# GMM-AA SYSTEM FOR ACOUSTIC SCENE CLASSIFICATION

*Vinayak Abrol, Pulkit Sharma, Anshul Thakur*

IIT Mandi, India

{vinayak_abrol, pulkit_s, anshul_thakur}@students.iitmandi.ac.in

## ABSTRACT

In this submission we propose to use Gaussian mixture modelling and Archetypal Analysis based system for DCASE17 acoustic scene classification task. We propose a feature learning approach via decomposing time-frequency (TF) representations with Archetypal Analysis (AA). In order to process large number of TF frames and capture the variations efficiently, firstly a class-specific GMM is build on frames of TF representations, followed by AA on GMM means to build class specific local dictionaries. Next, the TF representations are projected on the concatenated AA dictionary to get the non-negative sparse activations. Finally, the TF frames are reconstructed back using the computed activation vectors, and are then used to train a SVM classifier. The proposed method significantly outperforms the baseline system.

***Index Terms***— Archetypal analysis, dictionary learning, acoustic scene classification.

## 1. INTRODUCTION

The goal of Acoustic Scene Classification (ASC) task is to classify unstructured data from audio scenes, which may have variable-quality recordings, background/environmental noise, and variation of the acoustic content itself. In search of effective discriminative features representing a scene, existing works in the literature used the features inspired by speech e.g., Mel Frequency Cepstral Coefficients, zero-crossing rate etc. [1]. In addition, various other works have used image processing techniques to derive features e.g., histograms of oriented gradients from the time frequency representations [2, 3]. Most of the existing features focus on describing a specific aspect of the signal, and hence lack generalizability and flexibility. However, a few existing works for ASC have successfully employed feature learning techniques such as nonnegative matrix factorization (NMF) to derive adaptive representations of the data [4]. Similarly, the method proposed in this work combines the advantage of Gaussian mixture modeling and archetypal analysis (AA) to derive features for ASC.

This report is organized as follows: Section 2 provides a detailed description of the method proposed for the ASC task. Our experimental setup is provided in section 4 and results are reported in Section 4. Finally, the report is summarized in section 5.

## 2. OUR SUBMISSION

Our challenge contribution employs Gaussian mixture modeling and archetypal analysis based approach for learning discriminative features for ASC task. Final classification using the derived features is performed using SVM classifier. Our system outperformed the DCASE17 baseline system in terms of average classification accuracy at 88% on development data using 4fold evaluation strategy as provided in challenge.

### 2.1. Proposed GMM-AA based method for ASC

Designing a good ASC system requires a suitable choice of features which can efficiently describe the acoustic environments. In this submission we have considered supervised feature learning using time-frequency (TF) representations namely the constant Q-transform (CQT), which has been shown to perform well for ASC task. Let $\mathbf{X}^p \in \mathbb{R}^{m^p \times n}$ denote the CQT transform of a given acoustic scene recording, where $m^p$ and $n$ represent the number of time frames and the number of frequency bands, respectively. Next, a pooling step is applied on $\mathbf{X}$ to obtain a suitable representation for feature learning step. In this work, we applied a max-pooling operator on non-overlapping frames from 1sec of each recording. Thus, after the pooling step the data matrix $\mathbf{X}^p$ is transformed into $\mathbf{X} \in \mathbb{R}^{m \times n}$, where $m = 10$ for 10sec recording. This way one can, not only perform temporal integration i.e., capturing the long term dependencies present in audio signal, but can also reduce the amount of data frames to be processed. In order to efficiently model the variations present in the data, all the training examples of each class (after log compression) are next modeled using a GMM. Consider $m_i$ frames (obtained after pooling step) of an acoustic scene class (obtained from all the training examples) arranged in a matrix $\mathbf{X}_i \in \mathbb{R}^{m_i \times n}$ as rows, ($i = 1, 2, \ldots C$, $C$ being the total number of classes). The training data in $\mathbf{X}_i$ is modeled using a GMM $\sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{X}_{ik} | \boldsymbol{\mu}_{ik}, \Sigma_{ik})$ with $k = 1, 2, ..., K$ Gaussian mixtures, where $\boldsymbol{\mu}_{ik}$ denotes the mean corresponding to $k^{th}$ mixture of $i^{th}$ class.

The next step of the proposed approach perform feature learning from the learned GMM based generative model by decomposing the means of each class using archetypal analysis. AA performs convex-hull approximation via sparse-convex non-negative decomposition of the data, and this approach has been shown to extract underlying relevant basis/archetypes representing data [5]. Using AA, $t_i$ archetypes for each class are stored together to form a large overcomplete dictionary $\mathbf{D} \in \mathbb{R}^{t \times n}$ (a mixture of local dictionaries), such that $\sum_{i=1}^{C} t_i = t$. This learned dictionary $\mathbf{D}$ is used to compute a sparse-convex representation corresponding to the pooled representation $\mathbf{X}_i$ (both train and test) by solving the following optimization function:

$$\underset{\mathbf{A}, \mathbf{a}_j \in \Delta_t}{\operatorname{argmin}} \|\mathbf{X}^T - \mathbf{D}^T \mathbf{A}\|_F^2, \Delta_t \triangleq [\mathbf{a} \geq 0, \|\mathbf{a}\|_1 = 1], \tag{1}$$

where $\| \, . \, \|_F$ is the Frobenius norm, $\Delta$ denotes the simplex, $\geq$ denotes element wise operation, and $\mathbf{a}_j$ is a column of $\mathbf{A}$. The estimated activation matrix $\mathbf{A}_i$ is used to reconstruct the initial representation as $\hat{\mathbf{X}}_i^T = \mathbf{D}^T \mathbf{A}_i$. Finally, the reconstructed $\hat{\mathbf{X}}_i$ (for training data) is used to train a SVM classifier, to perform predictions on the test data. The idea of using reconstructed features for classification is motivated by recent works of Sainath et. al [6] in the field of speech recognition.

## 3. EXPERIMENTAL SETUP

Only first channel of each audio file is processed, and its CQT is extracted with the MATLAB CQT toolbox[1] (hopsize-882, 12-bands/octave), resulting in 500 frames and 146 frequency bands. After pooling each audio file is represented with 10 TF frames. Each class-specific GMM is build with 200 means using Microsoft identity toolkit[2]. AA is performed using the fast AA algorithm (based on SPAMS toolbox[3]) proposed in [5]. Here we have considered robust AA to deal with any outlier frames which may affect the performance of the system. For each class 40 archetypes are learned, resulting in a final concatenated dictionary of size $600 \times 146$. The SVM classifier is trained with one-vs-one strategy (box-constraint=3), and majority voting is employed for 10 frame representations of each audio recording on the obtained scores to decide the class label. All experiments are performed in MATLAB R2016b under Win10 OS on a desktop with 4th Gen i7-processor, 12Gb RAM and no GPU.

## 4. RESULTS

Our final system achieved an accuracy of 88% ± 0.5 (95% C.I.) averaged across four different folds using the provided development dataset. Thus we achieved an improvement of 13.2% over the DCASE17 baseline system. The confusion matrix for our GMM-AA system is shown in Table. 1.

On evaluation dataset, our system ranked $17^{th}$ (among teams) and achieved an accuracy of 65.7% with range 63.4 – 68.0% (95% C.I.). While our system performed well for most classes, it performed poorly on some confusing classes like forest_path, park, residential area etc., which we wish to investigate in the future.

## 5. DISCUSSION

The proposed work show the feasibility of using a GMM-AA based system using TF representations for ACS task. While GMM helps in efficiently modeling the data distribution, AA learns representative basis/archetypes which can be used to emphasis relevant information to discriminate between different audio scenes. The obtained atom activations are probabilistic in nature and they depict the amount of contribution coming from atoms of a particular class. This gives easy interpretation of data, i.e., larger support for correct class. Hence, the reconstructed feature will emphasis relevant class information. Further, note that data modeling for each class is done in an independent supervised way but not in a joint manner, and we hope doing so will further increase the accuracy of our system.

Similar to ours, the system proposed by Bisot et. al[4] (ranked $10^{th}$ among teams with accuracy of 69.8%) based on non-negative matrix factorization approach, shows that the conventional approaches can still potentially be applied to achieve better result. The improvement achieved by their system (compared to ours) was result of 1) joint supervised factorization 2) features learned on averaged signal from both channels and 3) fusion with DNN system trained on NMF features, where we feel the third approach is mainly making the difference in results. Further, many of the submissions in the challenge were based on deep learning approaches such as a convolutional neural network (CNN) and recurrent neural network (RNN). A careful analysis of results shows that simply applying CNNs with input features (mono/binaural) like CQT, mel-spectrogram etc., performed poorly as compared to the proposed approach. Although DCASE-17 provided more amount of audio data from last year, it still may not be sufficient for applying deep learning approaches efficiently. To address this some of the top performing systems employ data augmentation and/or score fusion with deep architectures learned on various different features (with/without background subtraction). In future, parallel to deep architectures (CNNs/RNNs), we wish to explore deep matrix factorization approaches for learning discriminative features for ASC.

## 6. REFERENCES

[1] J. T. Geiger, B. Schuller, and G. Rigoll, "Large-scale audio feature extraction and SVM for acoustic scene classification," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2013, pp. 1–4.

[2] V. Bisot, S. Essid, and G. Richard, "HOG and subband power distribution image features for acoustic scene classification," in *European Signal Processing Conference (EUSIPCO)*, 2015, pp. 719–723.

[3] A. Rakotomamonjy and G. Gasso, "Histogram of gradients of time-frequency representations for audio scene classification," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 1, pp. 142–153, 2015.

[4] V. Bisot, R. Serizel, S. Essid, and G. Richard, "Acoustic scene classification with matrix factorization for unsupervised feature learning," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 6445–6449.

[5] V. Abrol, P. Sharma, and A. K. Sao, "Finding archetypes by exploiting sparsity of convex representations," in *Signal Processing with Adaptive Sparse Structured Representations (SPARS) workshop*, June 2017.

[6] T. N. Sainath, B. Ramabhadran, M. Picheny, D. Nahamoo, and D. Kanevsky, "Exemplar-based sparse representation features: From timit to lvcsr," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 8, pp. 2598–2613, Nov 2011.

[1] https://www.cs.tut.fi/sgn/arg/CQT/

[2] https://www.microsoft.com/en-us/download/details.aspx?id=52279

[3] spams-devel.gforge.inria.fr

[4] http://www.cs.tut.fi/sgn/arg/dcase2017/documents/challenge_technical_reports/DCASE2017_Bisot_193.pdf

Table 1: Confusion matrix for our GMM-AA system. (TPR) True Positive Rate, (FAR) False Alarm Rate. Classes: 1. Beach 2. Bus 3. Cafe/Restaurant 4. Car 5. City center 6. Forest path 7. Grocery store 8. Home 9. Library 10. Metro station 11. Office 12. Park 13. Residential area 14. Train 15. Tram

| | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **True Class** | ***1*** | 84% | 1% | | 1% | <1% | <1% | <1% | 5% | | | | 1% | 3% | <1% | 4% | 84% | 16% |
| | ***2*** | 1% | 93% | | 1% | <1% | | | <1% | 1% | | | <1% | | 1% | 3% | 93% | 7% |
| | ***3*** | <1% | | 93% | | 1% | <1% | <1% | 4% | | 1% | | <1% | 1% | <1% | | 93% | 7% |
| | ***4*** | 1% | 1% | | 96% | | <1% | | <1% | <1% | | | | | <1% | 2% | 96% | 4% |
| | ***5*** | 1% | | 1% | | 83% | 1% | <1% | | | 1% | <1% | 1% | 12% | | <1% | 83% | 17% |
| | ***6*** | <1% | | 1% | | 1% | 81% | <1% | 14% | <1% | <1% | <1% | <1% | 2% | | | 81% | 19% |
| | ***7*** | <1% | <1% | 2% | | <1% | <1% | 93% | 2% | <1% | 2% | | | <1% | <1% | <1% | 93% | 7% |
| | ***8*** | 1% | <1% | <1% | | <1% | 2% | <1% | 94% | 1% | | <1% | <1% | 1% | | <1% | 94% | 6% |
| | ***9*** | 1% | <1% | <1% | <1% | 1% | <1% | <1% | 2% | 93% | | <1% | 1% | <1% | 1% | | 93% | 7% |
| | ***10*** | <1% | | 1% | | <1% | 3% | <1% | 4% | <1% | 88% | <1% | 1% | 1% | | <1% | 88% | 12% |
| | ***11*** | <1% | | <1% | | | 1% | <1% | 4% | <1% | <1% | 93% | <1% | <1% | | | 93% | 7% |
| | ***12*** | 5% | | | 1% | 1% | <1% | 3% | <1% | <1% | <1% | <1% | 81% | 7% | 1% | 1% | 81% | 19% |
| | ***13*** | 6% | | 1% | | 10% | 3% | 1% | 2% | 1% | 1% | | 8% | 68% | <1% | <1% | 68% | 32% |
| | ***14*** | 3% | 2% | <1% | <1% | | <1% | <1% | 2% | <1% | <1% | | | 1% | 90% | 2% | 90% | 10% |
| | ***15*** | 2% | 1% | 1% | 2% | | <1% | <1% | 1% | <1% | <1% | | <1% | 1% | 1% | 91% | 91% | 9% |
| | | ***1*** | ***2*** | ***3*** | ***4*** | ***5*** | ***6*** | ***7*** | ***8*** | ***9*** | ***10*** | ***11*** | ***12*** | ***13*** | ***14*** | ***15*** | ***TPR*** | ***FAR*** |
| | **Predicted Class** | | | | | | | | | | | | | | | | | |