TECHNICAL REPORT OF USTC SYSTEM FOR ACOUSTIC SCENE CLASSIFICATION

Xiao Bao, Tian Gao, Jun Du

National Engineering Laboratory for Speech and Language Information Processing,
University of Science and Technology of China, Hefei, Anhui, China
{baox, gtian09}@mail.ustc.edu.cn, jundu@ustc.edu.cn

ABSTRACT

This technical report describes our submission for acoustic scene classification task of DCASE 2016. We first explore the use of and Gaussian mixture models (GMM) and ergodic hidden Markov models (HMM). Next, we combine neural network based discriminative models (DNN, CNN) with generative models to build hybrid systems, including DNN-GMM, CNN-GMM, DNN-HMM and CNN-HMM. Finally, a system combination method is used to obtain the best overall performance from the multiple systems.

Index Terms— DCASE 2016, scene classification, GMM, ergodic HMM, DNN, CNN

1. 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.). DCASE 2016 [1] is an official IEEE Audio and Acoustic Signal Processing (AASP) challenge for acoustic scene classification and sound event detection within a scene tasks. The goal of acoustic scene classification is to classify a test recording into one of predefined classes that characterizes the environment in which it was recorded – for example "park", "street", "office". TUT Acoustic scenes 2016 dataset is used for the task. The dataset consists of recordings from various acoustic scenes, all having distinct recording locations. For each recording location, 3-5 minute long audio recording was captured. The original recordings were then split into 30-second segments for the challenge.

Over the last few years, scene classification has been gradually receiving attention in the field of audio signal processing, including the aspects of features, statistical models, decision criteria and meta-algorithms [2]. In this work, we focus on the exploration of statistical models.

2. SYSTEM OVERVIEW

The overall flowchart of our proposed system is illustrated in Fig. 1. Our system has two parts, namely GMM-based systems and ergodic HMM-based systems. In the training stage, first the input binaural audio signal is converted to a single channel by averaging and then is mixed with the left and right channel audio to form training dataset. Then, the training samples are further processed to extract MFCC and log Mel-filterbank (FBANK) features with 40 ms frame length and 20 ms frame shift. The MFCC features include static coefficients (0th coefficient included), delta coefficients and acceleration coefficients. And FBANK features also include delta coefficients and acceleration coefficients.

Next, MFCC features are used for both GMM and ergodic GMM-HMM training. These systems learn one acoustic model per acoustic scene class, and do the classification with maximum likelihood classification scheme. Then, subclass labels and state labels are generated from GMM and ergodic HMM models, respectively. Based on FBANK features as input and the above labels as learning targets, DNN (deep neural network) and CNN (convolutional neural network) models are trained to build hybrid systems, including DNN-GMM, CNN-GMM, DNN-HMM and CNN-HMM. Finally, system combination is implemented to utilize the complementarity of multiple systems.

In the testing stage, only averaged audios are used. Other processing configurations are the same with training samples. All HMM-based and neural network-based experiments are implemented by using Kaldi toolkit [3].

3. GMM-BASED SYSTEMS

3.1. Official Benchmark: GMM

A GMM-based baseline system is provided as the official benchmark which consists of MFCC features and GMM based classifier [4]. For each acoustic scene, a GMM class model with 16 Gaussians is trained based on MFCC features using expectation maximization algorithm. The testing stage uses maximum likelihood decision among all acoustic scene class models. Classification performance is measured using accuracy: the number of correctly classified segments among the total number of test segments.

3.2. DNN-GMM

A GMM class model with 16 Gaussians in above is trained for each acoustic scene. We assume these 16 Gaussians have different level of contribution and each Gaussian can be seen as a subclass of the scene. In this work, we use DNN to predict GMM subclasses and use a post-processing method to classify acoustic scenes. DNN has strong representation learning power to discriminate the total 240 (240=15*16) subclasses with FBANK features as input. The subclass for each frame feature is the Gaussian which has the maximum log-likelihood in the scene GMM model. Two hidden layers with 512 neurons and sigmoid activation function are used.

In the testing stage, we first calculate a prior score for each subclass from training set. We assume the subclass $S_{i,j}$ with largest number of frames is the dominant subclass of a training sample x_h, where $S_{i,j}$ is the jth subclass of the ith scene. The frequency of

Thanks to the National Natural Science Foundation of China under Grants No. 61305002 for funding.

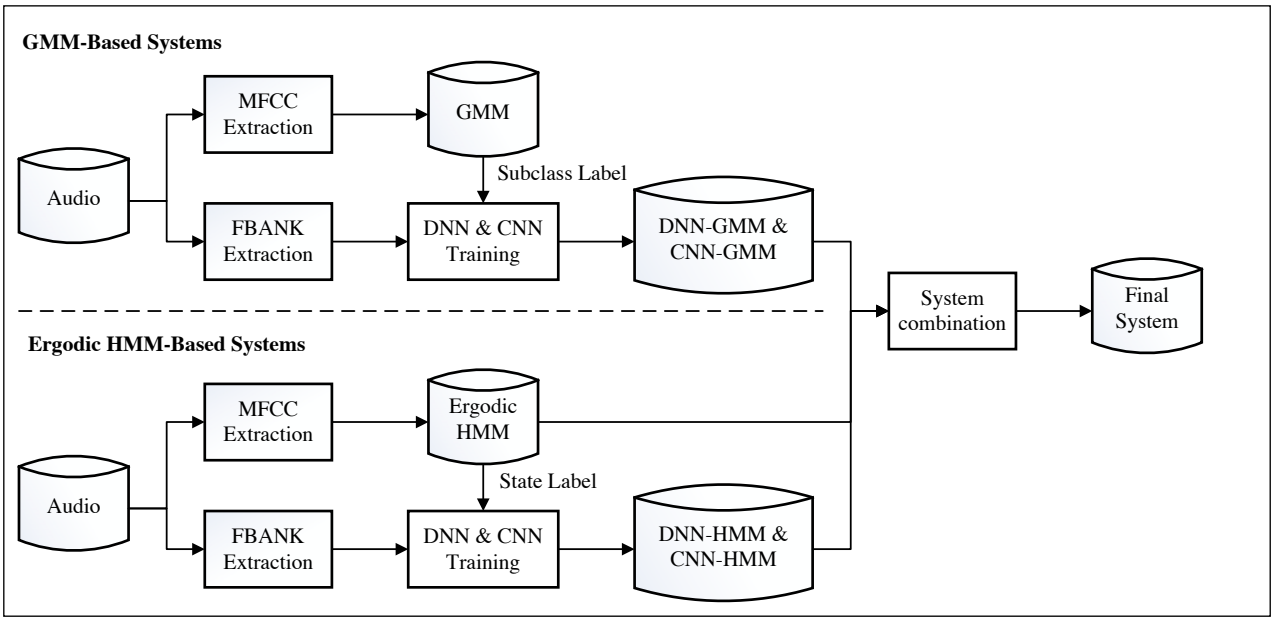

Figure 1: System overview.

$S_{i,j}$ in x_h is defined as follow,

$$F_{i,j}(\mathrm{x}_h) = \begin{cases} n_{i,j}(\mathrm{x}_h)/N_{frame}, & S_{i,j} \text{ is dominant subclass} \\ 0, & \text{else} \end{cases} \quad (1)$$

where, $F_{i,j}(\mathrm{x}_h)$ is the frequency of dominant subclass $S_{i,j}$ in training sample x_h, $n_{i,j}(\mathrm{x}_h)$ and N_{frame} representing the number of frames of $S_{i,j}$ and x_h, respectively. Then we give each subclass a prior score based on its frequency in training samples as follow,

$$P_{i,j} = \frac{1}{N_{sample}} \sum_{h=h_1 \ldots H} F_{i,j}(\mathrm{x}_h) \quad (2)$$

where, $P_{i,j}$ is a prior score of subclass $S_{i,j}$, $\mathrm{x}_{h_1} \ldots \mathrm{x}_H$ are training samples whose dominant subclass is $S_{i,j}$ and N_{sample} is the number of these training samples. So, we can use a matrix $\boldsymbol{Prior}$ to represent the prior scores of subclasses as follow,

$$\boldsymbol{Prior} = \begin{bmatrix} P_{1,1} \\ P_{1,2} \\ \vdots \\ P_{15,16} \end{bmatrix} \quad (3)$$

When a testing frame feature x is fed to the well trained DNN, posterior probability of 240 subclasses will be generated. We use a threshold to reset each value to 1 or zero. A matrix $\boldsymbol{Post}(\mathrm{x})$ is used to represent the scaled posterior probability. Next, our decision score $\boldsymbol{D}(\mathrm{x})$ is defined as follow,

$$\boldsymbol{D}(\mathrm{x}) = \boldsymbol{Prior} \odot \boldsymbol{Post}(\mathrm{x}) = \begin{bmatrix} D_{1,1}(\mathrm{x}) \\ D_{1,2}(\mathrm{x}) \\ \vdots \\ D_{i,j}(\mathrm{x}) \\ \vdots \\ D_{15,16}(\mathrm{x}) \end{bmatrix} \quad (4)$$

where $D_{i,j}(x)$ is the decision score of jth subclass in ith scene. And the scene which the input frame x belonged to is determined as follow,

$$C(x) = \underset{i}{\operatorname{argmax}} \sum_{j=1}^{16} D_{i,j}(x) \quad (5)$$

where $C(x)$ is the frame-level scene decision. Finally, the scene category of a whole testing sample is determined by using majority decision method based on frame-level results.

3.3. CNN-GMM

A CNN provides shift invariance over time and space and is critical to achieve state-of-art performance of image recognition. It has also been shown to improve speech recognition accuracy over pure DNN on some tasks. Similar with the above DNN-GMM system, we use one dimensional CNN to build CNN-HMM system. Two convolution layers and two fully connect layers are used.

4. ERGODIC HMM-BASED SYSTEMS

4.1. Ergodic GMM-HMM

HMM [5] 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. And ergodic HMM is more suitable for scene classification due to the uncertain structure of scene audio [6, 7]. In this work, ergodic GMM-HMMs are used for classification by training an ergodic HMM for each class with MFCC features, and by selecting the class with the largest a posteriori probability. We use the maximum-likelihood based BaumWelch algorithm to train the HMMs for each class separately. The number of states is 6 per

HMM and all HMMs share 3000 Gaussians. The Baum-Welch iterations are set to a maximum of 40 for all HMMs, yielding good convergence of the likelihoods. Viterbi algorithm is used for decoding HMM state sequences.

4.2. DNN-HMM

DNN-HMM hybrid system [8] has been used 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 on many speech recognition tasks. In this work, we train a DNN-HMM hybrid system based on above ergodic GMM-HMM system for scene classification. In this framework, the dynamics of the audio is modeled with ergodic HMMs and the observation probabilities are estimated through DNNs. Each output neuron of the DNN is trained to estimate the posterior probability of continuous density HMM's state given the observations. We use FBANK features with 11 frame context information as the input and two hidden layers with 512 neurons.

4.3. CNN-HMM

CNN can provide shift invariance over time and space and is critical to achieve state-of-art performance of image recognition. Similar with DNN-HMM system, we use one dimensional CNN to build CNN-HMM system. Two convolution layers and two fully connect layers are used.

5. SYSTEM COMBINATION

In this work, we build multiple systems from different aspects. GMM, subclass of GMM and HMM are different choices of modeling unit. DNN and CNN are different feature representation methods. We use a voting strategy to combine these systems.

6. REFERENCES

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

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] V. Ramasubramanian, R. Karthik, S. Thiyagarajan, and S. Cherla, "Continuous audio analytics by HMM and viterbi decoding," in *IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2011, pp. 2396–2399.

[8] 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.