ACOUSTIC SCENE CLASSIFICATION USING MFCC AND MP FEATURES

Manjunath Mulimani, Shashidhar G. Koolagudi

National Institute of Technology, Karnataka
Dept. of Computer Science & Engineering
Surathkal, India
manjunath.gec@gmail.com, koolagudi@yahoo.com

ABSTRACT

This paper, clearly describes our experiments for efficient acoustic scene classification task as a part of "Detection and Classification of Acoustic Scenes and Events-2016 (DCASE-2016)" IEEE Audio and Acoustic Signal Processing (AASP) challenge. Identification of features from given audio clips to appropriate acoustic scene classification is a challenging task because of heterogeneity by thier nature.

In order to identify such features, in this paper we have implemented few methods using Matching Pursuit (MP) algorithm in order to extract Time-Frequency (TF) based features. MP algorithm is used to select atoms iteratively among the set of parameterized waveforms in the dictionary that best correlates the original signal structure. Using these selected set of atoms mean and standard deviation of amplitude and frequency parameters of first few (n) atoms are calculated separately, resulting into four MP feature sets. Combination of twenty MFCCs along with four MP features enhanced the recognition accuracy of acoustic scenes using GMM classifier.

Index Terms— Matching Pursuit algorithm (MP), Mel-frequency cepstral coefficient (MFCC), Gaussian mixture model (GMM).

1. INTRODUCTION

Aim of any acoustic scene classification methods is to characterize the various scenes in the surrounding environment. Identification of acoustic scenes through the analysis of unstructured sound patterns is an interesting audio signal processing problem because of versatile applications such as, design of context aware mobile devices, robotics, cars, intelligent monitoring systems and so on.

Significant research has been done in the area of acoustic scene classification in the past few decades. Availability of well collected data sets with different environmental sounds, such as DCASE-2016 dataset [1] has attracted many researchers recently. Along with this standard dataset East Anglia dataset [2], Litis Ronen dataset [3] are also available. Much of the existing work has concentrated on developing various feature extraction algorithms specific to the acoustic scenes. Some of the explored features are time dependent temporal features, frequency dependent spectral features and combined Time-Frequency (TF) features [4][5] obtained using the Matching Pursuit (MP) algorithm[6]. The MP algorithm decomposes an audio signal into a linear expansion of waveforms that are selected from the predefined dictionary [6]. The dictionary is a collection of parameterized waveforms. Each waveform is called as an atom [4]. The atom is characterized based on the type of its dictionary such as Gabor functions, Haar wavelets, Fourier functions and so on, other parameters such as time, frequency, phase, window type, length and so on, are also used to characterize the atoms. Here, we used Gabor functions with MP algorithm those gives both time and frequency features. MP algorithm is an iterative algorithm [7]. The flow chart of MP algorithm is shown in Fig. 1, Where r_i is the residual of a signal at i^{th} iteration, a_i is an atom at index i, y is the original audio signal initialize it to r_1. In the first iteration, MP algorithm calculates the inner product of residual with atoms in the dictionary. The atom with the highest value of the inner product is selected first. This selected atom is subtracted from the residual. This process continues with updated residual and atoms in the dictionary until specified number of iterations reached [4]. This overall process is called as signal decomposition. In each iteration, the MP algorithm chooses the atom that best matches with signal structure. Gabor dictionary with MP algorithm gives TF features. Work carried out in [4][5], uses these features for environmental sound classification. Few of these feature extraction methods implemented in this paper for the task of acoustic scene classification. Matching Pursuit Tool Kit's (MPTK's) [7] MP algorithm is used in our experiments.

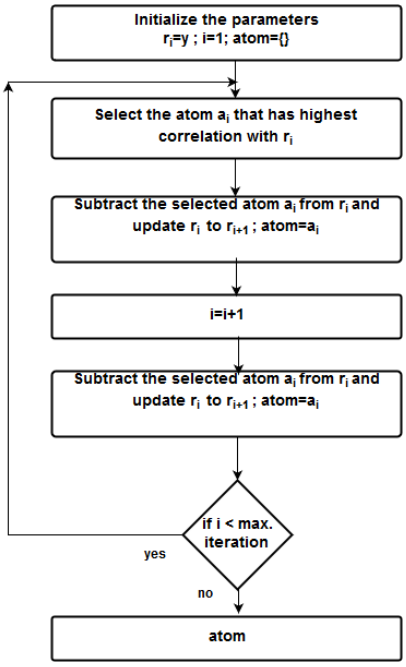

Figure 1: Flow chart of MP algorithm

2. PROPOSED METHOD

Steps involved in the proposed methodology described below in brief.

2.1. Dataset

In this technical report, we considered DCASE-2016 development dataset [1][8], it consists of 30 seconds audio clips of 15 different acoustic scenes such as beach, bus, cafe, car, city-center, forest-path, grocery, home, library, metro, office, park, residential area, train and tram.

2.2. Feature Extraction

First, each 30 seconds of audio signal is divided into a number of frames. Each frame of 40 ms in length contains 1764-samples with the overlap is 20 ms in length. A Gaussian window is applied for windowing the frames. For each frame, following features are extracted [9].

2.2.1. MP Features

As discussed earlier, MPTK's [7] MP algorithm is used in our experiments. MPTK uses different dictionary blocks with various scales. Here, single block dictionary consists of 1764 length Gabor atoms is created. Next, each 1764-samples signal frame is decomposed using MP with the dictionary of Gabor atoms that are also 1764-samples in length as same as in [4]. During decomposition, MP algorithm chooses the atom that best matches the signal structure in each iteration. Experimentally found that 1000 iterations large enough to approximate each frame signal structure. As we already discussed, the atom with the highest energy is selected first in each iteration. The most important information to describe the signal is found in first few atoms. Here, first, 10 atoms with their parameters amplitude and frequency are considered for feature evaluation. Mean and standard deviation of amplitude and frequency of first 10 atoms calculated, resulting into 4 MP features.

2.2.2. MFCC Features

20-MFCC features extracted from each frame. 4-MP features appended with 20-MFCC features, resulting into 24 features used for classification. 20-MFCC features followed by mean of amplitude, standard deviation of amplitude, mean of frequency, standard deviation of frequency used in order.

2.3. Classification

Here, same baseline system [1] GMM is used for classification. 16 Gaussians, 40 iterations, diagonal covariance matrix, 0.001 minimum covariance are considered for classification.

3. RESULTS

GMM is trained on DCASE-2016 development dataset with 4-fold cross validation. The resultant confusion matrix is shown in Table.1 The proposed method achieved a total classification accuracy of 66.8%. The accuracy of each class on development dataset is shown in Table. 2. Three classes office, tram, city achieved more than 90% of accuracy. Ten classes metro-station, home, car, residential and beach, library, cafe, bus, forest-path, grocery-store achieved overall more than 75% of accuracy and 60% of accuracy respectively. Train and park achieved the least accuracy. Park is misclassified as residential area 44% of the time. Train is misclassified as tram 39% of the time. Reason for park and residential area misclassification is both are outdoor scenes and both have most of similar audible events. Similarly train and tram are vehicles seem to have most of similar characteristics. The performance of our proposed system tested on evaluation dataset. Overall 65.6% accuracy achieved. Classwise performance on evaluation dataset as shown in Table. 3. Table 4 and 5 corresponds to the classwise performance of baseline system with MFCC features on development and evaluation datasets respectively.

Due to the high time complexity, here we only considered first 10 atoms of the MP algorithm for MP feature evaluation. Varying this number may improve the accuracy further.

4. REFERENCES

[1] A. Mesaros, T. Heittola, and T. Virtanen, "TUT Database for Acoustic Scene Classification and Sound Event Detection," in *24th European Signal Processing Conference 2016 (EUSIPCO 2016). Budapest, Hungary*, 2016.

[2] L. Ma, B. Milner, and D. Smith, "Acoustic environment classification," *ACM Trans. Speech Lang. Process.*, vol. 3, no. 2, pp. 1–22, July 2006. [Online]. Available: http://doi.acm.org/10.1145/1149290.1149292

[3] A. Rakotomamonjy and G. Gasso, "Histogram of gradients of time-frequency representations for audio scene detection," *arXiv preprint arXiv:1508.04909*, 2015.

[4] S. Chu, S. Narayanan, and C.-C. J. Kuo, "Environmental sound recognition with time–frequency audio features," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 6, pp. 1142–1158, 2009.

[5] B. Ghoraani and S. Krishnan, "Time–frequency matrix feature extraction and classification of environmental audio signals," *IEEE transactions on audio, speech, and language processing*, vol. 19, no. 7, pp. 2197–2209, 2011.

[6] S. G. Mallat and Z. Zhang, "Matching pursuits with time-frequency dictionaries," *IEEE Transactions on signal processing*, vol. 41, no. 12, pp. 3397–3415, 1993.

[7] S. Krstulovic and R. Gribonval, "MPTK: Matching Pursuit made tractable," in *Proc. Int. Conf. Acoust. Speech Signal Process. (ICASSP'06)*, vol. 3, Toulouse, France, May 2006, pp. III–496 – III–499.

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

[9] I. A. Challenge and A. Scenes, "IEEE AASP SCENE CLASSIFICATION CHALLENGE USING HIDDEN MARKOV MODELS AND FRAME BASED CLASSIFICATION May Chum , Ariel Habshush , Abrar Rahman , Christopher Sang The Cooper Union Electrical Engineering Department 41 Cooper Square New York , NY 10003," no. 3, pp. 3–5.

	beach	bus	cafe	car	city	forest	groc	home	lib	metro	offi	park	res	train	tram
beach	**54**	1	1	0	6	0	0	0	1	0	0	0	9	1	5
bus	0	**45**	0	1	0	0	0	0	0	0	0	0	0	13	19
cafe	0	0	**46**	0	0	0	13	6	1	7	0	0	0	0	5
car	0	1	0	**59**	0	0	0	0	0	0	0	0	0	0	18
city	0	0	1	0	**71**	0	0	1	0	0	0	0	5	0	0
forest	1	0	3	0	0	**45**	0	6	0	3	9	2	9	0	0
grocery	0	0	13	0	2	0	**44**	0	5	14	0	0	0	0	0
home	0	0	1	0	0	1	1	**62**	13	0	0	0	0	0	0
library	0	1	0	0	0	0	4	15	**49**	1	1	0	3	3	1
metro	0	0	4	0	0	0	2	1	1	**67**	2	0	0	0	1
office	0	0	0	0	0	0	0	6	0	0	**72**	0	0	0	0
park	2	0	2	0	1	5	2	3	11	0	4	**13**	33	1	1
residential	1	0	1	0	2	5	0	0	3	1	0	8	**55**	1	1
train	0	8	7	1	0	0	0	0	0	0	0	0	2	**30**	30
tram	0	5	0	0	2	0	0	0	0	0	0	0	0	0	**71**

Table 1: Confusion matrix of proposed system on development dataset

Class	beach	bus	cafe	car	city	forest	groc	home	lib	metro	offi	park	res	train	tram	overall
Accuracy	69.2%	58.1%	58.8%	75.4%	91.1%	56.2%	55.8%	78.2%	63.1%	85.2%	92.1%	16.4%	71.2%	39.0%	92.0%	66.8%

Table 2: Class wise Accuracy of proposed system on development dataset

Class	beach	bus	cafe	car	city	forest	groc	home	lib	metro	offi	park	res	train	tram	overall
Accuracy	73.1%	96.2%	69.2%	100%	73.1%	50%	65.4%	76.9%	7.7%	76.9%	96.2%	96.2%	23.1%	15.4%	65.4%	65.6%

Table 3: Class wise Accuracy of proposed system on evaluation dataset

Class	beach	bus	cafe	car	city	forest	groc	home	lib	metro	offi	park	res	train	tram	overall
Accuracy	69.3%	79.6%	83.2%	87.2%	85.5%	81.0%	65.0%	82.1%	50.4%	94.7%	98.6%	13.9%	77.7%	33.6%	85.4%	72.5%

Table 4: Class wise Accuracy of baseline system on development dataset

Class	beach	bus	cafe	car	city	forest	groc	home	lib	metro	offi	park	res	train	tram	overall
Accuracy	84.6%	88.5%	69.2%	96.2%	80.8%	65.4%	88.5%	92.3%	26.4%	100%	96.2%	53.8%	88.5%	30.8%	96.2%	77.2%

Table 5: Class wise Accuracy of baseline system on evaluation dataset