ACOUSTIC SCENE CLASSIFICATION USING BLOCK BASED MFCC FEATURES

Vikaskumar Ghodasara, Shefali Waldekar, Dipjyoti Paul and Goutam Saha

Electronics & Electrical Communication Engineering Department,

Indian Institute of Technology Kharagpur, India

ghodasara.vikas@gmail.com, shefaliw@ece.iitkgp.ernet.in, dipjyotipaul@ece.iitkgp.ernet.in and gsaha@ece.iitkgp.ernet.in

ABSTRACT

Acoustic Scene Classification (ASC) is receiving wide spread attention due to its wide variety of applications in smart wearable devices, surveillance, life log diarization etc. This work describes our contribution to the Acoustic scene classification task of the DCASE2016 Challenge for Detection and Classification of Acoustic Scenes and Events. In this work, we apply block based MFCC along with few traditional short term audio features with mean and standard deviation as statistics and Support Vector Machine (SVM) as a classifier to ASC. It is observed that block based MFCC feature performs better than classical MFCC. For evaluation purpose, we used three different datasets.

Index Terms**—** Acoustic scene classification, SVM, MFCC, Block based MFCC

1. INTRODUCTION

The problem of recognizing acoustic environment sound from which the sound was recorded is known as the problem of audio scene classification. Discrimination between different types of audio signals is an easy task for the human auditory system. However, this task is considered not to be so trivial for machines. Yet, no real-time algorithm exists that can replicate the human auditory system. Computational Auditory Scene Analysis (CASA) is the domain in which research, study and design related to “machine listening” is going on. Computational Auditory Scene Recognition (CASR) is one of the parts of CASA. Audio scene classification is a complex problem due to the wide variety of individual sound events occurring in an audio scene while only few of them give some information about the scene.

Early works in ASC were inspired by speech recognition systems; features like Mel Frequency Cepstral Coefficients (MFCC) [1] [2] have been widely used and they are often used as a baseline system for ASC [3] [4]. These features are used along with several other traditional features such as low level features (zero-crossing rate, spectral centroid, spectral roll-off, voicing related features, band energies) [5]. In this work, we also follow the established practice of using MFCC features and traditional audio features but follow a novel method of block based MFCC calculation which has been successfully used in automatic speaker recognition application [6]. Our work significantly improves the state of the art results on large ASC datasets.

The rest of the paper is organized as follows. Section 2 describes the feature extraction. Section 3 details the experimental setup and datasets. Section 4 describes the experiment evaluation and results. Finally section 5 presents conclusions.

2. FEATURE EXTRACTION

2.1. Mel Frequency Cepstral Coefficients (MFCC)

The mel frequency cepstrum has proven to be highly effective in recognizing structure of music signals and in modeling the subjective pitch and frequency content of audio signals. Psychophysical studies have found the phenomena of the mel pitch scale and critical band, and the frequency scale-warping to the mel scale has led to the cepstrum domain representation. The mel scale is defined as

$$F_{mel} = 2595 \times \log_{10}\left(1 + \frac{f}{700}\right) \tag{1}$$

where F_{mel} is the logarithmic scale of normal frequency scale represented by f.

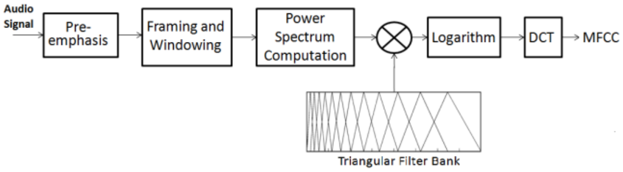

Figure 1.1: Different stages for MFCC feature extraction process

Pre-emphasis, which is a high pass filtering process, is used to boost the spectrum of voiced sounds which suffer a steep roll-off of 6 dB/octave towards high frequency region. Pre-emphasis is followed by framing and windowing of the filtered signal. Speech is quasi-stationary signal due to slow varying nature of vocal tract. Therefore, to estimate the spectral characteristics, signal should be analyzed over shorter duration frames (20-30ms). Adjacent frames are overlapped to preserve the boundary information. This is followed by estimation of power spectrum, multiplication with filter banks, logarithm of the result and discrete cosine transform (DCT) to generate de-correlated feature vectors.

2.2. Block Based Mel Frequency Cepstral Coefficients

In MFCC computation, as seen in previous section, DCT is applied on all the log energy coefficients to de-correlate the features. In [6], it was observed that two non-overlapping blocks with first block covering approximately the log energies of frequency bands equal to span of first formant and second block on remaining two formants, give better performance in speaker recognition task. In [7], it is shown that the block based MFCC features perform better in speech spoofing attack detection task.

2.3. Proposed Feature Extraction Method

In [6], it is observed that first formant frequency range is around 1 kHz and remaining two formants between 1 kHz – 4 kHz. Accordingly, the first DCT block was chosen to cover frequencies up to 1 kHz and the second DCT block covered remaining frequencies. Note that our work is different from speech based speaker recognition task, because audio frequency of interest is more than speech frequency range. Here, we compute MFCC over three blocks; first two blocks were same as in [6], the third block range is from 4 kHz to half of the sampling frequency, because in case of sound other than speech, the sampling frequency required is greater than 8 kHz (the usual sampling frequency for speech). The Cepstral coefficients $\{X_i\}$, using non-overlapping three blocks, with first and second block size q and r respectively, can be expressed as:

$$\{X_i\}_{i=1}^{q-1} = \sqrt{\frac{2}{q}} \sum_{j=0}^{q-1} \psi(j+1) \times \cos\left(\frac{i\pi(2j+1)}{2q}\right) \quad (2)$$

$$\{X_i\}_{i=q+1}^{\alpha-1} = \sqrt{\frac{2}{r}} \sum_{j=0}^{r-1} \psi(j+q+1) \times \cos\left(\frac{\pi(i-q)(2j+1)}{2r}\right) \quad (3)$$

$$\{X_i\}_{i=\alpha+1}^{p-1} = \sqrt{\frac{2}{p-\alpha}} \sum_{j=0}^{p-(\alpha+1)} \psi(j+\alpha+1) \times \cos\left(\frac{\pi(i-\alpha)(2j+1)}{2(p-\alpha)}\right) \quad (4)$$

where $\alpha = q + r$ and ψ is vector of p filter banks log energies.

For baseline system MFCC and 1st order derivative of MFCC, termed as delta (Δ) are considered as features. We use block based MFCC for ASC task, which is shown to outperform traditional MFCC features in speech-music classification task of MIREX2015 previously [8].

In this work, to have a better DCT block range in MFCC calculation, we tried out various combinations. Out of those combinations, DCT blocks applied on log energies covering frequencies up to 1 kHz, 1 kHz to 4 kHz, 4 KHz and beyond are found to give better overall classification accuracies.

3. EXPERIMENTAL SETUP & DATASETS

We evaluate the block based MFCC features along with traditional audio features (zero crossing rate, energy, entropy of energy, spectral centroid, spectral spread, spectral flux, spectral roll off point, spectral entropy, harmonic ratio, fundamental frequency) and the Support Vector Machine (SVM) classifier. When classes are not linearly separable in the original feature space, the feature space can be transformed to a higher-dimensional space using a nonlinear kernel function. In the present work, the system employs SVMs with a Radial Basis Function (RBF kernel). LIBSVM toolbox [9] is used in our work.

The following three data sets are used in our experiment.

1. DATASET-1 is a combination of development and evaluation datasets of Dcase2013 challenge on ASC [3].This dataset contains 10 different classes; for each class 20 samples, each 30 seconds long. Each class corresponds to a specific location, such as in a supermarket, in a restaurant, or at the office.

2. DATASET-2 is the LITIS Rouen dataset [10]. It is one of the largest publicly available data set for ASC. This data set was recorded with a smart phone. It contains 19 different classes; total 3026 samples, each of 30 seconds length. For each class, numbers of samples are non-uniform. Each class corresponds to a specific location, such as in a kid game hall, in a café, or at shop.

3. DATASET-3 is the development dataset of Dcase2016 challenge on ASC [4]. This dataset contains 15 different classes; for each class 78 samples, each 30 seconds long. Each class corresponds to a specific location, such as in a library, in a grocery store, or at a beach.

We used the same multiple-fold cross validation as suggested by the creators of the datasets to generate comparable results. The results presented are averaged over all the folds. In order to estimate the best regularization parameter and the best gamma parameter for the radial basis kernel, we perform a grid search on these parameters for each cross validation iteration.

4. EXPERIMENTAL EVALUATION & RESULTS

A series of experiments were performed to investigate the behavior of the system and to analyze the contribution of different parameters and system components to the classification performance. Implemented system is first evaluated with DATASET-1. For evaluation, 5-fold cross validation was performed, which is the official protocol of the challenge.

Block based MFCC + Δ block based MFCC + 10 short term audio features (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 computed on 50% overlapping, 20ms frames with varying number of filter banks for log energy calculation and varying number of blocks for block based MFCC calculation. Over all frames, mean and standard deviation were evaluated and these were considered as feature vector for that audio track. Table 3.1, Table 3.2 and Fig. 3.1 summarize the ASC performance for DATASET-1. We get the best overall accuracy in our experiments as 86% for three blocks and 60 filter bank combination.

Table 3.1: Average acoustic scene classification accuracy by varying number of filter banks and number of block in block based MFCC features (DATASET-1)

No. Of filter banks	**MFCC (B1)**	**Block based MFCC (2 blocks) (B2)**	**Block based MFCC (3 blocks) (B3)**
20	78.50 ± 3.79	73.00 ± 3.26	74.00 ± 8.02
30	79.00 ± 2.85	77.50 ± 5.86	82.50 ± 3.95
40	81.50 ± 5.18	81.50 ± 1.36	84.00 ± 5.18
50	81.00 ± 3.79	83.50 ± 3.79	85.00 ± 6.12
60	81.00 ± 5.47	82.50 ± 3.95	**86.00 ± 8.02**
70	80.00 ± 3.06	81.50 ± 4.18	85.00 ± 4.33
80	78.50 ± 2.85	80.50 ± 4.81	83.00 ± 4.11

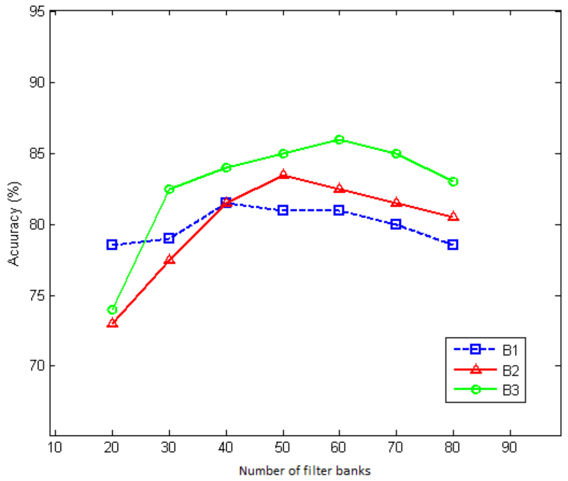

Figure 3.1: Average ASC accuracy by varying number of filter banks and number of block in block based MFCC features (DATASET-1) (B1 means DCT applied as whole, B2 means DCT applied in two blocks and B3 means DCT applied in three blocks in MFCC feature extraction)

Table 3.2: Confusion matrix of overall classification (DATASET-1)

	Bus	Busy street	Office	Open air market	Park	Quiet street	Restaurant	Supermarket	Tube	Tube station
Bus	**16**	0	0	0	1	0	0	0	3	0
Busy street	1	**19**	0	0	0	0	0	0	0	0
Office	0	0	**18**	0	2	0	0	0	0	0
Open air market	0	0	0	**17**	0	0	1	2	0	0
Park	1	0	0	0	**17**	2	0	0	0	0
Quiet street	0	2	0	0	2	**16**	0	0	0	0
Restaurant	0	0	0	0	0	0	**19**	0	0	1
Supermarket	0	1	0	0	0	0	0	**19**	0	0
Tube	2	0	0	0	0	0	0	0	**17**	1
Tube station	0	1	0	0	0	1	1	2	1	**14**

DATASET-2 was evaluated with same framework. For evaluation, 20-fold cross validation was performed, which is provided by creators of this dataset. Table 3.3 summarizes the result where we see the best performance of 93.51% occurring for three blocks and 80 filter bank combination.

Table 3.3: Average acoustic scene classification accuracy by varying number of filter banks and number of block in block based MFCC features (DATASET-2)

No. of filter banks	**MFCC (B1)**	**Block based MFCC (2 blocks) (B2)**	**Block based MFCC (3 blocks) (B3)**
20	86.73 ± 1.26	87.57 ± 1.22	87.37 ± 1.01
30	90.24 ± 1.16	89.73 ± 0.98	89.93 ± 1.18
40	91.10 ± 0.97	91.29 ± 0.86	91.57 ± 0.65
50	92.23 ± 1.01	92.18 ± 0.99	92.89 ± 0.83
60	92.19 ± 0.99	92.87 ± 0.73	92.64 ± 0.84
70	92.68 ± 1.03	92.92 ± 0.81	93.01 ± 1.07
80	93.08 ± 0.68	93.36 ± 0.93	**93.51 ± 1.09**
90	93.17 ± 0.75	93.48 ± 0.92	93.24 ± 0.84

In both these datasets we observe that block based MFCC features perform better as compared to traditional MFCC. On DATASET-3 with 4-fold cross validation we get overall accuracy of 80.42%. In this case, we chose the combination of three blocks and 60 filter banks which gives best performance for DATASET-1, since DATASET-1 and DATASET-3 are quite similar except that DATASET-3 has more classes.

Table 3.4: Confusion matrix of overall classification (DATASET-2)

	Plane	Bus	Busy street	Cafe	Car	Student hall	Train station hall	Kid game hall	Market	Metro - paris	Metro - rouen	Billiard pool hall	Quiet street	Restaurant	Pedestrian street	Shop	Train	High speed train	Tube station
Plane	**96**	0	0	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Bus	0	**756**	1	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0
Busy street	0	5	**476**	2	0	0	9	0	16	16	25	0	31	0	0	0	0	0	0
Café	0	0	4	**425**	0	0	5	0	4	14	0	1	0	0	20	4	0	0	3
Car	0	0	0	0	**970**	0	0	0	0	5	0	0	0	0	0	2	0	3	0
Student hall	0	0	0	0	0	**331**	2	0	0	0	0	0	6	10	6	5	0	0	0
Train station hall	0	0	2	1	0	0	**1047**	0	1	5	4	5	0	1	4	6	0	0	4
Kid game hall	0	0	0	0	0	0	0	**580**	0	0	0	0	0	0	0	0	0	0	0
Market	0	0	1	5	0	0	2	0	**1071**	0	0	1	1	2	4	13	0	0	0
Metro - paris	0	2	6	0	0	0	17	0	5	**471**	52	0	0	2	0	4	1	0	0
Metro - rouen	0	3	16	0	0	0	0	0	3	34	**937**	0	3	0	0	2	0	0	2
Billiard pool hall	0	0	6	0	0	0	0	0	0	0	0	**613**	0	0	0	1	0	0	0
Quiet street	0	0	22	7	0	2	2	0	0	1	1	3	**290**	0	16	12	0	0	4
Restaurant	0	0	0	0	0	7	1	0	1	0	0	0	0	**531**	0	0	0	0	0
Pedestrian street	0	0	8	32	0	0	12	0	20	0	0	0	18	2	**371**	17	0	0	0
Shop	0	1	8	13	0	2	10	0	30	2	1	0	5	4	12	**730**	0	0	2
Train	0	0	2	0	0	0	1	0	0	8	0	0	1	0	0	0	**644**	4	0
High speed train	0	0	0	0	0	0	0	0	0	6	13	0	4	0	0	0	0	**557**	0
Tube station	0	0	1	1	0	0	3	0	8	10	2	0	1	0	15	3	0	0	**456**

Table 3.5: Confusion matrix of overall classification (DATASET-3)

	Beach	Bus	Café/restaurant	Car	City centre	Forest path	Grocery store	Home	Library	Metro station	Office	Park	Residential area	Train	Tram
Beach	**59**	0	0	2	1	3	0	0	1	0	0	0	9	0	3
Bus	0	**59**	1	7	0	0	0	0	0	0	0	4	0	4	3
Café/restaurant	0	0	**67**	0	0	0	5	2	0	1	0	1	0	1	1
Car	0	2	0	**68**	0	0	0	0	0	0	0	0	1	7	0
City centre	0	0	1	0	**72**	0	0	0	0	3	0	0	2	0	0
Forest path	0	0	1	0	0	**75**	0	1	0	0	0	0	0	1	0
Grocery store	0	0	5	0	0	0	**58**	0	8	6	0	0	0	1	0
Home	2	0	0	0	0	0	0	**61**	6	0	3	4	0	2	0
Library	0	0	1	0	0	0	4	0	**67**	1	2	2	0	1	0
Metro station	0	0	1	0	3	0	2	0	0	**72**	0	0	0	0	0
Office	0	0	0	0	0	0	0	4	2	0	**71**	0	1	0	0
Park	5	0	0	0	3	1	0	0	2	0	0	**57**	9	1	0
Residential area	11	0	0	0	4	3	0	0	0	0	0	13	**47**	0	0
Train	0	11	7	6	0	0	0	0	1	0	0	0	3	**45**	5
Tram	2	0	1	1	0	0	4	0	2	0	0	1	0	4	**63**

From Table 3.2, it is seen that “Park” and “Quiet street”, and “Bus” and “Tube” classes misclassify with each other. From Table 3.4, it is observed that “Busy street” and “Quiet street”, “Shop” and “Market” and “Metro-paris” and “Metro-rouen” classes misclassify with each other. From the Table 3.5, it is clear that “Park”, “Residential area” and “Beach” classes misclassify among each other. Similarly, “Bus” and “Train” as well as “Train” and “Car” class misclassify with each other. “Forest path”, “City center”, “Metro station” and “Office” classes have classification accuracy of more than 90%.

5. CONCLUSION

In this paper, we proposed an acoustic scene classification system based on block based MFCC features and few traditional audio features. The number of extracted features is determined by the number of filter banks used in MFCC feature extraction. In order to compare our block based MFCC and SVM classifier system with other proposed methods, we evaluated it on the basis of the publicly available datasets for scene classification, namely Dcase2013 challenge development and evaluation dataset and LITIS Rouen dataset. With 86 % overall classification accuracy on DATASET-1, our classifier significantly outperforms the best algorithm submitted to the challenge (76% [11]) and with 93.51% overall accuracy on DATASET-2 it outperforms previous best result which is 93.4% [12]. On DATASET-3, we get overall accuracy of 80.42%. Number of filter banks and parameters of SVM kernel that give best performance are dependent on the training data. With different training samples and different number of classes of acoustic scenes we will obtain different parameters.

6. REFERENCES

[1] Stowell, Dan, Dimitrios Giannoulis, Emmanouil Benetos, Mathieu Lagrange, and Mark D. Plumbley. "Detection and classification of acoustic scenes and events." *IEEE Transactions on Multimedia* 17, no. 10 (2015): 1733-1746.

[2] Annamaria Mesaros, Toni Heittola, and Tuomas Virtanen, *Tut database for acoustic scene classification and sound event detection*, In 24rd European Signal Processing Conference 2016 (EUSIPCO 2016). Budapest, Hungary, 2016.

[3] http://c4dm.eecs.qmul.ac.uk/sceneseventschallenge/.

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

[5] J. T. Geiger, B. Schuller, and G. Rigoll, “Large-scale audio feature extraction and svm for acoustic scene classification,” in *Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2013.

[6] Sahidullah Md, and Goutam Saha. "Design, analysis and experimental evaluation of block based transformation in MFCC computation for speaker recognition." Speech Communication 54, no. 4 (2012): 543-565

[7] D. Paul, M. Pal and G. Saha, "Novel speech features for improved detection of spoofing attacks," *2015 Annual IEEE India Conference (INDICON)*, New Delhi, 2015, pp. 1-6.

[8] music-ir.org/mirex/wiki/2015:Music/Speech_Classification_and_Detection_Results

[9] Chang, Chih-Chung, and Chih-Jen Lin. "LIBSVM: a library for support vector machines." ACM Transactions on Intelligent Systems and Technology (TIST) 2.3 (2011):27.

[10] LITIS Rouen Dataset: https://sites.google.com/site/alainrakotomamonjy/home/audioscene

[11] 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.

[12] V. Bisot, S. Essid, and G. Richard, “Hog and subband power distribution image features for acoustic scene classification,” in 2015 European Signal Processing Conference (EUSIPCO), Nice, France, August 2015, pp. 724–728.