WAVELET-BASED AUDIO FEATURES FOR ACOUSTIC SCENE CLASSIFICATION

Technical Report

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 a submission for IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Events (DCASE) 2018 for Task 1 (acoustic scene classification (ASC)), sub-task A (basic ASC) and sub-task B (ASC with mismatched recording devices). We use two wavelet-based features in a score-fusion framework to achieve the goal. The first feature applies wavelet transform to log mel-band energies, while the second does a high-Q wavelet transformation on the frames of raw signal. The two features are found to be complementary so that the fused system relatively outperforms the deep-learning based baseline system by 17% for sub-task A and 26% for sub-task B with the development dataset provided for the respective sub-tasks.

***Index Terms*— Constant-Q transform, fusion, mel-scaled features, SVM, wavelet transform**

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, outdoor, or a moving vehicle. 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 system presented in this report, we use some spectral and temporal features from audio processing fields. The motivation behind using the spectral features, namely, *Mel-frequency discrete wavelet coefficients* (MFDWC) [2], was to be able to discriminate between acoustic scenes by the spectral characteristics of the specific audio events that characterize them. We use *constant-Q cepstral coefficients* (CQCC) [3] in an attempt to mimic the human hearing system better than mel-scaled features. A set of *short-term (ST) time and frequency* features are also used to complement the two features. Our proposed system employs a fusion-based framework. The classification results from the aforementioned features extracted from monophonic audio streams are *score-fused* to get the final classification.

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

2. BASIC SYSTEM CONFIGURATION

2.1. Features

The proposed systems use the following as features.

- *Mel-frequency discrete wavelet coefficients (MFDWC)* [2]: In all fields of speech processing, mel-frequency cepstral coefficients (MFCC) are the most exploited features. One of the important steps in MFCC extraction is discrete cosine transform (DCT). However, the basis vectors of DCT have approximately the same resolution in time and frequency. Also, because they span the whole frequency range, corruption of a band due to noise affects all the coefficients. These shortcomings can be overcome by using discrete wavelet transform (DWT) instead because it has better time and frequency localization capacity. Unlike Fourier based transforms, wavelet transform uses short basis functions for high-frequency content and long basis functions for low-frequency content of a signal. This makes it suitable for working with audio signals captured from acoustically different surroundings that cover the entire audio frequency range of 20Hz to 20 kHz. DWT applied to mel-filterbank log-energies results in MFDW coefficients. Wavelet based features are especially efficient in characterizing the impulsive parts of the audio [4]. The feature extraction scheme is same as that of MFCC, except that the DWT is applied in place of the DCT.
- *Constant-Q cepstral coefficients (CQCC)* [3]: Audio perception in humans exhibits higher frequency resolution at lower frequencies and a higher temporal resolution at higher frequencies. This is equivalent to having a set of filters with constant Q-factor across the entire spectrum. Geometrically spaced frequency bins can be employed to achieve this objective. Constant-Q transform (CQT) implements the same and is commonly used in music signal processing. It is similar as wavelet transform but with a high Q-factor.The coupling of CQT with traditional cepstral analysis resulted in constant-Q cepstral coefficients (CQCC) [3].
- *Short-term (ST) time and frequency features* [5]: Short-term features, such as zero crossing rate, 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 [5]. Since acoustic scenes are a collection of multiple environmental sounds, these features are expected to

add to the information captured by cepstral features.

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 [6]. 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 first convert the audio to monophonic by averaging the two channels [7]. 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 weightage 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

The block diagram for the proposed system is shown in Fig. 1 which is similar to first system in [9]. In this system, 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. The weights for fusion were obtained from test portion of the dataset 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.

3.1. Experimental Framework

We have used the the development dataset of TUT Acoustic Scenes 2016 (TUTAS16D) [10], TUT Acoustic Scenes 2017 (TUTAS17D) [11] and TUT Urban Acoustic Scenes 2018 (basic (TUTUAS18D) for sub-task A and mobile (TUTUASM16D) for sub-task B [12] in our experiments. The first two datasets differ from each other in the length of the audio streams (30 sec for first and 10 sec for second) and size of the datasets (the second one is larger). The third and the fourth datasets, which are the development data for the present challenge, differ from the first two in the classes. From all the data samples, MFDWC were extracted by applying Hamming window on 40 ms frame having 50% overlap. Pre-emphasis to the audio signals was done by a factor of 0.97. Filterbank of 80 filters was used for MFDWC (triangular filters). Delta (Δ) features, evaluated with a 3-frame window, were appended only for MFDWC. The parameters for CQCC features given in [3] were used here. ST features extraction was same as that in [7]. Frame-wise mean and standard deviation of the features were given as input to SVM classifier with RBF kernel. According to the previous DCASE challenges' ASC task setup, development data is partitioned into k folds, where k=4 for both 2016 and 2017. Fold-wise mean classification accuracy was used as the performance metric. However, in the present challenge, only one train and test partition was given i.e. k=1, and therefore the results are evaluated for the given test subset.

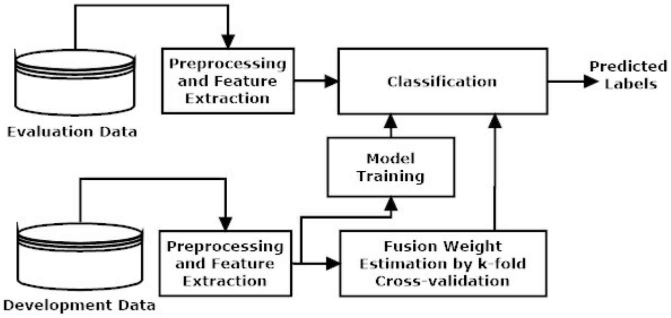
Evaluation Data
Preprocessing and Feature Extraction
Classification
Predicted Labels
Model Training
Development Data
Preprocessing and Feature Extraction
Fusion Weight Estimation by k-fold Cross-validation

Figure 1: Block diagram of fusion-based proposed system

4. RESULTS

The results of the three features on the three datasets are shown in Table 1. It can be seen here that although the features perform differently, they all surpass the chance-level accuracy of the three datasets (6% for first two datasets and 10% for the third dataset). Nevertheless, the three features carry complementary information and that is why the fusion resulted in improvement.

In the present challenge, class-wise mean accuracy is used as the metric. The mean accuracy of all classes reported for the logMBE-CNN baseline system for sub-task A is 59.7% [13]. Thus, by obtaining a mean class-wise accuracy of 69.81%, our proposed system has relatively outperformed the deep-learning based baseline by 17% with the development dataset of sub-task A. Class-wise performance comparison of the two systems for this sub-task is depicted in Fig. 2. The baseline system's worst performance is on the 'public_square' class, while most misclassifications have occurred with 'shopping_mall' being categorized as 'airport'. The darker shades in the diagonal of the proposed system's confusion matrix exhibit its superiority in all scene classes. It has shown its worst performance in 'street_pedestrian' with many samples of this class being wrongly put into 'public_square' category. The classes 'park' and 'street_traffic' are well-classified by both systems.

The performance of the proposed system for the three devices of sub-task B is shown in Table 2 alongwith the reported results of the baseline system. It can be seen that the proposed fusion-based system performed better than the baseline for all three devices. According to the rules of the challenge for this sub-task, the ranking of the systems will be done by the average performance with devices B and C only. The reported baseline accuracy in this case is 45.6%. Our proposed framework achieved 57.78%, which is 26% relatively

Table 1: Proposed system performance on three datasets in terms of mean accuracy (%).

Feature	TUTAS16D	TUTAS17D	TUTUAS18D	TUTUASM18D
ST-SVM	42.56±7.11	49.40±2.37	45.95	42.15
CQCC-SVM	73.94±7.24	76.01±1.75	63.07	60.53
MFDWC-SVM	79.15±0.42	82.27±1.70	65.57	63.24
Proposed	84.98±3.23	84.42±2.48	70.02	66.68

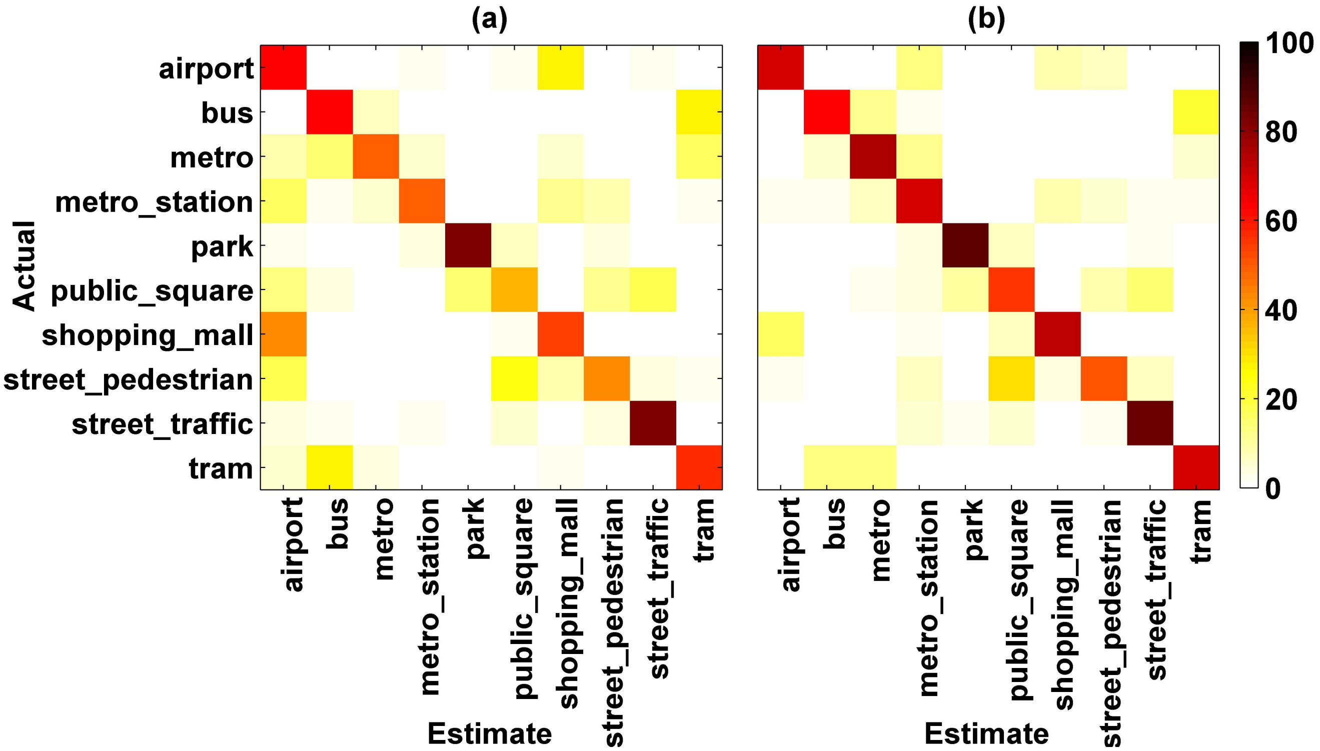

Figure 2: Confusion matrix of results with (a) logMBE-CNN based baseline system and (b) score-fusion based proposed system with TUT Urban Acoustic Scenes 2018 development dataset, sub-task A.

better.

Table 2: Proposed system performance for the three recording devices in sub-task B in terms of mean accuracy (%).

System	Device A	Device B	Device C
ST-SVM	43.76	21.67	40.00
CQCC-SVM	62.23	48.33	50.56
MFDWC-SVM	63.94	59.44	57.22
Proposed	68.51	58.89	56.67
Baseline	58.90	45.10	46.20

5. CONCLUSION

In this technical report, we have described a system for acoustic scene classification task (Task 1, Sub-task A and Sub-task B) of DCASE challenge 2018. The first sub-task is concerned with the basic problem of ASC, in which all available data (development and evaluation) are recorded with the same device. On the other hand, sub-task B addresses the situation in which an application will be tested with a few different types of devices, possibly not the same as the ones used to record the development data. Our system applied fusion of well-known audio processing features to produce classification better than the baseline system on both the sub-tasks.

6. 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] J. N. Gowdy and Z. Tufekci, “Mel-scaled discrete wavelet coefficients for speech recognition,” in *Acoustics, Speech, and Signal Processing, 2000. ICASSP’00. Proceedings. 2000 IEEE International Conference on*, vol. 3. IEEE, 2000, pp. 1351–1354.

[3] 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.

[4] A. Rabaoui, M. Davy, S. Rossignol, and N. Ellouze, "Using one-class SVMs and wavelets for audio surveillance," *IEEE Transactions on information forensics and security*, vol. 3, no. 4, pp. 763–775, 2008.

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

[6] 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.

[7] S. Waldekar and G. Saha, "Classification of audio scenes with novel features in a fused system framework," *Digital Signal Processing*, 2018.

[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] S. Waldekar and G. Saha, "IIT Kharagpur submissions for DCASE2017 ASC task: Audio features in a fusion-based framework," DCASE2017 Challenge, Tech. Rep., September 2017.

[10] 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.

[11] 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.

[12] A. Mesaros, T. Heittola, and T. Virtanen, "A multi-device dataset for urban acoustic scene classification," *arXiv preprint arXiv:1807.09840*, 2018.

[13] http://dcase.community/challenge2018/.