ENVIRONMENTAL SOUND RECOGNITION USING SPECTROGRAM IMAGE FEATURES

Amogh Hiremath[1]

[1]*Research Engineer, Research, Philips India Limited, Karnataka, India*

Abstract

Most of the prior research which has been carried out on audio recognition has been done in speech and music. Only in recent years, dozens of emerging works have been conducted on Environmental Sound Recognition and has gained importance. For the purpose of audio classification, many previous efforts utilize acoustic features such as Mel-frequency Cepstral Coefficients (MFCCs), Zero Crossing Rate (ZCR), Root Mean Square Error (RMSE), spectral centroid, spectral bandwidth and other frequency domain features derived from the spectrogram of the audio. In this paper, we use a slightly different approach of feature extraction, where we summarize short audio clips of about five seconds by segmenting out the most prominent part of the audio signal. We then compute spectrogram image of the segmented audio, and divide it into different sub-bands with respect to the frequency axis. For each of the sub-bands, we extract first order statistics and Gray Level Concurrence Matrix (GLCM) features. In the classification stage, we combine two SVM (Support Vector Machines) classifiers. The first classifier uses first order statistics and GLCM features. The second classifier uses acoustic features such as MFCCs, ZCR, RMSE, spectral centroid, spectral bandwidth and other frequency domain features derived from the spectrogram of the audio to obtain the final result. We evaluate our approach on two publicly available datasets, namely, ESC-10 and Freiburg-106 with a five-fold and a ten-fold cross validation for ESC-10 dataset and Freiburg-106 dataset respectively. Experiments show that the proposed approach outperforms the baselines and provides similar results compared to the state-of-art.

Keywords: *Environmental Sound Classification, First Order Statistics, GLCM, Spectrogram, SVM*

----------***----------

1. INTRODUCTION

Segmentation and classification of audio has gained popularity and is of major concern for many applications in recent years. The storage requirement for the audio signal is minimal and can be processed in relatively low-bandwidth. Video surveillance applications include analysis of both sound and image. Analysis of images is highly dependent on lighting conditions, however, audio analysis can be done even in darkness and in presence of fog. Most of the research on audio recognition presently has been carried out on speech and music, whereas large number of non-speech sounds are also of great importance. Non-speech audio analysis can be used for warning systems, robot navigation systems, crime investigations, and security systems. Environmental sound analyzers can be used to address noise pollution problems by keeping track of noise levels. There is a growing demand of sound classifiers with high accuracy in audio search applications. Environmental sound classification also plays a crucial role in defining the biodiversity of a region by automatic segmentation of birds and animal sounds.

Listening to everyday environmental sounds is different than human speech and music. Listening to speech and music involves taking heed to emotional content besides acoustic features. Whereas, perception of environmental sounds concerns more about events and the source that generates them. Unlike human speech where audio signals are structured, and can be broken down into elementary phonemes, non-speech audio signals are unstructured and more random. They also sometimes lack repetitive patterns like rhythm and melody as in music. Over the last decade, a number of approaches have been proposed for environmental sound classification and have become popular leading to considerable amount of research.

In most cases, the classification of audio signals is a three-step process. The first step usually involves certain preprocessing where audio denoising and silence removal is performed to segment the audio of interest. In the second step, relevant features are extracted followed by classification of the audio in the third step.

The audio signal is usually broken down into smaller segments with a particular time window and features are extracted for each segment. Mel-Frequency Cepstral Coefficients (MFCCs), Zero Crossing Rate (ZCR), Spectral Centroid are some of the most widely used features for audio analysis. Many previous efforts have also been made in classifying audio signals using other features such as MPEG-7 descriptors[1,2], Linear Prediction coefficients[3], features derived from statistics of spectrogram image of an audio [4] and Log–Gabor Filters[5]. The bag of phrases approach is introduced in [6], where a codebook is generated using Gaussian Mixture Model and then the codebook is used to obtain a new set of features for the classification.

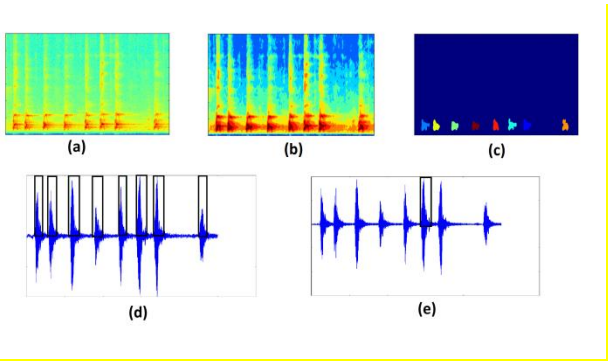

Fig-1: Sound Segmentation of dog bark. (a) Spectrogram of a dog bark; (b) k-means clustering of 1(a) spectrogram with k = 10; (c)Connected components of the labels retained from 1(b); (d) Dog bark segments using the segments obtained by applying k-means for spectrogram image; (e) Selecting the longest segment from 1(d).

In this paper, we summarize short audio clips of about five seconds by segmenting out the most prominent part of the audio signal. We model two SVM classifiers using spectrogram features and other acoustic features and, combine the two classifiers with a rule to get a resultant classification model. We evaluate our approach on two publicly available datasets by performing a 5-fold cross validation on ESC-10 dataset and a 10-fold cross validation on Freiburg-106 dataset.

The contents for the rest of the paper is organized as follows. In Section 2 we will discuss certain preprocessing steps required to segment the most prominent part of the signal. Section 3 describes the feature extraction techniques and the different classification models built using the extracted features. Results and conclusion will be portrayed in Sections 4 and 5 respectively.

2. PREPROCESSING

Accurate recognition of human speech and non-speech sounds require certain preprocessing of audio signals where the unwanted noise and silence has to be removed. Proper segmentation also reduces the computational time for further audio analysis. End point detection and silence removal are some of the well-known techniques for segmenting the speech. Lamel et al.[8] describes endpoint detector for isolated word segmentation, where a histogram of the low 10db of log energy levels are considered to estimate the background noise. Tzanetakis and Cook [9] describe a methodology for temporal segmentation using different features such as Spectral features, MFCCs, LPC coefficients and pitch. Jasmine et al. [10] model silence and noise parameters by assuming that the first 200ms of the speech signal contain noise. The assumption might be true for speech signals, but it might lead to improper modelling of noise for recordings of the environment where the audio recording might be captured starting from the middle of an event. On the other hand, Oliveira et al. [11] use morphological operations on spectrogram image and actively select the frames above a calculated threshold.

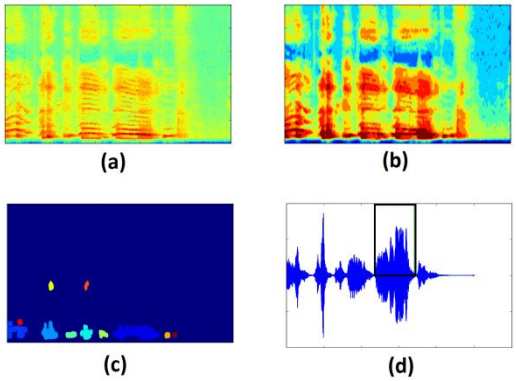

Fig-2: Sound Segmentation of baby cry; (a) Spectrogram of baby cry; (b) K-Means clustering of 2(a) spectrogram with K = 10; (c)Connected components of the labels retained from 2(b);(d)Longestsegment of baby cry audio clip.

In this work, we first resample the audio to 24,000 Hz and apply a high pass filter with a cut off frequency of 500Hz to remove the low frequency noise in the audio signals. We then compute the decibel scaled spectrogram image of the audio given by eq (1).

$$X(n,k) = \sum_{m=-\infty}^{\infty} x(m)\, w(n-m) e^{-j\frac{2\pi}{N}km} \qquad (1)$$

where, (n,k) represents the pixel co-ordinates of the spectrogram with N being the FFT (Fast Fourier Transform) size, $y_n(m) = x(m)w(n-m)$ being short-time section of signal $x(m)$ at time n, and a hanning window $w(n)$ is used as the windowing function.

The spectrogram is calculated with an FFT size of 512 which gives a frequency resolution of 46.875 Hz corresponding to the sampling rate of 24,000Hz. For the hanning window, we use a window length of 20ms with 75% overlap. Most of the previous works use the same window length and overlap as it provides a good tradeoff for the frequency, temporal resolution and the size of the image. Further, we rescale the spectrogram to a maximum value of 255. To remove the noise in the spectrogram, a median filter of radius three is applied. Figure-1(a) shows a spectrogram of a dog bark rescaled to amplitude in the range [0,255]. On the entire rescaled image, we use k-means with ten cluster centers to vector quantize the image to ten levels. Subsequently, we preform binary thresholding with the threshold being the second highest value among the cluster centers. Figure-1(b) shows spectrogram image quantized to ten levels. In this way, we create a binary mask where we retain the part of the image which correspond to two cluster centers with the highest pixel values. Figure-1(c) shows the connected components extracted from the binary mask obtained after thresholding. By considering the left most and right most location of each of the connected component, we extract corresponding segments in the time domain. Figure-1(d) shows the obtained segments in the time domain and the most prominent part of the audio clip is extracted as longest segment with respect to time as shown in Figure-1(e).

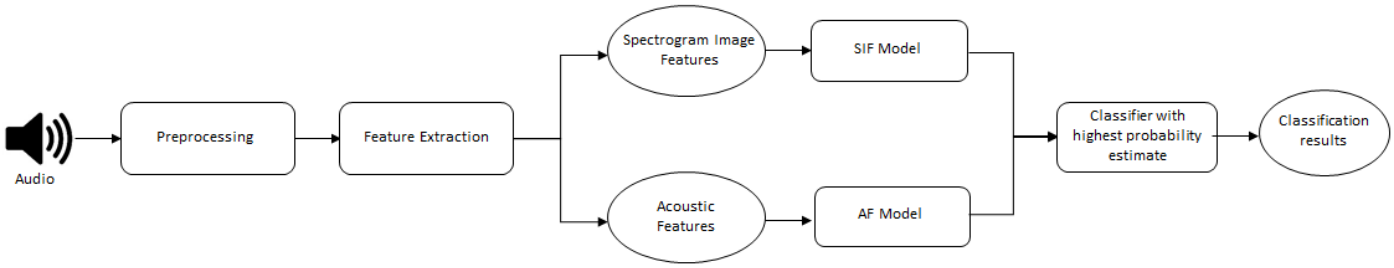

Fig-3: Architecture of the proposed MEASIF classification Model.

3. FEATURE EXTRACTION AND CLASSIFICATION

Appropriate development of feature extraction method is a tedious task, as it requires tailoring a new feature set for every new classification. For sound recognition in an audio surveillance, Sharan and Moir [12] made use of texture features of spectrogram image by concatenating the Gray Level Concurrence Matrix (GLCM) columns which they refer to as Spectrogram Image Texture Feature (SITF). Similarly, in this paper we explore environmental sound classification using two sets of features with, GLCM features and first order statistics of spectrogram image being the first set of features, and the commonly used acoustic features being the second.

In order to extract first set of features, we use the spectrogram image of the largest segment as obtained in Section 2. Wedivide the spectrogram image into four sub-bands with respect to frequency axis, and compute GLCM features and first order statistics for each of the sub-bands. To compute GLCM features, we use a combination of angles $\in \{0°, 45°, 90°, 135°\}$ and displacement vectors $\in \{3,5\}$ and extract thirteen textural descriptors out of fourteen descriptors as defined in [13]. Also, fourteen first order pixel statistics are extracted as described in pyradiomics module [14]. The GLCM descriptors include energy, contrast, correlation, sum of squares, inverse of difference moment, sum average, sum entropy, sum variance, entropy, difference variance, difference entropy, and two descriptors of information measure of correlation. The first order statistics obtained are directly applied to the sub-band of the spectrogram and it includes minimum, mean, median, variance, energy, entropy, tenth percentile pixel value, ninetieth percentile pixel value, inter quartile range, mean absolute deviation, robust mean absolute deviation, root mean square error, skewness and kurtosis. Details about each of the first order statistics features and their implementation is well stated in [14]. The final feature vector is concatenation of first order pixel statistics and GLCM features of all the sub-bands.

For second set of features, we extract commonly used acoustic features such as MFCCs, Delta MFCCs, ZCR, RMSE, spectral centroid, spectral bandwidth, spectral contrast and spectral rolloff with a window length of 20ms and 75% overlapping between the frames. The average and variance of these features across the frames constitute for acoustic feature vectors.

For the purpose of classification, we use four different models namely, **SIF Model** (Spectrogram Image Feature Model), **AF Model** (Acoustic Features Model), **ASIF Model** (Acoustic and Spectrogram Image Features Model) and a **MEASIF** (Modified Ensemble of Acoustic and Spectrogram Image Features Model). The SIF Model and AF Model are trained separately with the acoustic features and spectrogram image features respectively. A feature set containing a combination of both acoustic features and the spectrogram image features are used to train ASIF Model. Figure-3 shows the architecture of MEASIF Modelwhere we combine SIF Model and AF Model by a modified ensemble of both the models. The modified ensemble involves choosing the best model among the two by considering the class with the maximum probability estimate or the confidence score attained for each of the prediction.

4. EXPERIMENTS AND ANALYSIS

We evaluate our approach on two publicly available datasets, namely, ESC-10 and Freiburg-106. Validation is performed using five-fold cross validation on ESC-10 [7] dataset and a ten-fold cross validation on Freiburg-106 dataset. We compare our approach using four different models as described in Section 3. All the models are trained on SVM classifier with “rbf” kernel and the cost parameter “C” set to 1e4.

4.1 Datasets

The ESC–10 dataset consists of 400 labeled environment recordings equally divided in to 10 categories. Each audio clips are five seconds long, sampled at 44.1 KHz and compressed with Ogg vorbis compression at 192 kbit/s. The human accuracy for ESC-10 is 95.7%. The Frieburg-106 dataset [15] was collected using a consumer level dynamic cardioid microphone. It contains 1,479 audio based human activities of 22 categories.

4.2 Experimental Results

To evaluate and compare the recognition results, we use f-score [16] as the metric where both the precision and recall are taken into account as shown in eq (2).

$$f\text{-}score = 2.\left(\frac{precision\ .\ recall}{precision\ +recall}\right) \quad (2)$$

Table-1 shows the average f-score obtained for ESC – 10 dataset with four different models. The AF Model, SIF Model and ASIF Model gives an average f-score of 83.5%, 83% and 87% respectively. Whereas, the MEASIF Model which is a combination of AF and SIF Model, outperforms with an average f-score of 87.75%. Figure-4 shows the confusion matrix obtained by MEASIF Model and it can be seen that the model performs better on clock tick and chain saw events. However, least performance is observed in classification of helicopter sounds. This could be due to its close resemblance with other long duration ambient sounds such as rain and sea waves. Our approach outperforms the baseline recognition rates [7] of about 67.5% using SVM and about 72.7% using random forest classifier. By combining two models trained on self-engineered features such as acoustic features and spectrogram features, we are able to achieve recognitions rates compared to the state-of-art results [16] in which a recognition rate of 89.9% was attained using convolutional neural networks. It can also be observed that our approach has outperformed in inter-class recognition rates for events such as clock tick, fire crackling, and person sneeze compared to that of the state-of-art.

The recognition rates for Freiburg-106 dataset are shown in Table-2, for which a 10-fold cross validation is performed. Similar to ESC-10 dataset, MEASIF outperforms with an average f-score measure of 97.62% compared to the baseline recognition rate with an average f-score of 92% [17] using Non-Markovian Ensemble Voting. The recognition rate attained by our approach is on par with the state-of-art results where L.Hertel et al.[16] make use of convolutional network and obtain an average f-score of 98.3%.

Table -1: Recognition rates for ESC-10 dataset using f-score in percentage

No.	Event	AF Model	SIF Model	ASIF Model	MEASIF Model
1	Dog Bark	92.5	86.75	95	90.91
2	Rain	79.52	81.48	83.33	85.37
3	Sea Waves	86.75	77.5	80.95	86.75
4	Baby Cry	84.93	85	89.74	89.47
5	Clock Tick	75.95	86.42	87.80	87.80
6	Person Sneeze	84.71	84.34	90.47	88.37
7	Helicopter	78.48	75.32	80	79.48
8	Chainsaw	79.49	89.16	88.60	90
9	Rooster	90	82.05	92.5	92.68
10	Fire Crackling	82.5	81.08	81.08	86.48
	Average	83.5	83	87	87.75

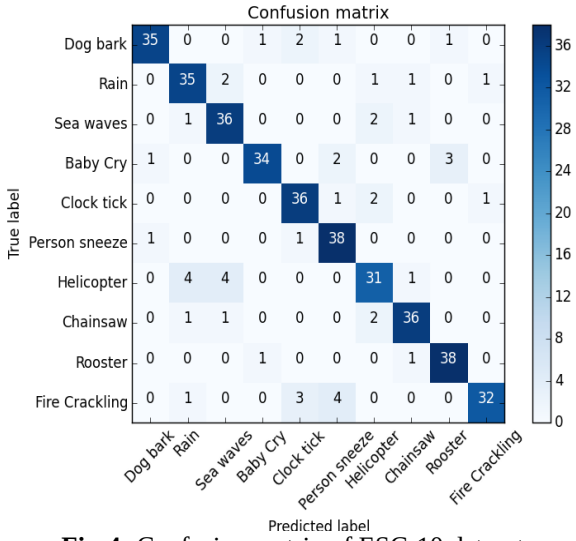

Fig-4: Confusion matrix of ESC-10 dataset

Table -2: Recognition rates for Freiburg dataset using f-score in percentage

No.	Event	AF Model	SIF Model	ASIF Model	MEASIF Model
1	Background	74.6	74.9	97.5	86.04
2	Food Bag Opening	97.4	98.2	96.34	98.15
3	Blender	100	96.7	100	100
4	Cornflakes Bowl	93	92.2	88.88	94.28
5	Cornflakes Eating	96.4	91.3	95.23	95.23
6	Pouring cup	91.7	90	90.47	93.33
7	Dish Washer	100	100	98.88	100
8	Electric Razor	100	98.7	100	100
9	Flatware Sorting	92.2	84.1	94.59	88.60
10	Food Processor	100	97.5	94.44	100
11	Hair Dyer	100	100	100	100
12	Microwave	98.5	99	100	98.92
13	Microwave Bell	95	90.7	100	95.83
14	Microwave Door	96.9	97.9	96.62	97.72
15	Plates Sorting	95	96.2	94.32	97.08
16	Stirring Cup	98.4	95.3	96.72	99.15
17	Toilet Flush	96.8	88.5	93.75	96.82
18	Tooth Brushing	98	98	100	98.24
19	Vacuum Cleaner	96.7	95.3	98.68	98.06

20	Washing Machine	97.9	94.3	98.50	99.24
21	Water Boiler	99.2	97.8	96.96	100
22	Water Tap	97.5	97.4	99.11	97.81
	Average	96.7	94.9	97.08	97.62

5. CONCLUSION

In this paper, a method for environmental sound recognition is proposed. Experimental results show that high sound recognition rates can be achieved by considering only a small segment of the audio instead of the whole audio signal. We can also infer that, the MEASIF model, which is combination of two different SVM classifiers trained with acoustic features and spectrogram image features performs better than a ASIF Model, trained with all the features together. Our proposed approach outperforms the baseline methods and achieves recognitions rates proportionate to that of state-of-art results.

REFERENCES

[1] G. Muhammad, K. Alghathbar, "Environment Recognition from Audio Using MPEG-7 Features," in Proc. 4th Int. Conf. Embedded and Multimedia Computing, Jeju, Korea, 2009, pp. 1-6.

[2] X. Valero, F. Alías, "Applicability of MPEG-7 low level descriptors to environmental sound source recognition," in European Acoustics Association Euroregio, Ljubljana, Slovenia, 2010.

[3] Khan MKS, Khatib WGA. (2006, Aug.). Machine-learning Based Classification of Speech and Music. Multimedia Syst.2006, 12(1), pp. 55–67.

[4] J. Dennis, H. D. Tran, and H. Li. (2011, Feb.). Spectrogram image feature for sound event classification in mismatched conditions. IEEE Signal Processing Letters, 18(2), pp. 130-133.

[5] S. Souli, Z. Lachiri, "Environmental sound classification using log-Gabor filter," In 11th Int. Conf. of Signal Processing, Beijing, China, 2012, pp. 144-147.

[6] H. Phan, L. Hertel, M. Maass, R. Mazur, A. Mertins, "Audio phrases for audio event recognition," in European Signal Process. Conf., Nice, France, 2015.

[7] K. J. Piczak, "ESC: Dataset for environmental sound classification", in Proc. 23rd ACM Int. Conf. Multimedia, Brisbane, Australia, 2015, pp. 1015-1018.

[8] L. F. Lamel, L. R. Rabiner, A. E. Rosenberg, and J. G. Wilpon. (1981, Aug.). "An improved endpoint detector for isolated word recognition," IEEE Trans. Acoustics Speech Sig. Process, 29(4), pp. 777-785.

[9] G. Tzanetakis, P. Cook, "A Framework for Audio Analysis Based on Classification and Temporal Segmentation", in Proc. EUROMICRO Informatics Theory and Practice for the New Millennium, Milan, Italy, 1999, pp. 2061- 2067.

[10] J. Meribah Jasmine, S. Sandhya, Dr. Ravichandran, Dr. D. Balasubramaniam.(2006, Apr.) "Silence Removal from Audio Signal Using Framing and Windowing Method and Analyze Various Parameter", IJIRCCE, 4(4).

[11] A. G. de Oliveira et al.(2015, Nov.). "Bird acoustic activity detection based on morphological filtering of the spectrogram", Applied Acoustics, 98(1), pp. 34-42.

[12] R. V. Sharan, T. J. Moir, "Robust audio surveillance using spectrogram image texture feature", Proc. IEEE Int. Conf. Acoust. Speech Signal Process., Brisbane, Australia, 2015, pp. 1956-1960.

[13] R. M. Haralick, K. Shanmugam, and I. Dinstein. (1973, Nov.). "Textural features for image classification," IEEE Transactions on Systems, Man and Cybernetics, 3(6), pp. 610- 621.

[14] Joost JM van Griethuysen, Andriy Fedorov, Chintan Parmar, Ahmed Hosny, Nicole Aucoin, Vivek Narayan, Regina GH Beets-Tan, Jean-Christophe Fillion-Robin, Steve Pieper, Hugo JWL Aerts, "Computational Radiomics System to Decode the Radiographic Phenotype"; Submitted 2017

[15] J. A. Stork, L. Spinello, J. Silva, and K. O. Arras, "Audiobased human activity recognition using on-Markovian ensemble voting," in Proc. RO-MAN, Paris, France, 2012, pp. 509–514.

[16] L. Hertel, H. Phan, and A. Mertins, "Comparing time and frequency domain for audio event recognition using deep learning," in Int. Joint Conf. Neural Networks IJCNN, Vancouver, Canada, 2016.

[17] J. Stork, L. Spinello, J. Silva, and K. O. Arras, "Audio-based human activity recognition using non-markovian ensemble voting," in Proc. RO-MAN, Paris, France, 2012, pp. 509-514.

BIOGRAPHIES

Amogh Hiremath is a Research Engineer in Philips Innovation Campus, Bangalore. He has obtained his M.Tech in Communication Engineering from NITK, Surathkal and B.E in Electronics and Communication Engineering from SDMCET, Dharwad