AUDIO FEATURES IN A FUSION-BASED FRAMEWORK FOR ACOUSTIC SCENE CLASSIFICATION

Shefali Waldekar, Goutam Saha

Electronics and Electrical Communication Engineering Dept.,
Indian Institute of Technology Kharagpur, India,
{shefaliw, gsaha}@ece.iitkgp.ernet.in

ABSTRACT

This report describes two submissions for Acoustic Scene Classification (ASC) task of the IEEE AASP challenge on Detection and Classification of Acoustic Scenes and Events (DCASE) 2017. The first system follows an approach based on a score-level fusion of some well-known spectral features of audio processing. The second system uses the first proposed system in a two-stage hierarchical classification framework. The two systems respectively show 18% and 21% better performance on the development dataset, and 10% and 6% better performance on the evaluation dataset, relative to that of the MLP-based baseline system of DCASE 2017.

Index Terms— Fusion, hierarchical classification, spectral features, SVM

1. INTRODUCTION

Acoustic scene classification (ASC) [1] is a closed-set classification task, where semantic labels are assigned to audio streams according to the environments they represent. These environments could be indoor (home, office, library etc.), outdoor (busy-street, forest, beach etc.), or a moving vehicle (car, bus, train etc.). Applications of ASC can be in context-aware and intelligent wearable devices, hearing-aids, robotic navigation systems, and audio archive management systems.

With application point of view, it is required that the machine listening algorithms be such that they are able to work with different types of audio, that is, speech, music, as well as environmental sounds. In the two systems presented in this report, we use some spectral and temporal features from audio processing fields. The motivation behind using the spectral features, namely, *non-overlap block transform coefficients* (NOBTC) [2] and *subband centroid frequency coefficients* (SCFC) [3], was to exploit the specific spectral characteristics of the audio events in a scene. We also use *constant-Q cepstral coefficients* (CQCC) features [4] in an attempt to mimic the human hearing system better than mel-scaled features. Our first proposed system employs a fusion-based framework. The classification results from the use of aforementioned features from binaural audio streams are *score-fused* to get the final classification. In the second proposed system, a two-stage hierarchical framework is put into place by employing the first proposed system in both stages. However, stage-one does coarse three-class classification into indoor, outdoor and vehicle classes, while stage-two gives the final classification output.

The rest of this report is organized as follows: In Section 2, we give the description of the elements that are core to both the proposed systems. Next, in Section 3 and Section 4 we elaborate on the formation of the two systems respectively. In Section 5, we present the experimental configuration and show the results. It is followed by the conclusion of the work in Section 6.

2. BASIC SYSTEM CONFIGURATION

2.1. Features

The proposed systems use the following as features.

- *Non-overlapped block transform coefficients (NOBTC)* [2]: In all fields of speech processing, mel-frequency cepstral coefficients (MFCC) are the most exploited features. Discrete cosine transform (DCT) is an important step in MFCC feature extraction. The feature extraction scheme in [2], for speaker recognition, captures speech information in a more efficient manner than the standard MFCC, because it applies DCT in blocks based on dominant formant frequency zones. The spectrum of audio scene signals too can be divided into frequency zones due to the presence of various acoustic events. Non-overlapped block transform coefficients (NOBTC) are a type of block-based MFCC, in which the DCT blocks do not overlap. NOBTC extracted with 60 filters, distributed in three blocks, gave best results in [5]. This choice of the number of filters takes into account the fact that we are dealing with general audio signals whose frequency components can span the whole audible frequency range of 20Hz to 20 kHz.
- *Subband centroid frequency coefficients (SCFC)* [3]: Spectral centroids are the centre of masses for the frequency bands under study and are perceptually connected to 'brightness' of a sound. These are also found to be quite robust to noise. We used subband centroid frequency (SCF), which is the weighted average frequency for a given subband, as a feature. The weights are the normalized energy of each frequency component in that subband. We have divided the frequency band uniformly on the Hz-scale with 60 overlapping (50% overlap) rectangular filters, thus resulting in 60 subbands per frame.
- *Constant-Q cepstral coefficients (CQCC)* [4]: While audio perception in humans asks for higher frequency resolution at lower frequencies, it also exhibits a higher temporal resolution at higher frequencies. This is equivalent to having a set of filters with constant Q-factor across the entire spectrum, and that can be achieved by geometrically spaced frequency bins. Constant-Q transform (CQT) implements the same and is commonly used in music signal processing. The coupling of CQT with traditional cepstral analysis resulted in constant-Q cepstral coefficients (CQCC) [4].

- *Short-term (ST) time and frequency features* [6]: Short-term features, such as zero crossing rate (ZCR), energy, entropy of energy, spectral centroid, spectral spread, spectral flux, spectral roll-off point, spectral entropy, harmonic ratio, and fundamental frequency, are found to possess the ability to discriminate between various sounds [6]. Since acoustic scenes are a collection of multiple environmental sounds, these features are expected to add to the information captured by cepstral features.

In our experiments, it was found that the inclusion of delta features improved the performance for NOBT coefficients, but not so with CQCC and SCFC. The addition of double-delta features did not benefit in any case.

2.2. Classifier

In our system, we have used SVM with RBF (radial basis function) kernel. Since SVM is a binary classifier, in order to determine a decision criterion for multi-class ASC, we have combined multiple SVMs following one-versus-one approach. Thus, for N classes, $N(N-1)/2$ classifiers are made, where each one trains on data from two classes. The decision criterion estimates the class of an unknown sample by evaluating the distance between the feature-point and the separating hyperplanes learned by the SVMs. Each binary classification is deemed to be a voting where votes are cast for all data points. The class with the maximum votes acquires a data point in the end.

2.3. Fusion strategy

SVM requires that each data sample is represented as a vector. For this purpose mean and standard deviation are considered as a good representation of the whole data [7]. The audio of DCASE challenge was recorded in binaural format, i.e., the two channels carried different values. One possible way of working with such data is to process the channels separately and then combine their results in a way such as score-level fusion. In score-level fusion, the classifier output is combined such that appropriate weights are given to the decisions of different participating systems. In this case, the system performing better should be given more weight in the decision making. Weights can be fixed empirically, but the process is cumbersome and also not robust. We have used the weight optimization algorithm followed by FoCal Multi-class toolkit [8], which uses the classification performance of each classifier and applies logistic regression to derive appropriate weights for score fusion.

3. PROPOSED SYSTEM 1

The block diagram for the first proposed system is shown in Fig. 1. In this system, the data of both the channels are individually processed. The required features are extracted from windowed frames of pre-emphasized audio and then across-frames mean and standard deviation are calculated. These vectors are used to train the SVM corresponding to each feature. The scores from the feature-wise classifiers are fused to generate channel-wise scores, which in turn are fused to generate the final scores. During development, a k-fold cross-validation setup was used. The weights for fusion were obtained from test portion of the folds and were saved for later use in system testing. The data for testing comes from the evaluation dataset and follows a path similar to that of development. However, in this case, whole development data is used for training the SVMs.

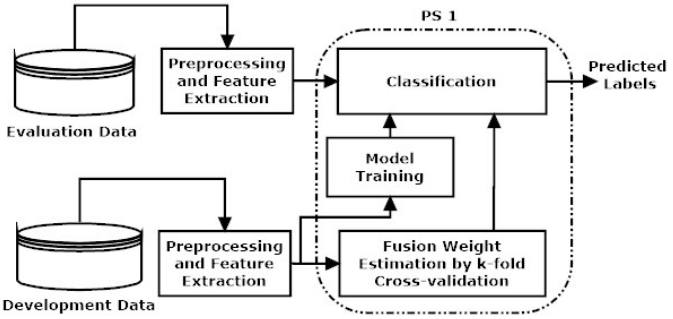

Figure 1: Block diagram of fusion-based Proposed System 1

The means of the fold-wise weights obtained during development are now used as weights for fusion of scores.

4. PROPOSED SYSTEM 2

The second proposed system is based on the concept of hierarchy. The data is divided into three classes as given in Table 1. The block diagram of this system is shown in Fig 2. For this system, we have used the architecture of the fusion-based Proposed System 1 (PS 1) in both stages. In the first stage, there is a three-class classifier which classifies the incoming audio streams for testing as belonging to either indoor, outdoor or vehicle class. The 'Coarse Labelling' block does the job of grouping the audio streams from training in three classes and changing labels accordingly. In the second stage, the results of the first stage are divided by class labels and supplied to three separate classifiers according to the class. Note that at this stage there is a six-class, a five-class and a four-class classifier for indoor, outdoor and vehicle class, respectively. The labels predicted by these systems are then put together by 'Label Combiner' to give the final predicted labels according to the test streams.

Table 1: Class-division for two-stage classification

Stage	**Class**		
One	Indoor	Outdoor	Vehicle
Two	cafe/restaurant, grocery store, home, library, metro station, office	lakeside beach, city center, forest path, urban park, residential area	bus, car, train, tram

5. RESULTS

5.1. Experimental Framework

We have used the development dataset of TUT Acoustic Scenes 2016 [9] and TUT Acoustic Scenes 2017 [10] in our experiments. The two datasets differ from each other only in the length of the audio streams (30sec for 2016 and 10 sec for 2017). From all the data samples, NOBTC, SCFC and ST features were extracted by applying Hamming window on 20 ms frame having 50% overlap. Pre-emphasis to the audio signals was done by a factor of 0.97. Filter-bank of 60 filters was used for NOBTC (triangular filters) and SCFC

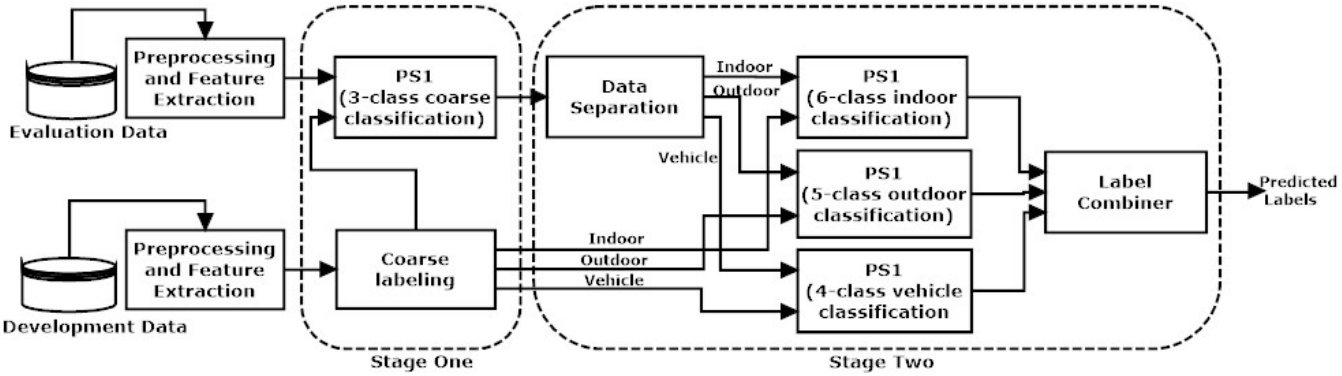

Figure 2: Block diagram of hierarchy-based Proposed System 2

(rectangular filters). The parameters for CQCC features given in [4] were used here. Delta (Δ) features, evaluated with a 3-frame window, were appended only for NOBTC. Frame-wise mean and standard deviation of the features were given as input to SVM classifier with RBF kernel. According to the DCASE challenges' ASC task setup, development data is partitioned into k folds, where k=4 for both 2016 and 2017. The fold-wise mean of classification accuracy was used as the performance metric.

5.2. Performance on Development Data

The results of Proposed System 1 on both the datasets is shown in Table 2. It can be seen here that different features perform differently on both the channels. When the data length is longer, i.e. 2016 challenge data, Channel 1 performed better classification than channel 2, while, with reduced data length the two channels' performance was equivalent. Nevertheless, both channels carried complementary information and that is why the fusion resulted in improvement.

Table 2: Proposed system 1: Mean accuracies and standard deviation with 4-fold Cross-validation on two datasets

	TUT Acoustic Scenes 2016		TUT Acoustic Scenes 2017	
System	**Channel 1**	**Channel 2**	**Channel 1**	**Channel 2**
ST+SVM	48.71±8.48	51.37±7.04	48.93±1.85	50.90±3.73
SCFC+SVM	74.03±2.05	71.80±1.43	74.16±2.51	73.09±2.97
CQCC+SVM	70.70±5.03	70.26±4.26	74.67±2.10	73.92±1.54
NOBT+Δ+SVM	76.23±1.39	70.60±4.32	78.04±1.78	75.57±4.0
Score (Feature)	82.91±3.08	79.66±6.11	82.57±3.86	82.93±2.13
Score (Channel)	85.30±3.88		86.32±2.54	

Table 3 shows the accuracy obtained at both the stages of Proposed System 2 on datasets of both the challenges. This system has shown better performance than the earlier one, with the improvement more prominent on 2016 dataset than on 2017 dataset. Again, length of the data could be given credit for the improvement.

For each acoustic scene, there are 312 segments (52 minutes of audio) in the development dataset of this challenge. The baseline system for the challenge is multilayer perceptron (MLP) based system which used mel-band log energies as features and reported an overall classification accuracy of 74.8%. Both the proposed systems have performed considerably better than baseline system on the given development dataset, wherein the second system's performance was superior. In Fig 3, we pictorially show the class-wise performance of both the proposed systems on the current challenge's development dataset. It can be observed in this picture that most of the misclassification have remained unaffected, probably due to the use of same classifiers in both stages. Nonetheless, misclassifications in broader classes have reduced in the two-stage framework. For example, 'cafe/restaurant' is not marked as 'city_center' or 'tram','library' is not marked as 'train' and 'tram' is not labeled as 'grocery_store' by the hierarchical system. On the other hand, some inter-broad-class misclassifications have newly occurred, like 'city_center' as 'park' and 'forest_path' as 'residential_area'.

Table 3: Proposed system 2: Mean accuracies and standard deviation for the two stages on both the development data

	TUT Acoustic Scenes 2016	TUT Acoustic Scenes 2017
Stage 1	96.27±1.26	96.78±0.82
Stage 2	90.86±1.79	88.78±2.29

5.3. Performance on Evaluation Data

The evaluation dataset for this challenge consists of total 1620 segments (270 min of total data). The reported accuracy of the baseline system on this data is 61.0% (58.7-63.4). The two systems presented in this report gave an accuracy of 67.0% (64.7-69.3) and 64.9% (62.6-67.2) respectively. It is noteworthy here that the Proposed System 2, whose performance was better than the Proposed System 1 on the development dataset, has performed poorly during evaluation. Also, both the system's relative improvement over baseline's performance has deteriorated on evaluation data.

6. CONCLUSION

In this technical report, we have described two systems for the acoustic scene classification task (Task 1) of DCASE challenge 2017. The first system applied fusion of well-known audio processing features to produce classification better than the baseline system. The second system employed the first system in a two-stage hierarchy and performed better than the first system on the development dataset. During evaluation, however, the performance of the first system was better. This could have been caused due to over-fitting, which calls for more analysis during system development.

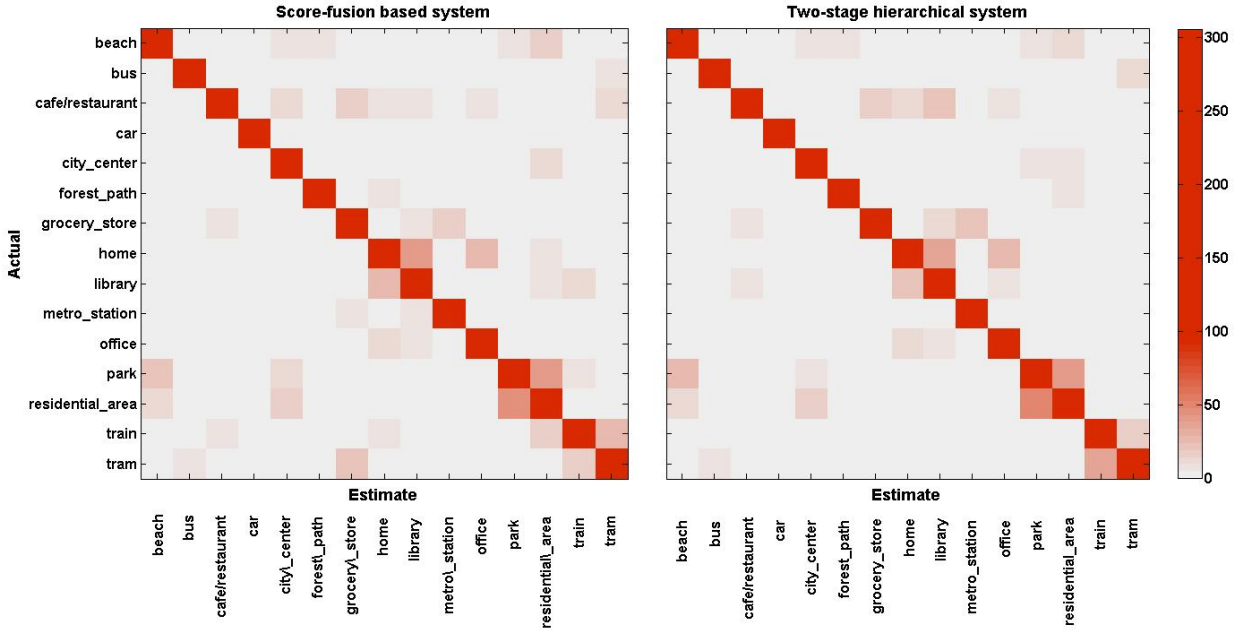

Figure 3: Confusion matrix of results with both proposed systems on TUT Acoustic Scenes 2017 development dataset

It was also observed that most of the classes remained misclassified to a similar extent by both the systems. We expect that use of different classification strategies would alleviate this problem.

7. REFERENCES

[1] 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.

[2] M. Sahidullah and G. Saha, "Design, analysis and experimental evaluation of block based transformation in MFCC computation for speaker recognition," *Speech Communication*, vol. 54, no. 4, pp. 543–565, 2012.

[3] J. M. K. Kua, T. Thiruvaran, M. Nosratighods, E. Ambikairajah, and J. Epps, "Investigation of spectral centroid magnitude and frequency for speaker recognition." in *Odyssey*, 2010, p. 7.

[4] M. Todisco, H. Delgado, and N. Evans, "A new feature for automatic speaker verification anti-spoofing: Constant-Q cepstral coefficients," in *Speaker Odyssey Workshop, Bilbao, Spain*, vol. 25, 2016, pp. 249–252.

[5] V. Ghodasara, S. Waldekar, D. Paul, and G. Saha, "Acoustic scene classification using block-based MFCC features," in *IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Events (DCASE 2016), Budapest, Hungary, Tech. Rep*, 2016.

[6] T. Giannakopoulos and A. Pikrakis, *Introduction to Audio Analysis: A MATLAB® Approach*. Academic Press, 2014.

[7] G. Roma, W. Nogueira, and P. Herrera, "Recurrence quantification analysis features for environmental sound recognition," in *Applications of Signal Processing to Audio and Acoustics (WASPAA), 2013 IEEE Workshop on*. IEEE, 2013, pp. 1–4.

[8] N. Brümmer, "FoCal multi-class: Toolkit for evaluation, fusion and calibration of multi-class recognition scorestutorial and user manual," *Software available at http://sites. google. com/site/nikobrummer/focalmulticlass*, 2007.

[9] A. Mesaros, T. Heittola, and T. Virtanen, "TUT database for acoustic scene classification and sound event detection," in *Signal Processing Conference (EUSIPCO), 2016 24th European*. IEEE, 2016, pp. 1128–1132.

[10] A. Mesaros, T. Heittola, A. Diment, B. Elizalde, A. Shah, E. Vincent, B. Raj, and T. Virtanen, "DCASE 2017 challenge setup: Tasks, datasets and baseline system," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2017 Workshop (DCASE2017)*, November 2017, submitted.