S-TRANSFORM AND GAUSSIAN MIXTURE MODEL FOR ACOUSTIC SCENE CLASSIFICATION

Santosh Kumar Srivastava,
Department of Computer Science and Engineering,
BRCM College of Engineering and Technology,
Haryana, India
santoshkumar.srivastava@gmail.com

Submitted: *Feb, 18, 2020* ***Revised:*** *May, 4, 2020* ***Accepted:*** *May, 19, 2020*

Abstract: In this study, Acoustic Scene Classification (ASC) system is designed with the help of *S*-transform and Gaussian Mixture Model (GMM). The *S*-transform is an extension of continuous wavelet transform that combines the progressive resolution with phase information. Thus, it exhibits the amplitude response of the frequency samples in contrast to wavelet transform. The *S*-transform coefficients are modeled by GMM using posterior probabilities of testing features. Also, preprocessing of acoustic signals is done by a series of operations; explosion, pre-emphasis filtration and windowing approach. The number of Gaussian components which is used to model the scene is varied (GMM-4, GMM-8, GMM-16, and GMM-32) and the performance of ASC system is analyzed using TAU Urban Acoustic Scenes 2019. The results show the effectiveness of the system with average recognition rate of 77.59%, 81.58%, 87.66% and 84.50% for GMM-4, GMM-8, GMM-16, and GMM-32 respectively.

Keywords: Acoustic scene classification, time-frequency representation, S-transform, probabilistic classifiers, Gaussian mixture model.

I. INTRODUCTION

The development of signal processing techniques has greatly extended to various ranges of applications such as radar and sonar related applications, audio and speech analysis, medical signal analysis. These techniques may provide more information than human imagination. The ASC system is a pattern recognition application which helps to classify the scene with the help of information in the acoustic signals. Texture descriptors based ASC system is described in [1]. At first, the acoustic signal is converted into spectrogram and then features are extracted. To improve the accuracy, visual features also combined with acoustic features.

Different neural networks and time-frequency features are discussed in [2]. These are evaluated either merged or singly. Convolution Neural Networks (CNN) and Deep Neural Networks (DNN) are used and three features such as constant Q-transform, Gammatone filter and log energy Mel filter are extracted. Feature selection methods are employed for ASC system from the aggregate of visual and acoustic features in [3]. Principal component analysis and correlation based selection approaches are used not only to select features but also to reduce feature space dimension.

A multi scale fusion based ASC system is described in [4]. A top down pathway strategy is used to integrate the multi scale semantic features obtained from a CNN with Xception architecture. A channel weighted CNN is employed for the extraction. Efficient non-negative feature learning is analyzed for ASC in [5]. It

uses feature fusion of low-level time-frequency representation by DNN and activation features non-negative matrix factorization approach.

Multi channel CNN and dense CNN are explored in [6] for ASC. Features are extracted in an end-to end manner for the detection and classification of acoustics scenes. An ASC system based on the ensembles of CNN is described in [7]. At first, the signals are converted into spectrograms and then nearest neighbour filters are applied on it to smooth similar patterns. A network ensemble model is built based on three different CNNs.

A bag-of-acoustic based system is discussed in [8] for ASC system using distributed microphone array. The extracted spatial features from each sound clip are quantized and regarded as a document for a particular acoustic scene classification. An approach for feature extraction in a constrained learning environment for ASC is described in [9]. DNN is used to stimulate the Fourier transform and their information loss is elevated by temporal transformer module. CNN and DNN are applied for ASC in [10] using log Mel and Mel frequency features. The parameters of CNN and DNN are varied and their performances are evaluated.

An efficient ASC system using *S*-transform and GMM is presented here. The S-transform coefficients are effectively extracted after preprocessing and then modeled by GMM for recognition. The organization of the work is as follows: The techniques used in ASC systems are discussed in details in section 2 and the next section discusses the results obtained by the ASC system. Section 4 concludes this work based on *S*-transform and GMM.

II. METHODS AND MATERIALS

The ASC system discussed in this section has mainly two stages as in many pattern recognition systems; feature extraction and classification. From these stages, the discriminant features are extracted and are classified into their respective scenes. Figure 1 shows a computer vision system. First, the acoustic signals are acquired and are preprocessed in the preprocessing stage in order to extract the features without redundancy. The next stages are feature extraction and classification or recognition where the patterns in the given signals are classified by the extracted features.

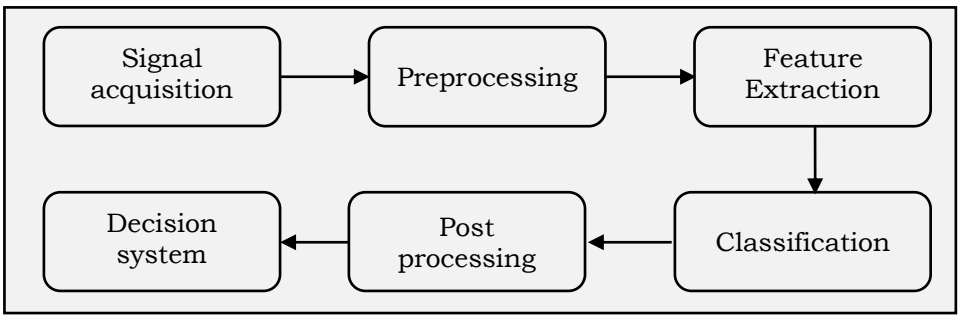

Fig. 1 Computer vision system

The ASC system based on *S*-transform and GMM classifier is considered as a multiclass problem and its definition is as follows: Let us consider *n*-acoustic scenes A*S* where the total number of different scenes is *n.* i.e., $AS = \{AS_1, AS_2, AS_3.....AS_n\}$ and the main aim is to identify the scene from *n*-scenes. To achieve this, a recognition system is designed which is defined by.

$$decision : R \mapsto \{AS_1, AS_2, AS_3 \ldots\ldots\ldots AS_n\} \quad (1)$$

where the *decision* is achieved by the GMM classifier and $\Re^t$ (t is the number of features) is the feature space created by *S*-transform. The flow of ASC system is shown in Figure 2.

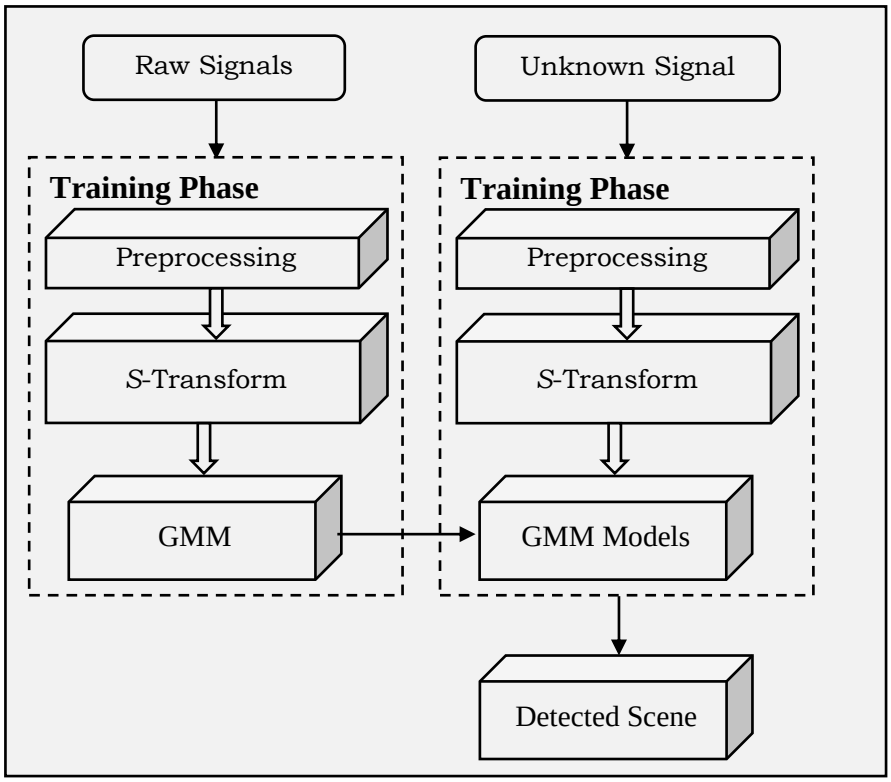

Fig. 2 ASC system using S-transform and GMM

A. Preprocessing

The overall performance of pattern recognition system can be improved by employing preprocessing steps. The ASC system consists of two preprocessing steps. At first, the original acoustic signal samples are exploded to 16 bit so that the small amplitudes are also considered for the extraction of features. Then, the ambiguity in the exploded signals is removed using pre-emphasis filter which is given below:

$$y_t = \alpha x_t + (1-\alpha) x_{t-1} \quad (2)$$

where x is referred to as the input acoustic signal and the α is referred to as the pre-emphasis filter coefficients. When applying this filter to the signal, the amplitudes of the lower frequency signals are decreased and the amplitudes of the higher frequency signals are increased. After filtering, windowing concept is applied. Let us consider the preprocessed signal *ASp*, q is the sample point of window applied, and k is the window length, then the resulting signal in a single frame is defined by,

$$AS_p(n,q) = AS_p(n)w(q-n) \tag{3}$$

where w is the windowing function which is a hamming window defined, by

$$w(n) = 0.54 - 0.46\cos\frac{2\pi(n-1)}{k-1} \qquad n = 0,1,.....k = -1 \tag{4}$$

Frame duration of 25 milliseconds with an overlapping of 10 milliseconds is analyzed for extracting frames from the filtered acoustic signal. After preprocessing, the signals are represented by *S*-transform for extracting features.

B. S-Transform

In 1996, the *S*-transform was first developed to analyze geophysical data [75] which is a time-frequency representation. The *S*-transform uniquely combines the progressive resolution that is absolutely referenced with the phase information which is main difference between other time-frequency representations such as Fourier transform. The absolutely referenced phase is explained as such the phase information given by the *S*-transform will always refer to the time t = 0, as like the Fourier transform. Due to this phase properties, *S*-transform can also be called as the locally referenced phase. The generalization of *S*-transform is as follows:

$$S(\tau, f, p) = \int_{-\infty}^{\infty} h(t)w(\tau - t, f, p)e^{-j2\pi ft}dt \tag{5}$$

where w is denoted as the *S*-transform window and p is denoted as a set of parameters determining the shape and properties of w which is given below.

$$w(t, f, p) = \frac{|f|}{\sqrt{2\pi}p}e^{-\frac{t^2f^2}{2p^2}} \tag{6}$$

The equation in (2), can also be derived from Fourier transform,

$$S(\tau, f, p) = \int_{-\infty}^{\infty} H(\alpha + f)W(\alpha, f, p)e^{j2\pi\alpha\tau}d\alpha \tag{7}$$

This *S*-transform window w has to satisfy the following four conditions as such;

$$\int_{-\infty}^{\infty} \Re\{w(\tau, f, p)\}d\tau = 1 \tag{8}$$

$$\int_{-\infty}^{\infty} \Im\{w(\tau, f, p)\}d\tau = 0 \tag{9}$$

$$w(\tau - t, f, p) = w(\tau - t, -f, p)^* \tag{10}$$

$$\frac{\partial}{\partial t}\Phi(\tau - t,\, f\,,\, p)\Big|_{t=\tau} = 0 \tag{11}$$

From the above equation, the first conditions states that when the equation is integrated over all τ, then the *S*-transform will converge to the Fourier transform as;

$$\int_{-\infty}^{\infty} S(\tau,\, f\,,\, p)d\tau = H(f) \tag{12}$$

The third condition states that the symmetrical property between the shapes of *S*-transforms analyzing function are either at the positive or negative frequencies. The important feature of the *S*-transform is that it combines the time-frequency space along with the frequency dependent resolution in reference to the information of the local spectrum phase setting. In contrast to the wavelet transform, the S-transform exhibits the amplitude response of the frequency samples. The S-transform can be derived by the Short Time Fourier Transform (STFT) transform as shown in the below equation. Let *h(t)* be the STFT of the signal s,

$$STFT(\tau,\, f) = \int_{-\infty}^{\infty} h(t)g(\tau - t)e^{-j2\pi ft}dt \tag{13}$$

where τ referred as the time spectral localization and f be referred as the Fourier frequency of the input signal respectively. And *g(t)* be referred as the windowing function. The *S*-transform can be derived by identifying the windowing function *g(t)*, with respect to the Gaussian function as

$$g(t) = \frac{|f|}{\sqrt{2\pi}} e^{-\frac{t^2 f^2}{2}} \tag{14}$$

By applying the above Gaussian function, the S-transforms can be defined as

$$S(\tau, f) = STFT(\tau, f) = \int_{-\infty}^{\infty} h(t)\frac{|f|}{\sqrt{2\pi}} e^{-\frac{t^2 f^2}{2}} e^{-j2\pi ft}dt \tag{15}$$

It is noted that if the window of *S*-transform is increased in time-domain, then the transform can provide better frequency resolution even for the low- frequency signals. It is considered that the *S*-transform have the complete information of the referenced phase signals.

C. GMM Classification

GMM classifier classifies the given classes of data by computing the posterior probability using the testing features with training database. In general, a class is described by $G_m = \{\gamma_1, \gamma_2, \gamma_3, \ldots\ldots \gamma_M\}$ with *M* Gaussian models. Expectation and Maximization (EM) is employed to compute the *M* Gaussian models and its relative weights [82]. The conditional probability is given by

$$p(T \mid \nabla) = \sum_{i=1}^{M} c_i \cdot \gamma_i(T) \tag{16}$$

where $\gamma_i(T)$ and c_i are the *N*-variate Gaussian function and mixture weights respectively. The best fit class is computed using Bayes rule and EM algorithm for testing features by finding the posterior probability [82].

III. RESULTS AND DISCUSSION

The TAU Urban Acoustic Scenes 2019 [13] database is used for performance evaluation of ASC system using *S*-transform and GMM classifier based system. Only the subtask A of task 1 (ASC) is analyzed. It consists of 10 different acoustic signals acquired from 12 large cities of Europe. They are Amsterdam, Vienna, Stockholm, Prague, Paris, Milan, Madrid, Lyon, London, Lisbon, Helsinki and Barcelona. The whole database is split into two sets; development set (acoustic signals from 10 cities) and evaluation set (acoustic signals from 12 cities). All signals are acquired using the same device. Table 1 shows the different scene class in the development dataset.

TABLE 1 Scene classes in database

#class	Description	#class	Description
1	Airport	6	Tram
2	Shopping Mall	7	Bus
3	Metro Station	8	Metro
4	Public square	9	Park
5	Street traffic	10	Street pedestrian

All the scenes are classified (10-classes) using the ASC system. Table 2 shows the accuracy of ASC system with baseline model. The system is analyzed with different Gaussian components in the power of 2 i.e., GMM-4, GMM-8, GMM-16, and GMM-32. Tables 3 to 6 show the confusion matrix of the system using different Gaussian components.

TABLE 2 Accuracy of ASC system with baseline

#class	Accuracy (%)				
	Baseline	**GMM-4**	**GMM-8**	**GMM-16**	**GMM-32**
1	73.40	71.26	74.35	78.86	76.25
2	68.48	66.44	70.98	75.96	73.92
3	83.68	81.61	85.98	93.33	90.34
4	67.44	65.12	70.03	78.04	73.90
5	89.05	86.82	90.55	94.53	92.54
6	82.34	80.28	83.72	91.51	87.39
7	84.82	82.65	86.51	91.81	88.67
8	71.59	69.52	73.44	85.22	78.06
9	93.01	90.67	94.30	96.89	95.60
10	83.68	81.59	86.01	90.44	88.34
Avg.	79.75	77.59	81.58	87.66	84.50

TABLE 3 Confusion matrix of ASC system using GMM-4

		Predicted class									
		1	2	3	4	5	6	7	8	9	10
True class	1	300	72	33	1	1	1	4	1	1	7
	2	49	293	58	5	1	1	1	1	8	24
	3	8	24	355	5	1	11	7	15	3	6
	4	1	2	14	252	25	3	1	1	35	53
	5	1	1	1	18	349	1	1	2	16	12
	6	1	1	5	1	1	350	44	31	1	1
	7	1	1	2	1	1	49	343	14	2	1
	8	3	1	17	1	3	85	19	301	1	2
	9	1	1	1	21	5	1	1	1	350	4
	10	11	21	9	27	4	1	2	1	3	350

TABLE 4 Confusion matrix of ASC system using GMM-8

		Predicted class									
		1	2	3	4	5	6	7	8	9	10
True class	1	313	70	31	0	0	0	2	0	0	5
	2	46	313	55	2	0	0	0	0	5	20
	3	5	21	374	3	0	9	5	12	1	5
	4	0	0	11	271	23	0	0	0	32	50
	5	0	0	0	15	364	0	0	0	13	10
	6	0	0	3	0	0	365	40	28	0	0
	7	0	0	0	0	0	46	359	10	0	0
	8	1	0	15	0	1	82	16	318	0	0
	9	0	0	0	17	3	0	0	0	364	2
	10	8	18	7	24	2	0	0	0	1	369

TABLE 5 Confusion matrix of ASC system using GMM-16

		Predicted class									
		1	2	3	4	5	6	7	8	9	10
True class	1	332	62	25	0	0	0	0	0	0	2
	2	40	335	50	0	0	0	0	0	0	16
	3	0	15	406	0	0	4	2	7	0	1
	4	0	0	5	302	15	0	0	0	30	35
	5	0	0	0	9	380	0	0	0	8	5
	6	0	0	0	0	0	399	22	15	0	0
	7	0	0	0	0	0	32	381	2	0	0
	8	0	0	9	0	0	51	4	369	0	0
	9	0	0	0	12	0	0	0	0	374	0
	10	2	12	3	24	0	0	0	0	0	388

TABLE 6 Confusion matrix of ASC system using GMM-32

		Predicted class									
		1	2	3	4	5	6	7	8	9	10
True class	1	321	68	28	0	0	0	0	0	0	4
	2	42	326	52	0	0	0	0	0	3	18
	3	0	17	393	1	0	7	4	9	0	4
	4	0	0	9	286	19	0	0	0	31	42
	5	0	0	0	12	372	0	0	0	10	8
	6	0	0	1	0	0	381	32	22	0	0
	7	0	0	0	0	0	39	368	8	0	0
	8	0	0	12	0	0	70	12	338	0	1
	9	0	0	0	15	1	0	0	0	369	1
	10	6	14	5	24	1	0	0	0	0	379

It is inferred from the tables that the ASC system outperforms the baseline model. The ASC system with GMM-16 components for the recognition of acoustic scenes has maximum average accuracy (87.66%) and outperforms baseline model over 8% approximately. The modeling of scenes with minimum number Gaussian components has no significant performance over the baseline model.

IV. CONCLUSION

In this work, the application of machine learning algorithm in the ASC system is considered. In particular, the *S*-transform coefficients are used as descriptors and GMM is used as a recognition algorithm. Both modules are implemented in MATLAB, a numerical computing environment with vast inbuilt functions. The potential of ASC system is assesses using TAU Urban Acoustic Scenes 2019. The success of ASC system is typically dependent on the extracted features and the recognition algorithms. Different numbers of Gaussian mixtures (GMM-4, GMM-8, GMM-16, and GMM-32) are tested to achieve maximum accuracy for ASC. Results prove that GMM-16 gives maximum average accuracy of 87.66% for 10-class scene classification using *S*-transform and GMM-16 based approach.

REFERENCES

[1]. G.Z. Felipe, Y. Maldonado, G.d. Costa and L.G. Helal, "Acoustic scene classification using spectrograms," 36th International Conference of the Chilean Computer Science Society, 2017, pp. 1-7.

[2]. L.D. Pham, I.V. McLoughlin, H. Phan and R. Palaniappan, A Robust Framework for Acoustic Scene Classification. INTERSPEECH, 2019, pp. 3634-3638.

[3]. J. Xie and M. Zhu, Investigation of acoustic and visual features for acoustic scene classification. Expert Systems with Applications, Vol. 126, 2019, pp.20-29.

[4]. L. Yang, X. Chen, L. Tao and X. Gu. Multi-scale Fusion and Channel Weighted CNN for Acoustic Scene Classification. 2nd International Conference on Signal Processing and Machine Learning, 2019, pp. 41-45.

[5]. V. Bisot, R. Serizel, S. Essid and G. Richard. Nonnegative feature learning methods for acoustic scene classification, 2017.

[6]. D. Wang, L. Zhang, K. Xu and Y. Wang, Acoustic scene classification based on dense convolutional networks incorporating multi-channel features. In Journal of Physics: Conference Series, Vol. 1169, No. 1, 2019, pp. 012037.

[7]. T. Nguyen and F. Pernkopf. Acoustic scene classification using a convolutional neural network ensemble and nearest neighbor filters. In Proceedings of the Detection and Classification of Acoustic Scenes and Events, 2018, pp. 34-38.

[8]. K. Imoto and N. Ono Acoustic scene classification based on generative model of acoustic spatial words for distributed microphone array. 25th European Signal Processing Conference, 2017, pp. 2279-2283).

[9]. T. Zhang and J. Wu, J. Constrained learned feature extraction for acoustic scene classification. IEEE/ACM Transactions on Audio, Speech, and Language Processing, Vol. 27, No. 8, 2019, pp.1216-1228.

[10]. K. Hussain, M. Hussain and M.G. Khan, An improved acoustic scene classification method using convolutional neural networks (CNNs). American Scientific Research Journal for Engineering, Technology, and Sciences, Vol. 44, No. 1, 2018, pp.68-76.

[11]. R.G. Stockwell, L. Mansinha and R.P. Lowe, Localization of the complex spectrum: the S transform. IEEE transactions on signal processing, Vol. 44, No. 4,1996, pp. 998-1001.

[12]. C.M. Bishop, "Pattern recognition and machine learning", Springer, Chapter 9, Vol. 1, 2006, pp.435.

[13]. Q. Kong, T. Iqbalm, Y. Xu and M.D. Plumbley, DCASE 2018 Challenge Surrey Cross-task convolutional neural network baseline, 2018.