN-GMM and ergodic DNN-HMM modeling are presented. In Section VI, we report experimental results and analysis. Finally we summarize our findings in Section VII.

II. System overview

The overall flowchart of our proposed system is illustrated in Fig. 2. Our system has three parts with different-level modeling units of DNN, including direct modeling, DNN-GMM modeling and ergodic DNN-HMM modeling. Before model training, the audio samples are processed to extract MFCC and log Mel-filterbank (FBANK) features.

For direct modeling part, DNN is trained with FBANK features as input and acoustic scene labels as learning target. In the testing stage, the final decision for an audio is taken by first averaging the output of the neural networks for each input frame feature which forms the audio and next choosing the scene class with the best result.

For DNN-GMM modeling part, GMM models are first trained with MFCC features for each acoustic scenes. Then, we locate each frame feature to a single Gaussian distribution with maximum likelihood among all Gaussians in the GMM. In this way, each frame feature of an audio can has a subclass label, which is no longer a unified scene category label. We use FBANK features as input and the subclass labels as learning targets to train DNN-GMM.

For ergodic DNN-HMM modeling part, ergodic HMMs are first trained using MFCC features per acoustic scene class, and perform classification with maximum likelihood classification scheme. Next, the state labels are generated from ergodic HMM models by force alignment [14]. Based on FBANK features as input and HMM states as learning targets, DNN is trained to build ergodic DNN-HMM hybrid system.

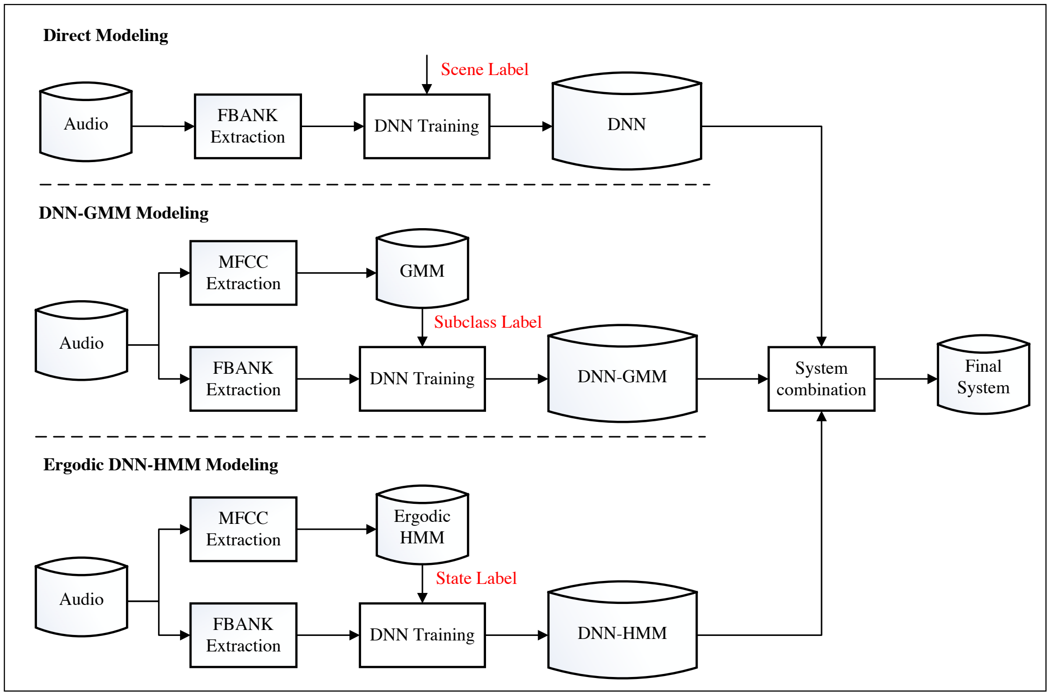


Fig. 2. System overview.

Finally, system combination is implemented to utilize the complementarity of multiple systems. We use a voting strategy at audio level to combine DNN, DNN-GMM and DNN-HMM systems. The voting strategy follows majority decision. All HMM-based and DNN-based experiments are implemented by using Kaldi toolkit [25]. GMM system is provided by DCASE2016 challenge [13].

III. Direct Modeling

In this section, we introduce the implementation of DNN and GMM for acoustic scene classification using acoustic scene labels as direct modeling units.

A. DNN

DNN is a popular neural network architecture which has been employed for classification tasks successfully. More precisely, when the objective is to classify a feature of interest $\mathbf{x}$ among Q classes, a DNN estimates the probabilities p_j, $j \in \{1...Q\}$, of each class given the input $\mathbf{x}$. The input features usually correspond to a time-frequency representation of the input signal, such as MFCC, Mel-filterbank. To provide acoustic context to the DNN, consecutive feature frames are concatenated in a sliding window approach. A graphical representation of DNN architecture for classification is given in Fig. 3.

For a H-layer DNN, it computes a non-linear function

$$g_W(\mathbf{x}) = g_H\left(\mathbf{W}_H g_H\left(\mathbf{W}_{h-1} \cdots g_1\left(\mathbf{W}_1 \mathbf{x}\right)\right)\right), \qquad (1)$$

where $\mathbf{x}$ are the input features, $g_h, h = 1, \ldots, H$ are *activation functions* and $\mathbf{W}_h, h = 1, \ldots, H$ are *DNN weights*. The layers with index $h = 1, \ldots, H-1$ are called *hidden layers*, so that the DNN from (1) is said to have $H-1$ hidden layers. The *hidden layer activations* in this paper are sigmoid functions. The layer with index H is called *output layer* with activations g_H. For classification task, the output is computed via the softmax activation function.

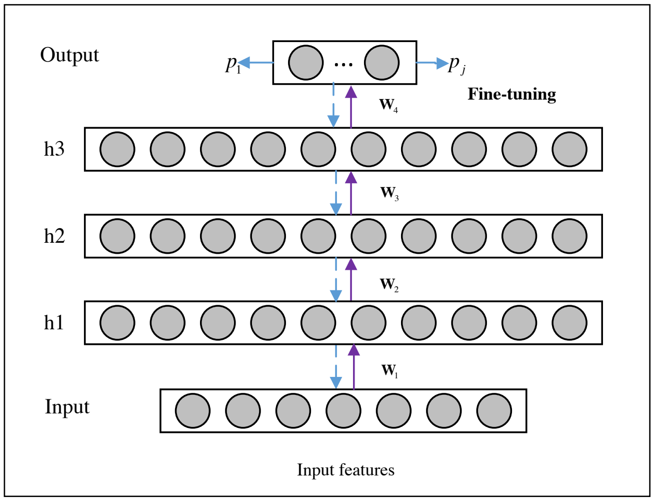


Fig. 3. A DNN for classification is described by an input, a given number of hidden layers and an output which describes the class probabilities.

The *DNN weights* of a DNN are trained by minimizing

Cross-Entropy cost function:

$$C = -\sum_{j=1}^{Q} q_j \log p_j \tag{2}$$

where C denotes Cross-Entropy cost function, p_j is the output of the softmax, and q_j is the corresponding target.

B. GMM

GMM is also often adopted as a generative model for acoustic scene classification. A GMM-based baseline system is provided as the official benchmark of DCASE2016 challenge, which consists of MFCC features and GMM based classifier [3], [13]. GMMs are used to infer global statistical properties of the features from local features vectors, which are interpreted as realizations of a generative stochastic process. For each acoustic scene, a GMM class model with N_g Gaussians is trained based on MFCC features using expectation-maximization (EM) algorithm.

Once the GMMs have been trained from the training data, and features have been extracted from an unlabeled audio, maximum likelihood decision is employed to evaluate which GMM class is statistically most likely to generate the observed features, hence determining the scene classification.

IV. DNN-GMM Modeling

A. Training for DNN-GMM

A GMM class model with N_g Gaussians in Section III-B is trained for each acoustic scene. The N_g Gaussians have different level of contribution to each frame feature. If we locate each frame feature to a single Gaussian distribution with maximum likelihood among all Gaussians in the GMM, we can use the single Gaussian as a new subclass label to replace original scene label as follow,

$$S_{i^*}^{j}(\mathbf{x}) = \underset{i}{\operatorname{argmax}}\, L_i^j(\mathbf{x}) \tag{3}$$

where, $S_{i^*}^{j}(\mathbf{x})$ is the subclass label of input training feature $\mathbf{x}$, i^* is the index of Gaussian with maximum likelihood in the GMM model, j is the index of scene class, and $L_i^j(\mathbf{x})$ is the likelihood of ith Gaussian in jth scene class. The number of total subclasses is $Q * N_g$, where Q is the number of scene category.

In this work, we use neural networks to predict GMM subclasses. Neural networks are trained to discriminate the total $Q * N_g$ subclasses $(S_1^1, \ldots, S_i^j, \ldots, S_{N_g}^Q)$ with FBANK features as input, resulting to DNN-GMM model.

When single scene category is subdivided into various subclasses, the distribution of labels is no longer uniform. Because the representative pattern of an audio is sparse, so prior scores for each subclass calculated from training set are stored as one part of model to further improve system performance. We first calculate a prior score for each subclass at the level of audio. We assume that subclass S_i^j with the largest number of frames is the dominant subclass of the training audio $\mathbf{X}_n$. The prior score of S_i^j for audio $\mathbf{X}_n$ is defined as follow,

$$p(S_i^j, \mathbf{X}_n) = \begin{cases} N_i^j(\mathbf{X}_n)/N_f, & S_i^j \text{ is dominant subclass} \\ 0, & \text{else} \end{cases} \tag{4}$$

where, $\mathbf{X}_n$ is an audio in the training set, $N_i^j(\mathbf{X}_n)$ and N_f representing the number of frames of dominant subclass S_i^j and of audio $\mathbf{X}_n$, respectively. Then the prior score for each subclass at the training set level is calculated as follow,

$$p(S_i^j) = \frac{1}{N_t} \sum_{n=1}^{N_t} p(S_i^j, \mathbf{X}_n) \tag{5}$$

where, $p(S_i^j)$ is the prior score of subclass S_i^j in the training set, N_t is the number of training audios whose $p(S_i^j, \mathbf{X}_n)$ is non-zero.

B. Decoding for DNN-GMM

When a testing frame feature $\mathbf{x}$ is fed to the well trained neural networks, posterior probabilities of subclasses are generated. We reset each value to 1 or 0 with a threshold to get rid of distractions. $p(S_i^j|\mathbf{x})$ is used to represent the scaled posterior probabilities. Next, we multiply scaled posterior probability by corresponding prior score as follow,

$$d(S_i^j|\mathbf{x}) = p(S_i^j) p(S_i^j|\mathbf{x}) \tag{6}$$

where $d(S_i^j|\mathbf{x})$ is the decision score of subclass S_i^j given $\mathbf{x}$. The scene category which the input frame $\mathbf{x}$ belonged to is determined as follow,

$$\hat{q}(\mathbf{x}) = \underset{j}{\operatorname{argmax}} \sum_{i=1}^{N_g} d(S_i^j|\mathbf{x}) \tag{7}$$

where $\hat{q}(\mathbf{x})$ is the frame-level scene decision. Finally, the scene category of the whole testing audio is determined by using majority decision method based on frame-level results.

In this work, the system which used prior score in the testing stage is denoted as DNN-GMM, and the system which didn't use prior score is denoted as DNN-GMM-NP.

V. Ergodic DNN-HMM Modeling

A. Ergodic HMM

HMM is an effective parametric representation for a time-series of observations, such as feature vectors measured from natural sounds. Left-to-right HMM has been successfully used for speech recognition. Ergodic HMM is more suitable for scene classification due to the uncertain structure of scene audio [7], [15]. In this work, ergodic HMMs are used for classification by training an ergodic GMM-HMM for each scene class with MFCC features. Each ergodic GMM-HMM with a set of states represents one scene class. The parameters of GMM-HMMs for all classes are learned according to the maximum likelihood estimation (MLE). By adopting the EM and Baum-Welch algorithms, the state prior probabilities, transition probabilities of HMMs, and the weight/mean/covariance parameters of GMMs, can be effectively estimated in an

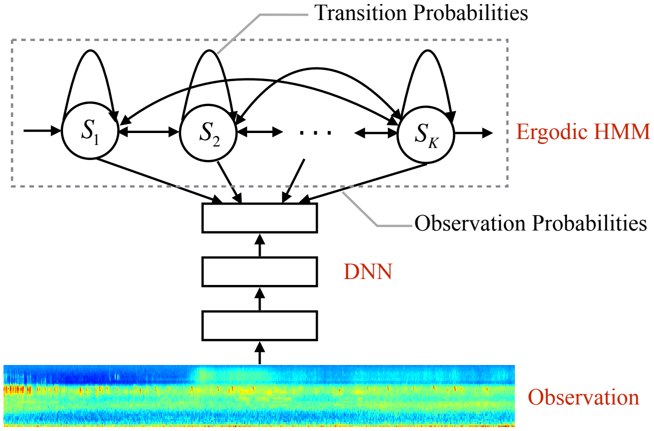


Fig. 4. Architecture of ergodic DNN-HMM hybrid system. The ergodic HMM models the sequential property of the observation, and the DNN models the scaled observation likelihood of all HMM states.

iterative manner and the Gaussian mixtures of all scene classes can be progressively learned and increased from scratch [25]. In the testing stage, Viterbi algorithm is employed for decoding HMM state sequences.

B. Training for DNN-HMM

DNN-HMM hybrid system [20] has been adopted for speech recognition in recent years. The hybrid system takes advantage of DNN's strong representation learning power and HMM's sequential modeling ability, and outperforms conventional GMM-HMM systems significantly. In this work, we train DNN-HMM hybrid system based on ergodic HMMs for scene classification as shown in Fig. 4. In the framework, the dynamics of the audio are modeled with ergodic HMMs, and the observation probabilities are estimated by DNNs. For the GMM-HMM training, the frame-level labels are not necessary as an embedded re-estimation procedure could be applied with the scene labels. However, for the DNN-HMM system, the state labels should be prepared for the subsequent training of DNN model. The procedure for the parameter learning via the Cross-Entropy criterion is as follow,

- **Step1: DNN-HMM Initialization**
  A set of ergodic GMM-HMMs for all scene classes learned using MLE criterion as in Section V-A are prepared. The HMM topology of DNN-HMM system is directly copied from that of GMM-HMM system, including the corresponding state prior probabilities and transition probabilities.
- **Step2: Forced-Alignment**
  The main purpose of forced-alignment here is to generate the frame-level state labels by matching an audio against the corresponding scene label via a general-purpose Viterbi recognizer with GMM-HMMs [14]. After applying to all training samples, the underlying state labels are derived as the learning targets of the DNN output layer.

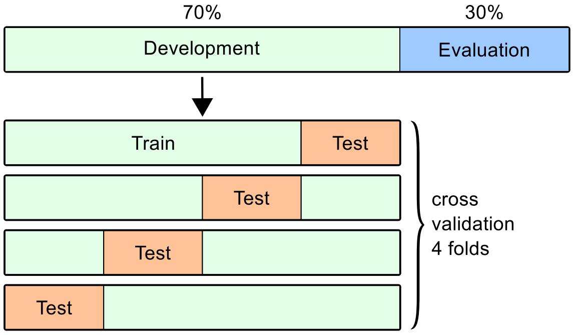


Fig. 5. Dataset partitioning into training and evaluation sets [13].

- **Step3: DNN Cross-Entropy Training**
  As shown in Fig. 4, each output neuron of the DNN is trained to estimate the posterior probability of continuous density HMM's state given the observations. The Cross-Entropy training procedure can be found in Section III-A.

C. Decoding for DNN-HMM

In the testing stage, after the decoding of DNN, posterior probabilities of states given the observations are generated. Since the HMM require the likelihood instead of the posterior probability during the decoding process, we first convert the posterior probability to the likelihood [20]. Viterbi algorithm is then employed for decoding HMM state sequences.

VI. Experiments

A. Experiment Setup

Our experiments are done on the task 1 of DCASE2016 challenge. DCASE2016 is an official IEEE Audio and Acoustic Signal Processing (AASP) challenge for acoustic scene classification and sound event detection within a scene tasks. TUT Acoustic scenes 2016 dataset [13] is used for acoustic scene classification task. The dataset consists of recordings from various acoustic scenes, all having distinct recording locations. There are 15 acoustic scenes for the task:

- Bus - traveling by bus in the city (vehicle)
- Cafe/Restaurant- small cafe/restaurant (indoor)
- Car - driving or traveling as a passenger, in the city (vehicle)
- City center - (outdoor)
- Forest path - (outdoor)
- Grocery store - medium size grocery store (indoor)
- Home - (indoor)
- Lakeside beach - (outdoor)
- Library - (indoor)
- Metro station - (indoor)
- Office - multiple persons, typical work day (indoor)
- Residential area - (outdoor)
- Train - (traveling, vehicle)
- Tram - (traveling, vehicle)
- Urban park - (outdoor)

The dataset was split into development set and evaluation set, such that the evaluation set consists of approximately 30% of the total amount. The development set was further partitioned into 4 folds of training and testing sets to be used for cross-validation during system development. This process is illustrated in Fig. 5. For each acoustic scene, 78 segments were included in the development set and 26 segments were kept for evaluation.

As for front-end, the binaural audio signals from DCASE2016 are converted to a single channel by averaging and then is mixed with the left and right channel audios to form training dataset. In the testing stage, only averaged audios are used. MFCC features were calculated using 40 ms frames with Hamming window and 50% overlap and 40 mel bands. The first 20 coefficients were kept, including the 0th order coefficient. Delta and acceleration coefficients were also calculated using a window length of 9 frames, resulting in a frame-based feature vector of dimension 60. FBANK features were calculated using 40 ms frames with Hamming window and 50% overlap and 40 Mel bands. Delta and acceleration coefficients were also calculated using a window length of 9 frames, resulting in a frame-based feature vector of dimension 120.

For GMM system, a GMM model with 16 components was trained for each scene based on MFCC features using expectation maximization algorithm. For GMM-HMM system, there are 7 emitting states per ergodic HMM to represent each scene class, all GMMs share 3000 Gaussian mixtures. The Baum-Welch iterations are set to a maximum of 40 for all HMMs, yielding good convergence of the likelihoods.

The detailed architecture of DNN in this paper is 1320 (120*11)-512-512-T, where, 1320 (120*11) denotes 120-dimension FBANK feature with 11 frames context information, 512-512 denotes two hidden layers with 512 neurons, and T denotes the number of neurons at output layer. For direct modeling, T is the number of scene classes 15. For DNN-GMM, T is 240 (15*16), and T is 105 (15*7) for DNN-HMM. The Kaldi toolkit was adopted for DNN training. The mini-batch size was 256. The learning rate of Cross-Entropy training was set to 0.001 for the first 10 iterations and then halved after every epoch. Total number of epoch is 20.

For evaluation, the scoring of acoustic scene classification will be based on classification accuracy: the number of correctly classified segments among the total number of segments. Each segment is considered an independent test sample. All the model configuration tunings were done on development dataset.

B. Results and analysis

Table I gives system performances for four folds on development dataset. The systems include GMM baseline provided by DCASE2016 challenge, DNN with direct modeling, DNN-GMM-NP, DNN-GMM, ergodic HMM and DNN-HMM. The performances on four folds were different. For example, the accuracy of GMM ranged from 67.2% to 81.9%. This indicates the dataset is not homogeneous. The change trend of other systems on development dataset were basically consistent except Fold4, where there are no better performance than GMM baseline on this fold. Considering the systems' generalization, we didn't adjust configurations more for Fold4. For DNN-GMM modeling, we can observe that DNN-GMM outperformed DNN-GMM-NP on four folds which indicates that prior scores of subclasses are useful. For ergodic HMM-based modeling, hybrid system DNN-HMM achieved improvement on three folds compared with GMM-HMM.

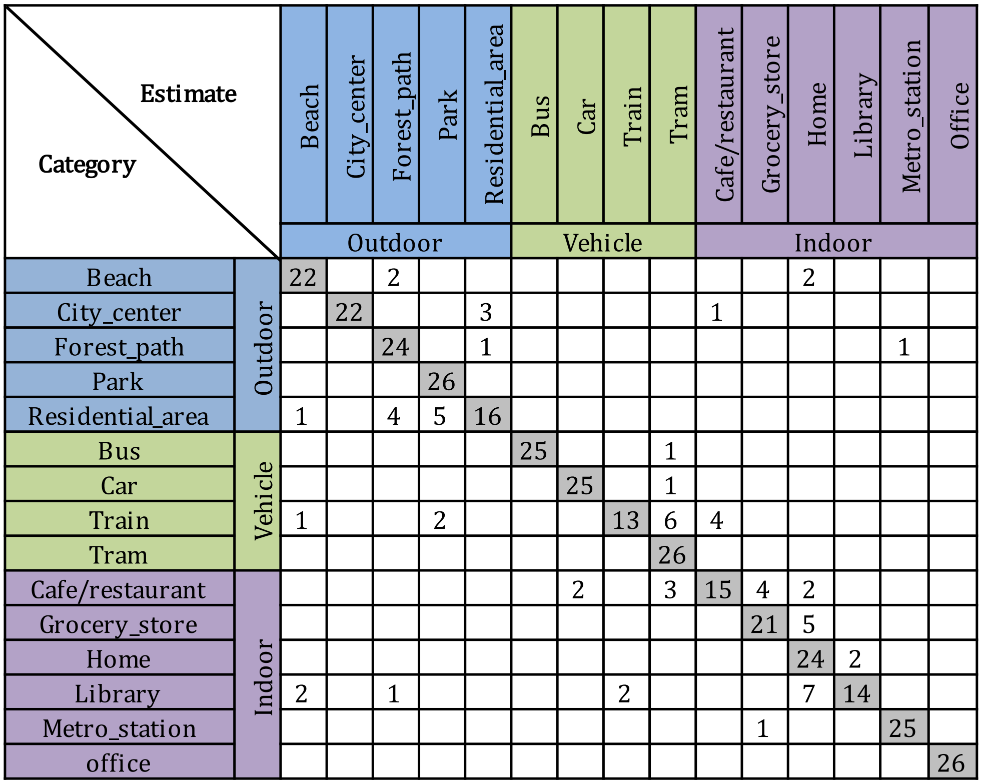
TABLE I
ACOUSTIC SCENE CLASSIFICATION PERFORMANCE FOR FOUR FOLDS ON DEVELOPMENT DATASET.

System	Accuracy			
	Fold1	Fold2	Fold3	Fold4
Direct modeling				
GMM [13]	67.2%	68.9%	72.3%	81.9%
DNN	77.0%	74.6%	71.2%	72.5%
DNN-GMM modeling				
DNN-GMM-NP	78.7%	73.0%	71.9%	72.5%
DNN-GMM	80.7%	74.5%	73.5%	73.4%
Ergodic HMM modeling				
GMM-HMM	75.2%	66.2%	68.3%	79.0%
DNN-HMM	80.2%	73.0%	70.6%	77.3%

TABLE II
ACOUSTIC SCENE CLASSIFICATION RESULTS OF DIFFERENT SYSTEMS ON DEVELOPMENT DATASET AND EVALUATION DATASET (SYSTEMS WITH * SIGN ARE USED TO DO SYSTEM FUSION).

System	Accuracy (Development dataset)	Accuracy (Evaluation dataset)
Direct modeling		
GMM [13]	**72.6%**	**77.2%**
DNN*	73.8%	80.3%
DNN-GMM modeling		
DNN-GMM-NP	74.0%	82.1%
DNN-GMM*	75.5%	**83.3%**
Ergodic HMM modeling		
GMM-HMM	72.2%	79.2%
DNN-HMM*	75.3%	81.8%
System combination		
Fusion	**76.4%**	**83.1%**

Table II lists the accuracies of different systems for acoustic scene classification on development dataset and evaluation dataset. The accuracies on development dataset were averaged over four folds. DNN with direct modeling yielded accuracy improvement on both development and evaluation dataset. When subclasses of GMM were applied as learning targets, DNN could further improve accuracy. The comparison of DNN-GMM-NP and DNN-GMM reveals the power of prior score. DNN-GMM with prior score obtained 83.3% accuracy on evaluation dataset which is the best performance of single system. For ergodic HMM modeling, the performance of GMM-HMM was good than GMM on evaluation dataset. When HMM states were employed as the learning targets of DNN, DNN-HMM could achieve better performance than DNN with direct modeling. DNN-HMM yielded almost the same accuracy with DNN-GMM on development dataset.



Category \ Estimate		Beach	City_center	Forest_path	Park	Residential_area	Bus	Car	Train	Tram	Cafe/restaurant	Grocery_store	Home	Library	Metro_station	Office
		Outdoor					Vehicle				Indoor					
Beach	Outdoor	22		2									2			
City_center			22			3					1					
Forest_path				24		1									1	
Park					26											
Residential_area		1		4	5	16										
Bus	Vehicle						25			1						
Car								25		1						
Train		1			2				13	6	4					
Tram										26						
Cafe/restaurant	Indoor							2		3	15	4	2			
Grocery_store												21	5			
Home													24	2		
Library		2		1					2				7	14		
Metro_station												1			25	
office																26

Fig. 6. Confusion matrix of Fusion system on evaluation dataset in Table II.

However, this difference has expanded rapidly on evaluation dataset from 81.8% to 83.3%. This indicates that subclasses incorporated with prior score have a better generalization ability than HMM states.

DNN, DNN-GMM and DNN-HMM were chosen to do system combination using voting strategy at audio level. The accuracy is seen on development dataset to improve from 72.6% for the GMM baseline system to 76.4% for the final Fusion system. On evaluation dataset, the final Fusion system yielded 25.9% relative error rate reduction compared with GMM, and the accuracy was almost the same with the best single system DNN-GMM. The large difference between DNN and DNN-GMM's performance (80.3% to 83.3%) may be the reason why system combination couldn't bring improvement on evaluation dataset. The corresponding confusion matrix of Fusion system on evaluation dataset is shown in Fig. 6. There are 3 scenes are all correct while scene Train has the lowest accuracy. 23.1% percent test samples of scene Train were assigned to the wrong scene Tram. How to classify the scenes with high similarity is still an open problem.

VII. Conclusion

In this paper, we have investigated high-resolution modeling units of DNN for acoustic scene classification based on GMM and ergodic HMM. Scene category, subclass of GMM and ergodic HMM state are employed from concrete to abstract to train DNN-based systems. Scene label is a straightforward choice for DNN to classify acoustic scenes. However, all frames tagged with the same label is not the best choice because the representative pattern of an audio is sparse. GMM is a generative model for acoustic scene classification. The multiple Gaussians in GMM can be seen as a subclass of the scene. We use the subclass as a bit abstract modeling unit of DNN. Ergodic HMM is more appropriate to model acoustic scenes than GMM. Using HMM states as modeling units, hybrid systems can be built. By comparison, we find high-resolution modeling units are more effective than scene category. Finally, a system combination method is employed to take advantage of the complementarity of different-level modeling units. The final system yields 25.9% relative error rate reduction compared with a GMM baseline on the evaluation dataset of DCASE2016 challenge.

Acknowledgement

This work was supported in part by the National Natural Science Foundation of China under Grant No. 61671422, National Key Research and Development Program under Grant No.2016YFB100130300, the ”Strategic Priority Research Program” of the Chinese Academy of Sciences under Grant No. XDB02070006, and MOE-Microsoft Key Laboratory of USTC.

References

[1] A. Bregman, *Auditory Scene Analysis*. MIT Press, 1990.

[2] D. Wang and G. J. Brown, *Computational auditory scene analysis: Principles, algorithms, and applications*. Wiley-IEEE Press, 2006.

[3] D. Barchiesi, D. Giannoulis, D. Stowell, and M. D. Plumbley, "Acoustic scene classification: Classifying environments from the sounds they produce," *IEEE Signal Processing Magazine*, vol. 32, no. 3, pp. 16–34, 2015.

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

[5] D. Giannoulis, D. Stowell, E. Benetos, M. Rossignol, M. Lagrange, and M. D. Plumbley, "A database and challenge for acoustic scene classification and event detection," in *21st European Signal Processing Conference (EUSIPCO 2013)*. IEEE, 2013, pp. 1–5.

[6] D. Stowell, D. Giannoulis, E. Benetos, M. Lagrange, and M. D. Plumbley, "Detection and classification of acoustic scenes and events," *IEEE Transactions on Multimedia*, vol. 17, no. 10, pp. 1733–1746, 2015.

[7] 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 Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 1, pp. 321–329, 2006.

[8] R. G. Malkin and A. Waibel, "Classifying user environment for mobile applications using linear autoencoding of ambient audio," in *ICASSP*, vol. 5. IEEE, 2005, pp. 509–512.

[9] W. Nogueira, G. Roma, and P. Herrera, "Sound scene identification based on mfcc, binaural features and a support vector machine classifier," *IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Events*, 2013.

[10] J. T. Geiger, B. Schuller, and G. Rigoll, "Recognising acoustic scenes with large-scale audio feature extraction and SVM," *IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Events*, 2013.

[11] B. Elizalde, H. Lei, G. Friedland, and N. Peters, "An i-vector based approach for audio scene detection," *IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Events*, 2013.

[12] K. Patil and M. Elhilali, "Multiresolution auditory representations for scene classification," *cortex*, vol. 87, no. 1, pp. 516–527, 2002.

[13] M. Annamaria, H. Toni, and V. Tuomas, "TUT database for acoustic scene classification and sound event detection," in *24rd European Signal Processing Conference 2016 (EUSIPCO 2016)*, 2016.

[14] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey *et al.*, "The HTK book (for HTK version 3.4)," *Cambridge university engineering department*, vol. 2, no. 2, pp. 2–3, 2006.

[15] V. Ramasubramanian, R. Karthik, S. Thiyagarajan, and S. Cherla, "Continuous audio analytics by HMM and viterbi decoding," in *ICASSP*. IEEE, 2011, pp. 2396–2399.

[16] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, 2006.

[17] G. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural computation*, vol. 18, no. 7, pp. 1527–1554, 2006.

[18] L. Deng and D. Yu, "Deep learning: methods and applications," *Foundations and Trends in Signal Processing*, vol. 7, no. 3–4, pp. 197–387, 2014.

[19] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath *et al.*, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012.

[20] D. Yu and L. Deng, "Automatic speech recognition - a deep learning approach," 2015.

[21] Q. Wang, J. Du, X. Bao, Z.-R. Wang, L.-R. Dai, and C.-H. Lee, "A universal VAD based on jointly trained deep neural networks," in *INTERSPEECH*, 2015, pp. 2282–2286.

[22] Y. Petetin, C. Laroche, and A. Mayoue, "Deep neural networks for audio scene recognition," in *Signal Processing Conference (EUSIPCO), 2015 23rd European*. IEEE, 2015, pp. 125–129.

[23] Q. Kong, I. Sobieraj, M. Plumbley, and W. Wang, "Deep neural network baseline for DCASE challenge 2016," DCASE2016 Challenge, Tech. Rep., September 2016.

[24] D. Battaglino, L. Lepauloux, and N. Evans, "Acoustic scene classification using convolutional neural networks," DCASE2016 Challenge, Tech. Rep., 2016.

[25] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The Kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*, no. EPFL-CONF-192584. IEEE Signal Processing Society, 2011.