Automatic Audio Event Recognition Schemes for Context-Aware Audio Computing Devices

Shivam Soni, Sudipta Dey and M. Sabarimalai Manikandan, *Member, IEEE*
Real-time Embedded Signal Processing Lab, School of Electrical Sciences
Indian Institute of Technology Bhubaneswar, Jatani, Odisha-752050, INDIA.
E-mail: ss81@iitbbs.ac.in, sd27@iitbbs.ac.in and msm@iitbbs.ac.in

***Abstract*—Automatic audio event recognition (AER) plays a major role in designing and building intelligent location and context-aware applications including audio surveillance, audio indexing and content retrieval, highlight extraction, drone and robotic navigation, machine health monitoring, audio-aware voice processing services, and urban sound pollution monitoring. In this paper, we present audio event recognition (AER) schemes using the Mel-frequency cepstral coefficients (MFCC) and machine classifiers such as multi-class support vectors machines (MC-SVM), fully connected feed-forward neural networks (FCFFNNs), and one-dimensional convolutional neural networks (1D-CNNs) that are capable of automatically recognizing seven sound classes including *aircraft, construction, music, nature (wind and rain), speech, vehicle, and train*. In this study, we created large scale audio database for both training and testing purposes. The performance of the three AER schemes are evaluated under different audio frame sizes (100 ms, 250 ms and 500 ms) using a wide variety of sounds recorded using different kinds of recording devices. Results show that the FCFFNN and 1D-CNN based AER schemes had the F1-score values of 95.72% and 96.34% for audio frame size of 250 ms whereas MC-SVM based AER scheme had the F1-score value of 85.84%. The 1D-CNN based AER scheme had a class-wise accuracy is greater than 84% for audio frame size of 250 ms whereas the FC-FFNN based scheme had a class-wise accuracy is greater than 80% for audio frame size of 250 ms. The computational analysis results show that the prediction time of 1D-CNN based scheme is faster than the FC-FFNN based AER scheme.**

***Index Terms*—Environmental sound classification, Audio scene recognition, Support vectors machines, feed-forward neural networks, and 1D convolutional neural networks.**

I. Introduction

Sounds convey lots of information about our everyday environments that can be used for extracting interesting portion of the events present in the recorded audio and video and predicting the faults different kinds of machines that are continuously used in many practical applications [1]- [13]. Further, the sounds can be used for monitoring some of suspicious activities in the unauthorized zones. The issue of recognizing risky occasions on streets or roads was considered by designing an audio surveillance framework that automatically identifies dangerous circumstances, for example, vehicle crashes and tire slipping [2]. Audio event detection (AED) is defined as the detection of individual sound events that are available inside an audio, resulting in a symbolic description such that each annotation gives the start time, end time and label for a single instance of a specific sound event [1]. Audio scene classification (ASC) is to classify a test recording into one of predefined sound scenes that characterizes the everyday environment in which a recording has been made, on the assumption that an audio scene is distinguishable from others based on its general acoustic properties [7].

Recently, ASC is a prominent research topic which is considered as a machine-learning task within a widespread single-class classification paradigm, wherein a set of class labels is known. In [1], E. Cakir et al. presented a convoluational recurrent neural network (CRNN) based polyphonic sound recognition scheme with MFCC feature. Results showed that the RNN based scheme considerably improves the recognition accuracy as compared to the feed forward neural networks (FFNNs), CNN and RNN based schemes on TUT database. In [3], J. Wang et al., presented environment sound classification for home automation using Gabor dictionary with matching pursuit (MP), principal component analysis (PCA) and linear discriminate analysis and multi-class support vector machine (MC-SVM) classifier for 17 class sounds recognition. In [4], S. Souli et al., proposed audio sounds classification using scattering features and support vectors machines for medical surveillance applications. In [5], W. Yang et al. showed that spectral MFCC features along with temporal dynamic feature can improve the performance for sound classification with local binary pattern. The method is evaluated on the TUT database (15 class) with proposed features with an ensemble classifier. In [6], S. Sigtia et al. presented a comparative study using various classifiers (DNN, GMM, SVM, RNN) in the context of IoT sound sensing application. The schemes were evaluated on the DCASE-16 database with 13 class of sounds. Results showed that the DNN yields the best classification accuracy than the GMM and SVM based schemes for a range of computational costs. In [9], O. Gencoglue et al. proposed the deep neural network (DNN) based recognition of isolated acoustic events such as footsteps, baby crying, motorcycle, rain etc. The scheme was evaluated for 61 sound classes by using the Mel energy feature vector. Results demonstrated that the DNN based audio classifier performs better than the HMM with GMM classifiers. In [10], H. D. Tran et al., proposed an online sound recognition based on probabilistic distance SVM with subband temporal envelop feature. The method was evaluated using the 10 class sound database. Results showed that the method outperforms than the MFCC features with SVM classifier. In [11], X. Valero et al. presented the biologically in-

ISBN: 978-1-7281-3296-9/19/$31.00 ©2019 IEEE

spired Gammatone cepstral coefficients (GTCCs) feature based non speech recognition scheme with SVM classifier. Results showed that the GTCCs feature results in better classification accuracy than that of the MFCC based scheme. In [12], F. Beritelli et al. presented environmental sound classification using the MFCC features and feed forward neural network classifier. The scheme was evaluated on 10 sound classes by varying window size. In [13], Rabaoui et al., proposed one class SVM based environmental sound classification for audio surveillance of 9 sound classes using the audio features such as d the ZCR, spectral centroid, spectral bandwidth, MFCCs extracted in various domains including temporal, frequency, cepstral, and wavelet. Results showed that it outperforms the binary class SVM and HMM based schemes.

In this paper, we present audio event recognition (AER) schemes using the mel-frequency cepstral coefficients (MFCC) and machine classifiers such as multi-class support vectors machines (MC-SVM), fully connected feed-forward neural networks (FCFFNNs), and one-dimensional convolutional neural networks (1D-CNNs) that are capable of automatically recognizing seven sound classes including aircraft, construction, music, nature (wind and rain), speech, vehicle, and train sounds. Major contributions of this paper are as follows.

- Exploring an effective deep learning networks based automatic audio event recognition scheme for audio surveillance application.
- Identifying the optimal audio frame size for yielding better recognition accuracy with considerable computational load under real-time implementation of audio surveillance system.
- Creating the large scale audio databases including aircraft, construction, music, nature (wind and rain), speech, vehicle, and train sounds by considering different audio recording devices such as commercial hand-held audio recorders, Smartphones, laptop PC and different recording environments.
- Developing cloud-based audio event recognition (AER) framework for remote audio sensing and monitoring services.
- Studying the real-time feasibility of the AER framework on Raspberry Pi computing platform.

The rest of this paper is organized as follows. Section II presents audio event recognition (AER) schemes using the MFCC features and machine learning classifiers. Section III presents the evaluation results of the three AER schemes on large scale audio databases. Finally, conclusions and future directions are drawn in Section IV.

II. Audio Event Recognition Schemes

In this section we describe the three audio event recognition schemes developed using the the mel-frequency cepstral coefficients (MFCC) and three machine classifiers such as multi-class support vectors machines (MC-SVM), fully connected feed-forward neural networks (FCFFNNs), and one-dimensional convolutional neural networks (1D-CNNs). In this study, our objective is to develop effective and efficient audio event recognition scheme for automatically recognizing 7 sound classes including *aircraft, construction, music, nature (wind and rain), speech, vehicle, and train* by choosing the suitable audio frame size. Since each of the sources can produce sounds with different durations which need to be considered in the creation of sound models for achieving better recognition accuracy with processing time. In some of the sound recognition methods like speech processing, selection of optimal speech frame size was well studied based on the glottal periods for processing and analysis of speech signals. But it is very difficult to have a complete study of the durations of each of the sounds that can be produced by different sources in real-life. Thus, we study the performance of the AER schemes under audio frame sizes of 100 ms, 250 ms and 500 ms.

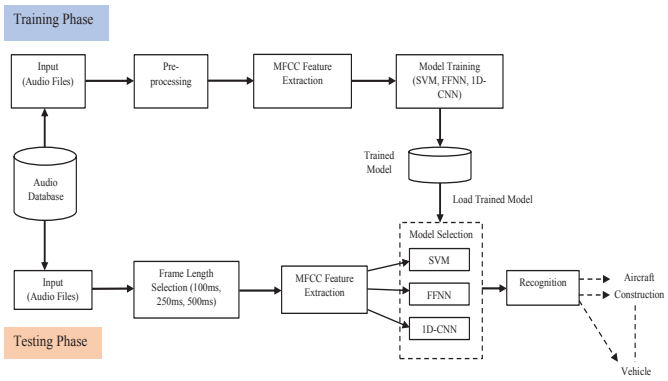

Fig. 1. Block diagram of the audio event recognition schemes using the MFCC features and machine learning classifiers.

The block diagram of the machine learning (ML) classifier based audio event recognition scheme is shown in Fig. 1, which consists of four major steps: preprocessing, feature extraction, sound models and recognition. Each of the audio event recognition steps is described in the next subsection.

A. Preprocessing

Since an acoustic sensor is exposed to severe environmental sound sources like wind, the recorded audio signals may be corrupted with the low-frequency components having frequency below 10 Hz which are produced by the senors' movements due to the winds. Since the preservation of these low-frequency components may not be interested for our seven sound classification, the recorded audio signal is passed through high-pass filter with cut-off frequency of 10 Hz. Then, the recorded audio signal is divided into frames with a predefined frame size (100/250/500 ms). In practice, sound intensity (or amplitude) depends on the location of an acoustic sensor from a sound source location which can be time-varying in the audio surveillance zone. Also the source location cannot be known aperiori in realistic scenarios. Some times the low sound intensity can be recorded when an acoustic sensor is far away from a sound source although the sensitivity of a sensor is very good. In order to avoid the microphone sensitivity variation, the amplitude normalization is performed on the zero-mean audio signal.

ISBN: 978-1-7281-3296-9/19/$31.00 ©2019 IEEE

B. MFCC Feature Extraction

Since our objective to investigate the performance of three ML classifiers for different frame size, we used the standard mel-frequency cepstral coefficients (MFCC) feature which is most commonly used in many environmental, speech and music sounds classification because human perception sensitivity with respect to frequencies is considered [12]. Each step of the MFCC feature extraction is described below:

- **Pre-emphasis filter** is used to amplify the high frequencies that balances the spectrum since high frequencies usually have smaller amplitudes as compared to the lower frequencies and may improves the signal-to-noise ratio (SNR). The pre-emphasis filter is implemented as [12]:

$$y[n] = x[n] - \alpha x[n-1], \tag{1}$$

  where α is fixed as 0.97.
- After pre-emphasis, window function such as the Hamming window is applied to each audio frame. The Hamming window function is defined as [12] :

$$h(n) = 0.54 - 0.46 \ \ cos\left(\frac{2\pi n}{N-1}\right) \tag{2}$$

  where, $0 \leq n \leq N-1$, where N represents the length of audio frame. The window function $h[n]$ is multiplied with the filtered signal $y[n]$ and result in $z[n]$.
- **Fourier Spectrum**: The magnitude spectrum is computed by taking the fast Fourier transform (FFT) of the windowed audio frame $z[n]$ that corresponds to different energy distribution over frequencies.
- **Mel-Frequency Spectrum**: The magnitude spectrum is multiplied by a set of 26 triangular band-pass filters which have the positions equally spaced along the Mel-scale which is related to the linear frequency f by the following equation [4]:

$$M(f) = 1125 \ \ \ln\left(1 + \frac{f}{700}\right). \tag{3}$$

  The Mel-frequency is proportional to the logarithm of the linear frequency that reflects the similar audio effects in the human's subjective perception.
- **MFCC**: Since the filter bank coefficients are highly correlated the discrete cosine transform (DCT) is applied to decorrelate the filter bank coefficients. The coefficients are computed as [1]:

$$C_n = \sum_{k=1}^{K} (\log \ \ S_k) \left[n\left(k - \frac{1}{2}\right)\frac{\pi}{K}\right] \tag{4}$$

  where, $n = 1, 2....., K$ and $K = 26$, number of triangular band-pass filters, S_k energy output of the K^{th} triangular band-pass filter.

Thus, above feature extraction step results in 26 coefficients. In this study, we use only the lower 13 of the 26 coefficients for each frame. The parameters used for extracting the MFCC are: pre-emphasis value of 0.97, 26 filters, lower band edge of 0 Hz, higher band edge of 8000 Hz, and 13 cepstral coefficients for each audio frame.

TABLE I
SPECIFICATIONS OF MULTI-CLASS SVM BASED AER SCHEME

Parameters	Frame Length		
	100 ms	**250 ms**	**500 ms**
Audio format	wav	wav	wav
Channel	Mono	Mono	Mono
Bit depth	16 bit PCM	16 bit PCM	16 bit PCM
Class	7	7	7
Events	28	28	28
Training data duration	28 h	28 h	28 h
Testing data duration	70 min	70 min	70 min
Sampling rate	16 kHz	16 kHz	16 kHz
Feature	MFCC	MFCC	MFCC
Number of samples in a frame	1600	4000	8000
NFFT	2048	4096	8192
Kernel	RBF	RBF	RBF
Penalty parameter C	1	1	1
Gamma	0.0769	0.0769	0.0769
Degree	3	3	3
Fold	5	5	5
Cache size	200	200	200
Tolerance	0.001	0.001	0.001
Epochs	Infinite	Infinite	Infinite
Training time	196 h	29 h	10 h
Trained model size	60.1 MB	25.3 MB	14.3 MB

C. Machine Learning Classifiers

In this study, we evaluate the performance of the three machine learning classifiers such as multi-class support vectors machines (MC-SVM), fully connected feed-forward neural networks (FCFFNNs), and one-dimensional convolutional neural networks (1D-CNNs) under different audio frame size of 100, 250 and 500 ms for automatically recognizing the 7 sound classes.

1) Multi-Class SVM: For a given training data (x_i, y_i) for $i = 1, 2, ...N$, the optimization problem for the SVM is formulated as follows [6]:

$$\underset{w,\xi,b}{\arg\min}\{\frac{1}{2}\|\mathrm{w}\|^2 + \mathrm{C}\sum_{\mathrm{i}=1}^{\mathrm{N}}\xi_{\mathrm{i}}\} \tag{5}$$

subject to the constraints

$$y_i(w \cdot x_i - b) \geq 1 - \xi_i, \qquad \xi_i \geq 0 \tag{6}$$

where ξ_i are non-negative slack variables. It is solved using a Lagrangian formulation of the problem, thus producing the multipliers α_i and the decision function [6]:

$$f(x) = sgn\left(\sum_{i=0}^{N} y_i \alpha_i x \cdot x_i + b\right) \tag{7}$$

where N is the number of training samples and x is a feature vector. A nonlinear kernel function $\mathrm{K}(x_i, x_j)$ is used to replace the dot products $x \cdot x_i$, with the effect of projecting the data into a higher dimensional space where it is linearly separable. The decision function is defined as [4]:

$$f(x) = sgn\left(\sum_{i=0}^{N-1} y_i \alpha_i K(x, x_i) + b\right) \tag{8}$$

ISBN: 978-1-7281-3296-9/19/$31.00 ©2019 IEEE

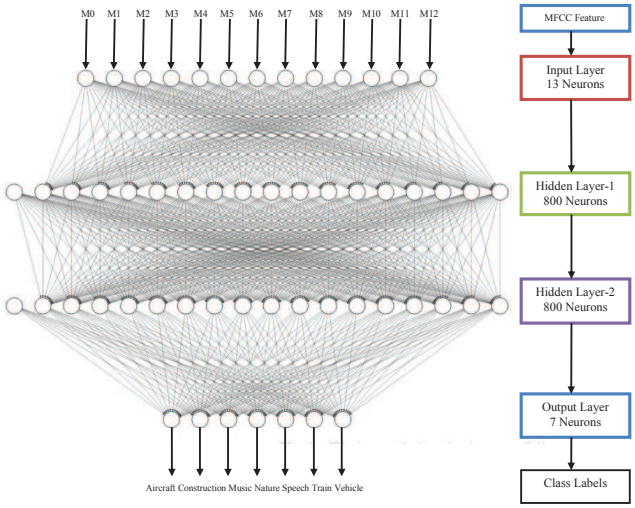

Fig. 2. Block diagram of the fully connected FFNN based AER scheme.

In this work, the Gaussian radial basis function is used. It is defined as $K_R(x, x_i) = exp(-\gamma\|x - x_i\|^2)$, and the SVM classifier was implemented using Scikit-learn, which is one-against-rest multi-class classifier [10]. The specifications of the MC-SVM based AER scheme are summarized in Table I.

2) Fully Connected FFNN Based AER Scheme: The FFNN architecture contains the input layer, hidden layer and output layer. Each layer gets its input from a previous layer and then computes and transforms the data. The transformed data sends to the next layer. Each layer consists of neurons which are having various modes of connections to other neurons in the same layer as well as the other layers depending on the type of network. Each layer connection has its own weights. The block diagram of feed forward neural network is shown in Figure 2. The non-linear transformation is defined as [6]:

$$h^i = f(W^i h^{i-1} + b^i), \quad for \;\; 0 \leq i \leq L \qquad (9)$$

where h^0 corresponds to the input x, W_i, b_i are the weight matrix and bias vector for the i^{th} layer, and the output of the final layer h^L is the desired output. The non-linearity f is usually a *sigmoid* or a *hyperbolic tangent function, tanh(x)*. However, the Rectified Linear Unit (ReLU) function is used for faster convergence of models in the training phase. The *ReLU* function is defined as $f(x) = max(0, x)$, which is simpler than the *sigmoid or tanh activation* because it only requires a piece-wise linear operator instead of a fixed point *sigmoid or tanh* [9]. The specifications of the FC FFNN based AER scheme are summarized in Table II.

3) One-dimensional CNN Based AER Scheme: Fig. 3 shows a block diagram of the one-dimensional CNN based AER scheme. The details of the ID-CNN based AER scheme are summarized in Table III. In this study, the number of layers and the tuning parameters are varied for the optimum performance. Hence, the proposed model consists of 4 convolutions, 1 max-pooling, 2 dropout, 1 flatten and 2 fully-connected

TABLE II
SPECIFICATIONS OF THE FC FFNN BASED AER SCHEME

Parameters	Frame Length		
	100 ms	**250 ms**	**500 ms**
Audio format	wav	wav	wav
Channel	Mono	Mono	Mono
Bit depth	16 bit PCM	16 bit PCM	16 bit PCM
Class	7	7	7
Events	28	28	28
Training data duration	70 h	70 h	70 h
Testing data duration	70 min	70 min	70 min
Sampling rate	16 kHz	16 kHz	16 kHz
Feature	MFCC	MFCC	MFCC
Number of samples	1600	4000	8000
NFFT	2048	4096	8192
No. of hidden layer	2	2	2
Fold	5	5	5
Activation functions	Softmax, Relu	Softmax, Relu	Softmax, Relu
Number of neurons	1600	1600	1600
Batch size	128	128	128
Learning rate	0.001	0.001	0.001
Optimizer	Adam	Adam	Adam
Epochs	15	15	15
No. of parameters	657607	657607	657607
Training time	277 min	72 min	57 min
Trained model size	7.92 MB	7.92 MB	7.92 MB

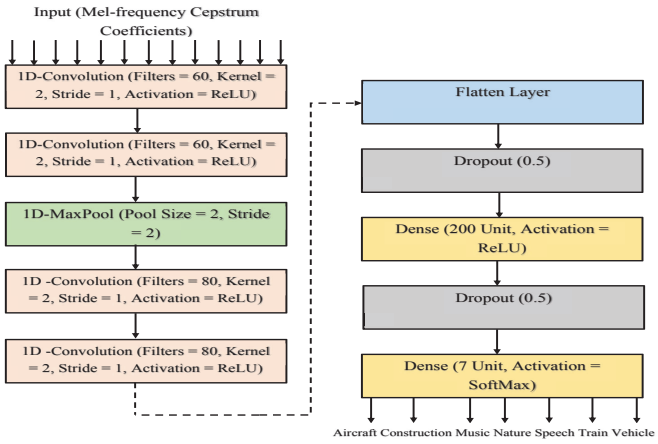

Fig. 3. One-dimensional CNN based audio event recognition (AER) scheme

layers. The filter shifts are set at 1 and 2 for convolution and max-pooling, respectively. The max-pooling operation reduces the dimensions of feature maps and also retains the important and significant features of the audio feature [8]. The flatten layers convert 2D output into 1D output. To avoid the the model overfitting, a dropout layer is used before each dense layer. In the last stage, the fully-connected layer is used to connect the neurons in the previous layers. The specifications of the 1D-CNN based AER scheme are summarized in Table IV.

III. RESULTS AND DISCUSSION

In this section, we evaluate the performance of the three machine learning classifiers based AER schemes using a wide variety of audio recorded using different kinds of audio recording devices.

ISBN: 978-1-7281-3296-9/19/$31.00 ©2019 IEEE

TABLE III
SPECIFICATIONS OF 1D-CNN BASED AER SCHEME FOR DIFFERENT FRAME LENGTHS

No.	Layer (Type)	Frame Length					
		100 ms		250 ms		500 ms	
		Shape	Parameters	Shape	Parameters	Shape	Parameters
1	Conv1D	(12 , 60)	180	(12 ,60)	180	(12 , 60)	180
2	Conv1D	(11, 60)	7260	(11 , 60)	7260	(11 , 60)	7260
3	MaxPooling	(5 , 60)	0	(5 , 60)	0	(5 , 60)	0
4	Conv1D	(4 , 80)	9680	(4 , 80)	9680	(4 , 80)	9680
5	Conv1D	(3 , 80)	12880	(3 , 80)	12880	(3, 80)	12880
6	Flatten	240	0	240	0	240	0
7	Dropout	240	0	240	0	240	0
8	Dense	200	48200	200	48200	200	48200
9	Dropout	200	0	200	0	200	0
10	Dense	7	1407	7	1407	7	1407

TABLE IV
THE SUMMARY OF DATABASE AND HYPER-PARAMETERS OF 1D-CONVOLUTIONAL NEURAL NETWORK MODEL USED IN THIS WORK

Parameters	Frame Length		
	100 ms	250 ms	500 ms
Audio format	wav	wav	wav
Channel	Mono	Mono	Mono
Bit depth	16 bit PCM	16 bit PCM	16 bit PCM
Class	7	7	7
Events	28	28	28
Training data duration	70 h	70 h	70 h
Testing data duration	70 min	70 min	70 min
Sampling rate	16 kHz	16 kHz	16 kHz
Feature	MFCC	MFCC	MFCC
Number of samples	1600	4000	8000
NFFT	2048	4096	8192
Convolution layer	4	4	4
Dropout layer	2	2	2
MaxPooling layer	1	1	1
Fold	5	5	5
Activation functions	Softmax, Relu	Softmax, Relu	Softmax, Relu
Number of neurons	1600	1600	1600
Batch size	128	128	128
Learning rate	0.001	0.001	0.001
Optimizer	Adam	Adam	Adam
Epochs	20	20	20
No. of parameters	79607	79607	79607
Training time	13 hours	2 hour	1 hour
Trained model size	1 MB	1 MB	1 MB

TABLE V
DETAILS OF AUDIO DATABASE

Title	Class	Duration (h)	Contribution	
			Self	Internet
Aircraft	Aircraft	13		100%(Youtube)
Construction	Hammer, Concrete Mixer, Handsaw, Drilling Machine, Axe	10	100%	
Music	Classic, Country, Disco, Hiphop, Jazz, Metal, Pop, Raggae, Rock	10		100%(GTZAN)
Nature	Rain Wind, Thunderstrom	12 2	70%	30%(FreeSound.org)
Speech	Male Female	7.5 3.5	80%	20% (Youtube)
Train	Train	12	80%	20% (Youtube)
Vehicles	Bus MotorCycle Tractor JCB Auto-rickshaw Van	3 1 5 1 0.25 1	90%	10%(Youtube)

A. Audio Database

The description of the audio database is summarized in Table V. Our audio signals are digitized at sampling rate of 44.1 kHz and 16-bit resolution. For training and testing purposes, the recorded audio signals are resampled to 16 kHz. We used H1n Handy recorder, EVISTR digital voice recorder and mobile handset (Samsung, Redmi) for recording audio under different environmental conditions. In addition with our audio signals, we also collected from the public multimedia websites (like youtube, GTZAN library, freesound.org). A total duration of the audio is about 81 hours.

B. Performance Evaluation

The performance of audio event detection methods are evaluated by using benchmark metrics such as precision, sensitivity, F1 score and overall accuracy [7]. The precision is defined as $Precison(Pr) = \frac{TP}{TP+FP}$ and sensitivity is defined as $Recall(Re) = \frac{TP}{TP+FN}$. The overall accuracy (Acc) is defined as $Accuracy(Acc) = \frac{TP+TN}{TP+TN+FP+FN}$ and F1-score is defined as $F1Score(F1) = \frac{2Re \times Pr}{Re+Pr}$, where TP, TN, FP, and FN denote true positives, true negatives, false positives, and false negatives, respectively.

C. Performance Comparison

In this study, we evaluated three audio event recognition schemes using the mel-frequency cepstral coefficients (MFCC) and machine classifiers such as multi-class support vectors machines (MC-SVM), fully connected feed-forward neural networks (FCFFNNs), and one-dimensional convolutional neural networks (1D-CNNs). For each frame length, the confusion matrix of three AER schemes are summarized in Table VI and their overall performances are summarized in Table VII. Results show that the FCFFNN and 1D-CNN based AER schemes had the F1-score values of 95.72% and 96.34% for audio frame size of 250 ms whereas MC-SVM based AER scheme had the F1-score values of 85.84%. The 1D-CNN based AER scheme had a class-wise accuracy is greater than 84% for audio frame size of 250 ms whereas the FCFFNN based scheme had a class-wise accuracy is greater than 80% for audio frame size of 250 ms. The computational analysis results show that the prediction time of 1D-CNN based scheme is faster than the FCFFNN based AER scheme.

IV. CONCLUSIONS AND FUTURE DIRECTIONS

In this paper, we presented three audio sound recognition (AER) schemes using the MFCC and ML classifiers such as multi-class SVM, fully connected FFNN and 1D-CNN for recognizing 7 sound classes. All three AER schemes were evaluated using a wide variety of audio signals in real-time by using our audio event monitoring GUI. Results demonstrated the significance of selection of frame length for achieving better recognition performance. Results showed that 1D-CNN outperforms other schemes for audio frame length 250 ms. Currently we are implementing the 1D-CNN based AER scheme on Raspberry Pi computing platform.

ISBN: 978-1-7281-3296-9/19/$31.00 ©2019 IEEE

TABLE VI
CONFUSION MATRIX FOR THREE AUDIO EVENT RECOGNITION (AER) SCHEMES FOR FRAME LENGTHS (FL) OF 100 MS, 250 MS, AND 500 MS.

	MC-SVM Based AER Scheme							FCFFNN Based AER Scheme							1D-CNN Based AER Scheme						
Audio frame length (FL) = 100 ms																					
	A	**C**	**M**	**N**	**S**	**T**	**V**	**A**	**C**	**M**	**N**	**S**	**T**	**V**	**A**	**C**	**M**	**N**	**S**	**T**	**V**
Aircraft (A)	**11838**	12	70	0	20	29	22	**11927**	3	16	0	22	11	12	**11872**	0	36	0	64	16	3
Construction (C)	72	**3804**	3731	2	169	632	3581	5	**11912**	29	0	36	5	4	3	**11537**	300	0	105	26	20
Music (M)	198	497	**10123**	107	237	688	141	51	389	**10598**	18	305	533	97	60	286	**10763**	64	328	416	74
Nature (N)	0	2	4	**11984**	0	1	0	0	0	6	**11983**	0	2	0	0	3	17	**11969**	0	2	0
Speech (S)	3	1458	1196	2	**9292**	27	13	1	854	1373	3	**9676**	55	29	5	680	640	2	**10628**	12	24
Train (T)	0	0	0	0	0	**11991**	0	0	0	0	0	0	**11991**	0	0	0	3	0	0	**11988**	0
Vehicle (V)	0	0	0	0	1	0	**11990**	0	0	0	0	0	1	**11990**	0	0	0	0	1	1	**11989**
Audio frame length (FL) = 250 ms																					
Aircraft	**11792**	6	45	0	100	19	29	**11954**	5	10	0	14	7	1	**11931**	2	22	0	16	20	0
Construction	61	**5371**	2434	8	610	228	3279	3	**11940**	29	0	15	2	2	2	**11604**	240	0	56	77	12
Music	98	524	**10485**	109	279	416	80	21	413	**10939**	22	219	310	67	23	261	**11223**	18	210	252	4
Nature	0	0	1	**11990**	0	0	0	0	0	0	**11991**	0	0	0	0	0	4	**11986**	0	1	0
Speech	10	1638	1090	3	**9196**	44	10	1	1054	1252	0	**9605**	47	32	7	886	883	3	**10179**	11	22
Train	0	0	0	0	0	**11991**	0	0	0	0	0	0	**11991**	0	0	0	0	0	0	**11991**	0
Vehicle	0	0	0	0	0	0	**11991**	0	0	0	0	0	0	**11991**	0	0	1	0	0	0	**11990**
Audio frame length (FL) = 500 ms																					
Aircraft	**11779**	0	31	0	141	14	26	**11978**	1	2	0	4	6	0	**11969**	0	1	0	19	2	0
Construction	125	**5340**	1301	2	1270	195	3758	0	**11899**	26	0	19	44	3	0	**11803**	135	0	46	7	0
Music	80	934	**10100**	118	245	367	147	12	536	**10935**	17	171	265	55	3	630	**10590**	33	485	243	7
Nature	0	0	5	**11986**	0	0	0	0	0	0	**11991**	0	0	0	0	0	0	**11991**	0	0	0
Speech	5	1367	1244	3	**9275**	67	30	6	1399	1503	0	**8965**	58	60	6	1486	1087	1	**9358**	20	33
Train	0	0	0	0	0	**11991**	0	0	0	0	0	0	**11991**	0	0	0	0	0	0	**11991**	0
Vehicle	0	0	0	0	0	0	**11991**	0	0	0	0	0	0	**11991**	0	0	0	0	0	0	**11991**

TABLE VII
PERFORMANCE OF THREE AUDIO EVENT RECOGNITION SCHEMES

		MC-SVM			FCFFNN			1D-CNN		
FL	**Class**	**Pr**	**Se**	**F1**	**Pr**	**Se**	**F1**	**Pr**	**Se**	**F1**
100 ms	Aircraft	97.74	98.72	98.22	99.52	99.46	99.48	99.43	99	99.21
	Construction	65.89	31.72	42.82	90.53	99.34	94.73	92.25	96.21	94.18
	Music	66.93	84.42	74.66	88.15	88.38	88.26	91.52	89.75	90.62
	Nature	99.08	99.94	99.5	99.82	99.93	99.87	99.45	99.81	99.62
	Speech	95.6	77.49	85.59	96.38	80.69	87.83	95.52	88.63	91.94
	Train	89.69	100	94.56	95.18	100	97.53	96.2	99.97	97.98
	Vehicle	76.14	99.99	86.45	98.82	99.99	99.4	99	99.98	99.48
	Avg.	**84.44**	**84.61**	**83.11**	**95.49**	**95.40**	**95.30**	**96.20**	**96.19**	**96.15**
250 ms	Aircraft	98.58	98.34	98.45	99.79	99.69	99.73	99.73	99.49	99.6
	Construction	71.24	44.54	54.81	89.02	99.57	94.1	90.99	96.77	93.79
	Music	74.59	87.44	80.5	89.44	91.22	90.32	90.7	93.59	92.12
	Nature	99	99.99	99.49	99.81	100	99.9	99.82	99.95	99.88
	Speech	90.28	76.69	82.93	97.48	80.1	87.93	97.3	84.88	90.66
	Train	94.43	100	97.13	97.03	100	98.49	97.07	100	98.51
	Vehicle	77.91	100	87.58	99.15	100	99.57	99.68	99.99	99.83
	Avg.	**86.58**	**86.71**	**85.84**	**95.96**	**95.80**	**95.72**	**96.47**	**96.38**	**96.34**
500 ms	Aircraft	98.24	98.23	98.23	99.84	99.89	99.86	99.92	99.81	99.86
	Construction	69.88	44.53	54.39	86	99.23	92.14	84.79	98.43	91.1
	Music	79.64	84.22	81.86	87.71	91.19	89.41	89.64	88.31	88.97
	Nature	98.98	99.99	99.48	99.85	100	99.92	99.71	100	99.85
	Speech	84.85	77.34	80.92	97.88	74.76	84.77	94.44	78.04	85.46
	Train	94.91	100	97.38	96.98	100	98.46	97.78	100	98.87
	Vehicle	75.16	100	85.81	99.02	100	99.5	99.66	100	99.82
	Avg.	**85.95**	**86.33**	**85.44**	**95.33**	**95.01**	**94.87**	**95.13**	**94.94**	**94.85**

V. ACKNOWLEDGMENTS

This research work was carried out as a part of the cluster projects of the Special Manpower Development Program for Chips to System Design under Ministry Of Electronics & Information Technology (MeitY) Grant, Govt. of India.

REFERENCES

[1] E. Cakir, G. Parascandolo, T. Heittola, H. Huttunen and T. Virtanen, "Convolutional Recurrent Neural Networks for Polyphonic Sound Event Detection," *IEEE/ACM Trans. Audio, Speech, and Language Process.,* vol. 25, no. 6, pp. 1291-1303, 2017.

[2] N. Almaadeed, M. Asim, S. Al-Maadeed, A. Bouridane, and A. Beghdadi, "Automatic Detection and Classification of Audio Events for Road Surveillance Applications," *Sensors (Basel)*, vol. 18, no. 6, 2018.

[3] J. Wang, C. Lin, B. Chen and M. Tsai, "Gabor-Based Nonuniform Scale-Frequency Map for Environmental Sound Classification in Home Automation," *IEEE Trans. Automation Science and Engineering,* vol. 11, no. 2, pp. 607-613, 2014.

[4] S. Souli and Z. Lachiri, "Audio Sounds Classification using Scattering Features and Support Vectors Machines for Medical Surveillance," *Applied Acoustics,* Elsevier, vol. 130, pp. 270-282, 2018.

[5] W. Yang and S. Krishnan, "Combining Temporal Features by Local Binary Pattern for Acoustic Scene Classification," *IEEE/ACM Trans. Audio, Speech, and Language Processing,* vol. 25, no. 6, pp. 1315-1321, 2017.

[6] S. Sigtia, A. M. Stark, S. Krstulovic and M. D. Plumbley, "Automatic Environmental Sound Recognition: Performance Versus Computational Cost," *IEEE/ACM Trans. Audio, Speech, and Language Processing,* vol. 24, no. 11, pp. 2096-2107, 2016.

[7] A. Mesaros, T. Heittola and T. Virtanen, "TUT Database for Acoustic Scene Classification and Sound Event Detection," in *Proc. 24th European Signal Processing Conference (EUSIPCO),* pp. 1128-1132, 2016.

[8] H. Zhang, I. McLoughlin and Y. Song, "Robust Sound Event Recognition Using Convolutional Neural Networks," in *Proc.IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)* pp. 559-563, 2015.

[9] O. Gencoglu, T. Virtanen and H. Huttunen, "Recognition of Acoustic Events using Deep Neural Networks," in *Proc. 22nd European Signal Processing Conf. (EUSIPCO),* pp. 506-510, 2014.

[10] H. D. Tran and H. Li, "Sound Event Recognition With Probabilistic Distance SVMs," *IEEE Trans. Audio, Speech, and Language Processing,* vol. 19, no. 6, pp. 1556-1568, 2011.

[11] X. Valero and F. Alias, "Gammatone Cepstral Coefficients: Biologically Inspired Features for Non-Speech Audio Classification," *IEEE Trans. Multimedia,* vol. 14, no. 6, pp. 1684-1689, 2012.

[12] F. Beritelli and R. Grasso, "A Pattern Recognition System for Environmental Sound Classification Based on MFCCs and Neural Networks," in *Proc. 2nd Int. Conf. on Signal Process. and Commn. Systems,* pp. 1-4, 2008.

[13] A. Rabaoui, M. Davy, S. Rossignol and N. Ellouze, "Using One-class SVMs and Wavelets for Audio Surveillance," *IEEE Trans. Info. Forensics and Security,* vol. 3, no. 4, pp.763-775, 2008.

ISBN: 978-1-7281-3296-9/19/$31.00 ©2019 IEEE