

# Forensic Audio Analysis and Event Recognition for Smart Surveillance Systems

Yusuf Ozkan and Buket D. Barkana  
Electrical Engineering Department  
University of Bridgeport, Bridgeport  
E-mails: {yozkan@my.bridgeport.edu; bbarkana@bridgeport.edu}

**Abstract**— In this study, we propose forensic audio analysis and event recognition framework for smart surveillance systems. Surveillance systems are getting a growing attention. Integration of audio surveillance systems to IP cameras offers advantages over video surveillance systems solely. For instance, an audio surveillance system can provide an additional information about the exact moment of an incident that might be occurring out of the range of surveillance cameras. In our work, mel-frequency cepstral coefficients (MFCCs), energy, pitch range (PR), and linear prediction coefficients (LPC) feature sets were extracted and evaluated on an open-access DASE database, which contains nine audio events including glass breaking, dog barking, scream, gunshot, explosions, police sirens, door slams, footsteps, and house alarm sounds. K-nearest-neighbor (KNN), support vector machines (SVMs), Gaussian Mixture Model (GMM), and classifier fusion classifiers are designed. All classifiers achieved promising accuracy rates.

**Index Terms**—Forensic audio analysis; surveillance systems; MFCCs; pitch range, LPCs

## I. INTRODUCTION

Security is one of the sectors that have been getting more importance and demand in the last decades. This demand is not only about data safety or cyber-security, but is also to keep public places and personal properties secure. Ongoing economic crises, unemployment, drug usage, and retrogression in big cities brought higher crime rates. This increased the demand for crime scene investigators and security systems in cities. Enforcement officers have great responsibilities to prevent these undesirable incidents. In the line of their duty, they maintain public order, enforce the law, detect, and prevent criminal activities. These duties require a broad surveillance network besides reliable intelligence service and advanced biometric solutions. For this reason, the number of Closed-Circuit Television (CCTV) cameras is increased dramatically in the last decade. The Internet Protocol (IP) cameras are integrated into CCTV cameras. In addition, the need for covering blind spots of cameras led the sector to add more cameras to the existing video surveillance systems. These demands increased the total cost of video surveillance systems. Integration of an audio surveillance system to IP cameras is an alternative to overcome some of the challenges of conventional systems. An audio surveillance system can provide additional information about the exact moment of an incident that might be occurring out of the range of cameras. Possible suspicious audio events may contain glass

breaking, house alarm and police siren sounds, gunshot, footsteps, dog barking, scream, etc. (Fig.1).



Figure 1. Possible suspicious/dangerous audio events

## II. RELATED WORK

Rabaoui et al. studied nine environmental sounds: human screams, gunshots, glass breaking, explosions, door slams, dog barks, phone rings, children voices, and machines. Wavelet-based features, Mel-frequency cepstral coefficients (MFCCs), individual temporal and frequency features were extracted and one-class support vector machines (1-SVMs) were used. Gaussian radial basis function (RBF) kernel width was scaled in a distribution-dependent way to avoid under-fitting/over-fitting problems. However, their proposed scaled-width kernel did not improve the recognition accuracies significantly [1]. Rabaoui et al. proposed 1-SVMs together with a dissimilarity measure and compared its performance with a Hidden Markov Model (HMM) system. They reported that their proposed system outperformed the HMM system [2]. Chu et al. studied matching pursuit (MP) algorithm and extracted four features to classify fourteen acoustic events. They compared the results of k-nearest-neighbor (KNN) and Gaussian mixture model (GMM) classifiers. The highest accuracy (83.9%) was calculated when a set of MP and MFCC features was applied to the GMM classifier [3].

Uzgent et al. proposed a new feature set based on pitch range (PR) calculations. They classified sounds by using SVM and artificial neural networks (ANNs). Their work showed that the combination of PR and MFCCs improved the classification

accuracy about 16% for seven environmental sound events including gunshot, glass breaking, scream, dog barking, rain, engine, and restaurant noise [4]. Choi et al. proposed a hierarchical recognition method to detect nine common suspicious acoustic events. MFCCs, timbre, spectral statistics (SS), LPCs, and their combination were extracted as features. They designed a GMM classifier. In addition, six background noise types were added to sound events in different signal-to-noise ratios (SNR). They reported that once the background noise was detected correctly, the recognition of suspicious events was improved by 7% [5].

Ruiz-Martinez et al. compared the performances of three classifiers, SVM, logistic regression (LR), and backpropagation artificial neural network (BP-ANN), to detect nine environmental sound events. MFCCs, matching pursuit, and their combination were used. The highest overall recognition accuracy was reported by using MFCCs features and SVM classifier. MFCCs + MP and MP features with the ANN and LR classifiers did not perform as well as expected [6]. Chang et al. studied six indoor sound events: coughing, screaming, infants' crying, glass breaking, laughing, and doorbell ringing. Nine temporal and 15 spectral features were extracted. The sequential floating forward selection (SFFS) was employed to find the most distinct features among the events prior classification. It was reported that four features, peak, valley, contrast of octave-based, and MFCCs, provided the most selective characteristics among these six classes. They achieved an overall accuracy of 86% [7].

Uzkent et al. introduced an abnormal sound detection method for audio surveillance systems. They developed pitch-range based feature set that contains two features. The features were extracted from the statistical orders (minimum, maximum, standard deviation and mean) of fundamental frequency (F0). SVM and k-means clustering classifiers were used to evaluate the performance of the PF set for three environmental sound events, which were restaurant, engine, and rain. Recognition rates ranged between 91.3-95.4% and 81-92.8% by SVM and the k-means clustering classifier, respectively [8]. In [2011b], their database consists of three sound events, gunshot, dog barking, and glass breaking. Overall accuracy was calculated between 79-92% by using the two PR features [9].

Nandwana et al. studied five abnormal sound events encountered on the roads, ambulance siren, railroad crossing bell, tire screech, car honk, and glass breaking. 39 features including, energy, 12 MFCCs, the velocity, and acceleration coefficients, were calculated. Frame size was selected as 25ms and overlapping was applied. They compared the classification results of Gaussian Mixture Model-Universal Background Model (GMM-UBM), GMM-supervector (SV), and identity vector (I-vector). GMM-SV obtained the highest accuracy of 87% [10]. Crocco et al. did literature survey regarding audio surveillance systems. They analyzed previous works in terms of their pros and cons [11].

### III. EXPERIMENTAL DATA

Open access database DASE of auditory suspicious events was used to evaluate the performance of our work [12]. DASE consists of the nine most common suspicious audio events and

total of 3105 recordings of gunshot (GS) (# of 332 samples), explosions (E) (# of 400 samples), scream (S) (# of 300 samples), glass breaking (GB) (# of 352 samples), dog barking (DB) (# of 204 samples), police sirens (PS) (# of 446 samples), door slams (DS) (# of 492 samples), footsteps (FS) (# of 379 samples), and house alarm (HA) (# of 200 samples) sounds.

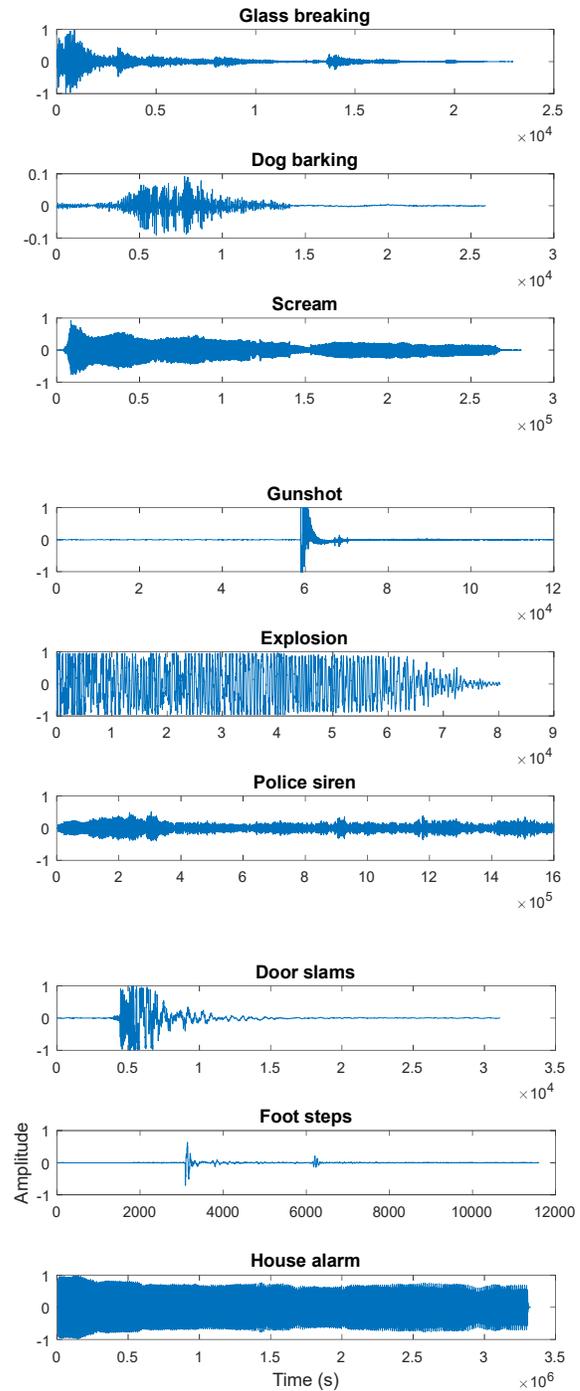


Figure 2. Example sounds in time domain. Gunshot, door slams, and footsteps show similarities in time domain. These sounds occur in a shorter-duration of time than the other sounds in the database and present more abrupt amplitude changes in a short time.

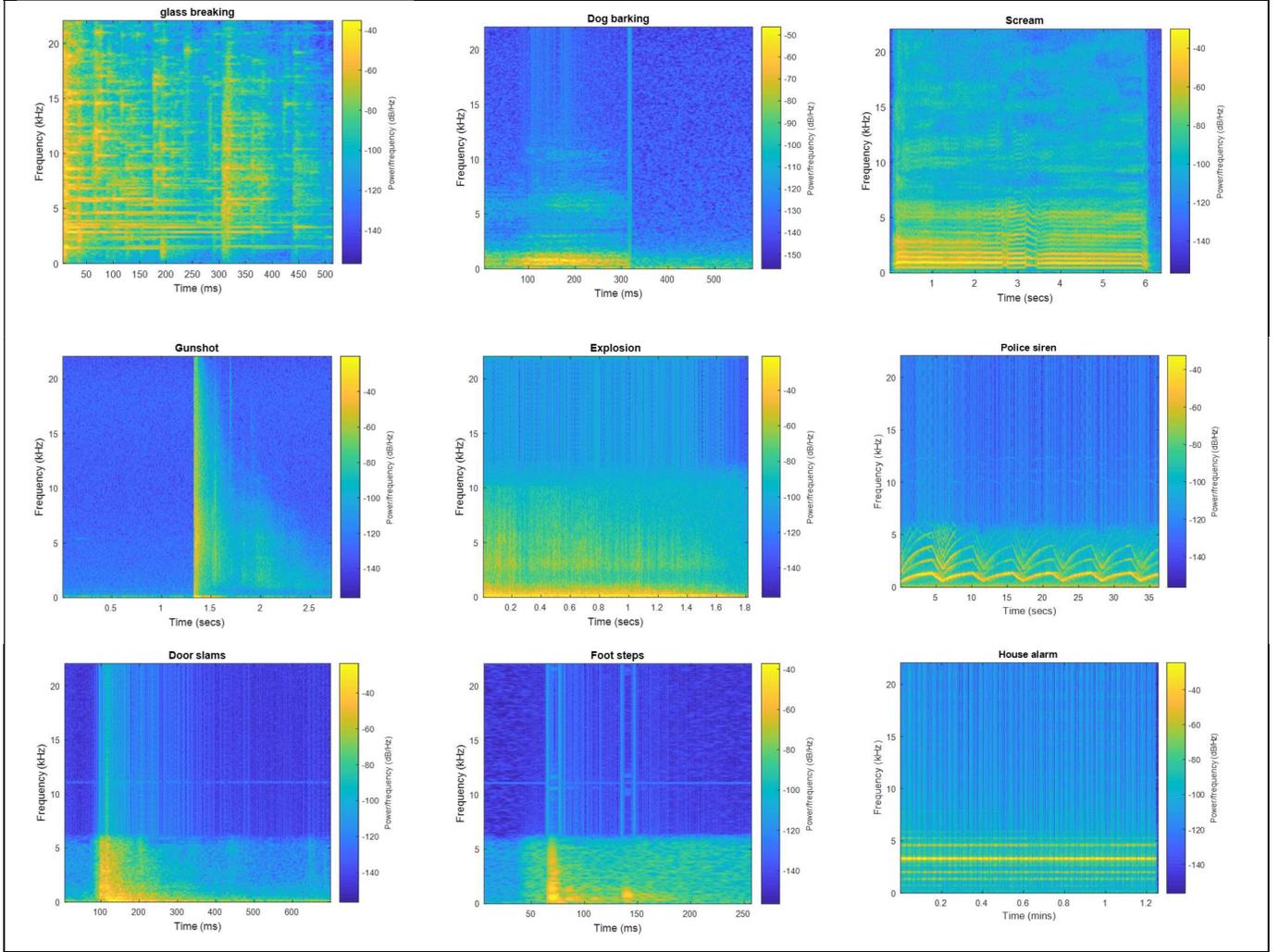


Figure 3. Examples of spectrograms of the sounds given in Fig. 2.

The database provides training and test sets. The signals are sampled at 44.1 kHz and digitized at 16-bit. Monophonic versions of the audios are used. We excluded ‘HA0121’ sound event in our calculations due to the difficulty of its PR features’ extraction. Example sounds and their spectrograms are shown in Figs. 2 and 3, respectively.

Spectrograms are calculated by using a Kaiser filter with a duration of 512 in samples, 500 samples overlapping, and 1024-point FFT. Sampling frequency is 44.1 kHz. Blue colors present low energy contents, whereas green and yellow colors present high-energy contents. Glass breaking, scream, gunshot, and explosion sounds have wider frequency content than the dog barking, police siren, door slams, and foot step sounds. The low and high frequency ends of the spectra of the glass breaking and house alarm sounds carry little energy. For the other sound events, the low frequency end carries higher energy.

In this work, we utilized MFCCs, pitch range (PR), LPCs, and energy features.

#### A. MFCCs and energy features

MFCCs are well-known spectral features and are widely used in signal processing. In this work, MFCCs are calculated by using a 50ms Hamming window with 90% overlapping. Cepstral coefficients are derived from the short-time spectrum of a signal. MFCC coefficients are calculated as given in (1-4) [13]. First, short-time Discrete Fourier transform (DFT)  $X(n, w_k)$  is defined in (1), where  $x[n]$  is a discrete audio signal,  $w[n]$  is a window.  $w_k$  is  $(2\pi/N)*k$  where  $N$  defines the length of DFT.  $E_{mel}$  defines the resulting energies in (2).  $V_l(w_k)$  is defined as the frequency response of the mel-scale filter.  $L_l$  and  $U_l$  indicate lower and upper frequencies.  $C_{mel}$  presents mel-cepstrum where  $R$  is the number of filters,  $E_{mel}(n, l)$  is the energy of each frame that is calculated for the sound frame at time  $n$ .

$$X(n, w_k) = \sum_{m=-\infty}^{\infty} x[m]w[n-m]e^{-jw_k m} \quad (1)$$

$$E_{mel}(n, l) = \frac{1}{A_l} \sum_{k=L_1}^{U_1} |V_l(w_k)X(n, w_k)|^2 \quad (2)$$

$$A_l = \sum_{k=L_1}^{U_1} |V_l(w_k)|^2 \quad (3)$$

$$C_{mel}[n, m] = \frac{1}{R} \sum_{l=0}^{R-1} \log\{E_{mel}(n, l)\} \cos\left(\frac{2\pi l}{R} lm\right) \quad (4)$$

### B. Pitch-range (PR) features

Pitch range features are useful to see the changes in the frequency of a sound signal. Two PR features are calculated as in (5-7) [14]. First, autocorrelation function (ACF) is defined in (5), where  $N_1$  is length of the sound frame.  $\tau$  is time offset. The frames are overlapped by 50 percent and the frame size is selected as 25ms. Time delay  $T_i$  is calculated for each frame,  $i=1, 2, \dots, M_1$ .  $M_1$  is the number of the frames. The pitch  $P_i$  is the reciprocal of the time delay per frame ( $p_i = \frac{1}{T_i}$ ). PR1 and PR2 features are calculated by using  $P_i$  set. Smaller frame size  $N_2$  ( $N_2 < N_1$ ) is chosen with no overlapping by analyzing the  $P_i$  set.  $P_j, j=1, 2, \dots, M_2$  represents the mean pitch value for each frame.  $M_2$  is the number of the frame size.  $\text{std}\{P\}$  and  $\text{std}\{\hat{P}\}$  are the standard deviation of  $P_i$  and  $\hat{P}$  in (7), respectively. Two PR features are calculated as in (6 and 7)

$$\text{ACF}(\tau) = \sum_{n=0}^{N_1-1} x(n)x(n+\tau) \quad (5)$$

$$\text{PR}_1 = \frac{\max(\hat{P})}{\min(\hat{P})} \quad (6)$$

$$\text{PR}_2 = \frac{\text{std}\{P\}}{\text{std}\{\hat{P}\}} \quad (7)$$

### C. Linear Predictive Coefficients

Linear prediction uses an all-pole model to represent the sound signals and to relate to formants [15]. Linear-prediction coefficients (LPCs) approximate the current sound sample as a linear combination of past sound samples. It is given in (8).  $p$  is defined as the number of previous samples that are used to approximate. In this work, we set the  $p$  value as 10. The coefficients are calculated by minimizing the root square error (RSE) between the actual and the predicted sound signals. Then coefficients are calculated as in (9).  $r = [r(1), r(2), \dots, r(p)]^T$  is the autocorrelation vector.  $R$  is a  $p$  by  $p$  Toeplitz autocorrelation matrix

$$\begin{aligned} x(n) &= a_1 x(n-1) + a_2 x(n-2) + \dots + a_n x(n-p) \\ n &= 0, 1, 2, \dots, N-1 \end{aligned} \quad (8)$$

$$a = R^{-1}r \quad (9)$$

## IV. CLASSIFICATION

K-nearest-neighbor (KNN), support vector machines (SVMs), Gaussian Mixture Model (GMM), and a classifier fusion were used as classifiers. The decision for the classifier fusion was made by using majority voting concept.

### A. Support Vector Machines

SVM is a frequently used supervised classification method. We implemented this method by using LibSVM toolbox [16]. We left most of the learning parameters in default settings. 3-fold cross validation and Radial Basis Function (RBF) kernel were used

### B. K-Nearest Neighbor Algorithm

KNN is a fundamental classification algorithm that determines the label of an input data based on its similarity to data in the training set.  $K$  is the number of the neighbors and is usually an odd number to avoid ties. We selected  $K=1$ . Euclidean distance, the Minkowski distance, and the Manhattan distance algorithms are mainly applied methods to determine the degree of proximity. In our implementation, Euclidean distance was adopted

### C. Gaussian Mixture Model

A Gaussian mixture model is formed in two steps. Firstly, data set and  $k$ , which is the number classes, technically called the number of mixtures, are chosen to form a universal background model (UBM). UBM has  $k$  mean and covariance matrices that represent each class. If the variance between classes is different, contours can be formed at the UBM step. Otherwise, an adaptation step has to be applied to enroll each component to the related cluster. When the correlation between classes is too high, it leads to ill-conditioned covariance formation, which leads GMM to fail. At the adaptation stage,  $k$  means and covariance matrixes are adjusted by using a labeled training set. This step yields  $k$  unique mean and covariance matrices. A Gaussian mixture density is a weighted sum of  $M$  components density given in (10) [17, 18].  $p_i$  is mixture weights.  $b_i(\vec{x})$  are mixture densities. Each mixture density is a D-dimensional Gaussian function given in (11).  $\vec{\mu}_i$  is a mean vector.  $\Sigma_i$  is a covariance matrix. A complete Gaussian mixture model includes mean vectors, mixture weights, and covariance matrices. These variables are defined as  $\lambda$  in (12)

$$p(\vec{x}|\lambda) = \sum_{i=1}^M p_i b_i(\vec{x}) \quad (10)$$

$$b_i(\vec{x}) = \frac{1}{(2\pi)^{D/2} |\Sigma_i|^{1/2}} \exp\left\{-\frac{1}{2}(\vec{x}-\vec{\mu}_i)' \Sigma_i^{-1} (\vec{x}-\vec{\mu}_i)\right\} \quad (11)$$

$$\lambda = \{p_i, \vec{\mu}_i, \Sigma_i\} \quad i = 1, \dots, M \quad (12)$$

In the trial and error stage, we evaluated several number of mixtures as 4, 8, 16, 32, 64, 128, and 256. Based on the performance, we set the number of mixtures to 4 in this work

#### D. Classifier Fusion

In this study, equally weighted classifier fusion was implemented. Majority voting is practiced. The predicted final class label is chosen as the class label, which was predicted most frequently by the classifiers. If all three classifiers predict different labels, the final decision was given based on the decision of the SVM classifier

#### V. RESULTS

Classifier accuracies for each sound event are given in the Table I. The highest accuracy rates are depicted in bold in the table.

TABLE I. OVERALL CLASSIFICATION ACCURACIES (%).

Sound events	MFCCs + PRs + LPC (25 dimensions)				MFCCs + PRs (15 dimensions)			
	KNN	SVM	GMM	Classifier fusion	KNN	SVM	GMM	Classifier fusion
<i>GB</i>	73.418	75.949	21.519	<b>77.21</b>	<b>73.418</b>	<b>73.418</b>	51.899	73.418
<i>DB</i>	96.667	<b>97.5</b>	46.667	<b>97.5</b>	<b>96.667</b>	<b>96.667</b>	43.333	<b>96.667</b>
<i>S</i>	77.273	<b>84.091</b>	14.773	80.682	<b>77.273</b>	<b>77.273</b>	44.318	<b>77.273</b>
<i>GS</i>	56.311	<b>60.194</b>	13.592	54.369	56.311	56.311	21.359	56.311
<i>E</i>	<b>55.738</b>	44.262	21.311	45.902	<b>55.738</b>	<b>55.738</b>	50.82	<b>55.738</b>
<i>PS</i>	84.211	83.459	<b>92.481</b>	89.474	84.211	<b>84.211</b>	65.414	84.211
<i>DS</i>	75.342	84.932	73.288	<b>88.356</b>	75.342	75.342	<b>77.397</b>	75.342
<i>FS</i>	72.973	<b>75.676</b>	54.955	<b>75.676</b>	<b>72.973</b>	<b>72.973</b>	62.162	<b>72.973</b>
<i>HA</i>	62.5	<b>62.5</b>	21.429	60.714	<b>62.5</b>	<b>62.5</b>	26.786	<b>62.5</b>
<i>Overall acc.</i>	70.92	77.36	46.38	<b>77.92</b>	74.92	<b>77.36</b>	52.28	74.91
Sound events	MFCCs (13 dimensions)							
	KNN	SVM	GMM	Classifier fusion				
<i>GB</i>	<b>78.481</b>	77.215	58.228	<b>78.481</b>				
<i>DB</i>	96.667	<b>98.333</b>	69.167	<b>98.333</b>				
<i>S</i>	75	79.545	45.455	<b>81.818</b>				
<i>GS</i>	55.34	<b>58.252</b>	18.447	54.369				
<i>E</i>	<b>55.738</b>	42.623	27.869	44.262				
<i>PS</i>	<b>86.466</b>	85.714	60.15	<b>86.466</b>				
<i>DS</i>	73.973	81.507	76.712	<b>86.986</b>				
<i>FS</i>	71.171	73.874	69.369	<b>75.676</b>				
<i>HA</i>	62.5	62.5	37.5	<b>67.857</b>				
<i>Overall acc.</i>	74.91	76.36	55.18	<b>77.92</b>				

In this work, feature variance of police sirens and door slams sounds were calculated correlated within the classes by using the three feature sets. It was observed that the classification accuracies of these classes were calculated higher compared to the other classes by the GMM classifier. However, the GMM classifier was outperformed by the KNN, SVM, and classifier fusion. Chu et al. [19] reported similar findings in their work that classified four sound effects in movies. The classifier fusion slightly achieved higher accuracies for most sound events than the other classifiers. We observed that the LPC-10 feature set did affect the classification accuracies negatively

#### VI. CONCLUSION

In this paper, we have presented a forensic audio analysis and event recognition framework for smart surveillance systems. Nine suspicious sound events were analyzed at the time and spectral domains and their characteristics were investigated. MFCCs, energy, PR, and LPC features were used in KNN, SVM, GMM, and classifier fusion. By using all feature sets, the classifiers achieved an overall accuracy of 74.92%, 77.36%,

46.38%, 77.92%, respectively. Classifier fusion outperformed the others in terms of overall accuracy.

It was noticed that house alarm, explosion, and gunshot audio events were classified poorly by all classifiers with all features. The house alarm audio events were misclassified by 16% as scream audio event and by 12.5% as police siren audio event. Explosion events were misclassified by 13% as gunshot and by 13% as door slam.

The highest classification rates were achieved in dog barking, door slam, and police siren events as of 98%, 87%, and 86%, respectively, by using classifier fusion and MFCCs feature set only. GMM classifier accuracies were calculated as of 46.38%, 52%, and 55.2% by using MFCC+PR+LPC, MFCC+PR, MFCC feature sets respectively.

#### REFERENCES

- [1] A. Rabaoui, M. Davy, S. Rossignol, Z. Lachiri, and N. Ellouze, "Improved one-class SVM classifier for sounds classification," in 2007 IEEE Conference on Advanced Video and Signal Based Surveillance, 2007, pp. 117-122)
- [2] A. Rabaoui, M. Davy, S. Rossignol, and N. Ellouze, "Using one-class SVMs and wavelets for audio surveillance," IEEE Transactions on information forensics and security, vol. 3, no. 4, pp. 763-775, 2008.

- [3] 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.
- [4] B. Uz Kent, B. D. Barkana, and H. Cevikalp, "Non-speech environmental sound classification using SVMs with a new set of features," *International Journal of Innovative Computing, Information and Control*, vol. 8, no. 5, pp. 3511-3524, 2012.
- [5] W. Choi, J. Rho, D. K. Han, and H. Ko, "Selective background adaptation based abnormal acoustic event recognition for audio surveillance," in *2012 IEEE Ninth International Conference on Advanced Video and Signal-Based Surveillance*, 2012, pp. 118-123.
- [6] C. A. Ruiz-Martinez, M. T. Akhtar, Y. Washizawa, and E. Escamilla-Hernandez, "On investigating efficient methodology for environmental sound recognition," in *2013 International Symposium on Intelligent Signal Processing and Communication Systems*, 2013, pp. 210-214.
- [7] C.-Y. Chang and Y.-P. Chang, "Application of abnormal sound recognition system for indoor environment," in *2013 9th International Conference on Information, Communications & Signal Processing*, 2013, pp. 1-5.
- [8] B. D. Barkana and B. Uz Kent, "Environmental noise classifier using a new set of feature parameters based on pitch range," *Applied Acoustics*, vol. 72, no. 11, pp. 841-848, 2011.
- [9] B. Uz Kent and B. D. Barkana, "Pitch-range based feature extraction for audio surveillance systems," in *2011 Eighth International Conference on Information Technology: New Generations*, 2011, pp. 476-480.
- [10] M. K. Nandwana and T. Hasan, "Towards Smart-Cars That Can Listen: Abnormal Acoustic Event Detection on the Road," in *INTERSPEECH*, 2016, pp. 2968-2971.
- [11] M. Crocco, M. Cristani, A. Trucco, and V. Murino, "Audio surveillance: A systematic review," *ACM Computing Surveys (CSUR)*, vol. 48, no. 4, p. 52, 2016.
- [12] B. D. Barkana, N. John, and I. Saricicek, "Auditory Suspicious Event Databases: DASE and Bi-DASE," *IEEE Access*, vol. 6, pp. 33977-33985, 2018.
- [13] T. F. Quatieri, *Discrete-time speech signal processing: principles and practice*. Pearson Education India, 2002.
- [14] B. D. Barkana and J. Zhou, "A new pitch-range based feature set for a speaker's age and gender classification," *Applied Acoustics*, vol. 98, pp. 52-61, 2015.
- [15] Y. Zhan, H. Leung, K.-C. Kwak, and H. Yoon, "Automated speaker recognition for home service robots using genetic algorithm and Dempster-Shafer fusion technique," *IEEE Transactions on Instrumentation and Measurement*, vol. 58, no. 9, pp. 3058-3068, 2009.
- [16] C.-C. Chang and C.-J. Lin, "LIBSVM: a library for support vector machines," *ACM transactions on intelligent systems and technology (TIST)*, vol. 2, no. 3, p. 27, 2011.
- [17] McLachlan, G., and D. Peel. *Finite Mixture Models*. Hoboken, NJ: John Wiley & Sons, Inc., 2000)
- [18] A. A. Mallouh, Z. Qawaqneh, and B. D. Barkana, "New transformed features generated by deep bottleneck extractor and a GMM-UBM classifier for speaker age and gender classification," *Neural Computing and Applications*, vol. 30, no. 8, pp. 2581-2593, 2018.
- [19] S. Chu, S. Narayanan, and C.-C. J. Kuo, "Towards parameter-free classification of sound effects in movies," in *Applications of Digital Image Processing XXVIII*, 2005, vol. 5909, p. 59091J: International Society for Optics and Photonics.