Journal of Advances in Mathematics and Computer Science

35(2): 25-41, 2020; Article no.JAMCS.54824
ISSN: 2456-9968
(Past name: British Journal of Mathematics & Computer Science, Past ISSN: 2231-0851)

Security Detection in Audio Events: A Comparison of Classification Methods

Alissar Nasser[1*]

[1]Faculty of Economic Sciences and Business Administration, Lebanese University, Hadath, Lebanon.

Author's contribution

The sole author designed, analysed, interpreted and prepared the manuscript.

Article Information

DOI: 10.9734/JAMCS/2020/v35i230247
Editor(s):
(1) Prof. Sheng Zhang, Bohai University, China.
Reviewers:
(1) David Lizcano, Madrid Open University, Spain.
(2) Ajay B. Gadicha, P. R. Pote College of Engineering and Management, India.
Complete Peer review History: http://www.sdiarticle4.com/review-history/54824

Original Research Article

Received: 29 December 2019
Accepted: 05 March 2020
Published: 14 March 2020

Abstract

The security of public places is becoming important with the increased rate of violence and subversion. Recently, several types of research have been proposed to automatically detect abnormal behavior in public places like a car crash, violence or other hazardous events in an attempt to improve security and save lives. Furthermore, most of the researches are using supervised classifications techniques to classify the audio signals. This paper proposes the use of the kernel principal component analysis (KPCA) to reduce the number of MFCC features extracted from the audio signal and then apply an unsupervised classification algorithm. Moreover, this paper presents the results of several supervised and unsupervised classification methods for audio events detection and compares these results with the result of the proposed approach. Experiments are done using a real data set recorded at the mean of public transportation. The obtained results reveal that K-means on 2 KPCA components gave good results for triggering a true alarm as well as detecting a false alarm; where the percentages of false and missed alarms were 4.5% and 7.8% respectively; whereas these values were 0.8% and 9.3% respectively for kernel k-means. Notwithstanding the DNN network gave the best results with a false alarm rate of 0% and 1.4% missed alarm.

Keywords: Audio event detection; MFCC; classification; unsupervised; supervised; kernel PCA; K-means; DNN; kernel Davies and Bouldin index.

**Corresponding author: E-mail: alissar.nasser@gmail.com, alissar.nasser.1@ul.gov.lb;*

1 Introduction

Nowadays CCTV surveillance systems are integrated into the means of public transport and in cities for the purpose of security and to decrease the rate of crime. The analysis of videos recorded by the CCTV system is not always sufficient to take into account the passengers activities like screaming or any violent action especially when the environment is too busy. For that reason video surveillance system must be accompanied by an audio surveillance system where the detection of any unusual audio event is becoming possible. Many researches proposed a join audio-visual approach for speech recognition and audio event detection. Authors in [1] presented an end-to-end audiovisual approach based on a 2- layer Bidirectional Gated recurrent Units which simultaneously learn to extract both audio and visual features. In [2], the authors propose a novel method to incorporate audios and videos by building an on-line Audio-video concurrence matrix that detects events using hierarchical clustering. In [3] the authors used lips reading to speech recognition. The authors in [4] presented a platform for audio-visual video analysis to assist agencies in analyzing and identifying suspects from large scale videos recorded after a terrorist attack.

Recently, the classification of audio events detection is becoming an active topic because of the increased rate of violence and thank to the availability of audio signals recorded in public places. However, audio signals are considered high-dimensional data and therefore they need special treatment before the step of classification. Most of the researches are using supervised classification techniques which require labeled dataset to learn the classifier. Labeling a dataset isn't an easy task and in a real-world application, labels are not available.

This drawback motivates the authors of the current paper to investigate the usefulness of the unsupervised classifications. Therefore, this paper studies the effectiveness of applying non-supervised techniques in the context of audio event detection and presents a new approach based in three steps; the first transform the audio signal into features; the second one is using the kernel principal component analysis (KPCA) to reduce the features from the first step and take the components that are sufficient for classification. In the third step an unsupervised classification algorithm is applied to classify the audio signal.

The automated event detection system passes through the feature extraction step before the process of classification. By feature extraction, we recognize the signal components that are essential for classification and discard unnecessary components. In the context of speech recognition, three main feature extraction techniques are used namely the Mel Frequency Cepstral Coefficients (MFCCs), the Linear Prediction Coefficients (LPCs) and Linear Prediction Cepstral Coefficients (LPCCs) [2,5,6]. Sharan and Moir in [7] provide an overview of the different features extraction techniques used in an automated sound system, from cepstral features like MFCC and Gammatone Cepstral Coefficients GTCC, to time-frequency features like Central moments; these methods were followed by the use of either a Gaussian Mixture Model (GMM) or a Hidden Markov Model (HMM) to classify the audio signal.

In the other side, several classification architectures have been proposed in the literature for audio events detection. Vacher and al. used a GMM classifiers trained on several features to detect screams or gunshots [8]; Laffite and al. measured the performances of diverse neural network architectures to detect shouts and scream in transportation means [9]; Valenzise and al. used a short audio frame where the input signal is used to modeling the background sounds [10]. Rouas and al. used a combination of GMM and SVM to reduce the effect of the background sounds on the classification results in order to detect scream in an outdoor environment [11], while, Ntalampiras and al. used a two-stage GMM classifier where in the first stage the audio is classified into normal or abnormal events and in the second stage classifier, it is classified into a specific class [12]; Foggia and al. used a pool of SVM classifiers learned on a bag of word approach to detect audio events for surveillance applications [13].

In this paper, the Mel Frequency Cepstral Coefficients is used to extract 39 coefficients correspond to the first 12 MFCC coefficients, three energy terms, the first and second derivatives of the 12 coefficients. The kernel PCA method is then used to reduce the 39 features into smaller number to capture the maximum

information in the data set; finally the unsupervised k-means algorithm is applied to separate abnormal events like screaming, shouting or any dangerous events that could happen in public transportation means from the noise and the normal speech. Moreover, the paper presents a comparative study of several classifiers from both supervised and unsupervised methods used in the literature for audio event detection. The audio signal used is recorded in public transport means by the French project SAMSIT.

The paper is organized as follows: Section 1 presents the MFCC features extraction method, the proposed method and briefly describes several classification methods. Section 2 provides the results of the experiments and the evaluation of the methods, and section 3 concludes the paper.

1.1 MFCC feature extraction

Automatic speech recognition is a well-established area of research, from which technologies for the development of real-world applications emerge. However, in a real application, recognition systems are subject to many sources of noise causing significant degradation of performance. The speech signal has a very large variability. The same person never pronounces a word twice in the same way. The speech rate may vary; the duration of the signal is then changed. Finally, the speech is a means of communication where many elements come into play, such as the place, the speaker's emotion, the relationship that is established between the speakers (stressful or friendly). The acoustics of the place (protected environment or noisy environment); these factors influence the form and content of the message.

Therefore, it is necessary to format the audio signal before any process of clustering. For this, some operations are performed before any treatment. The signal is first filtered and then sampled at a permitted frequency. A pre-emphasis step is made to record the high frequencies. Then the signal is segmented into frames. Each frame consists of a fixed number N of speech samples. In general, N is set such that each frame corresponds to about 16-20 ms of speech (duration during which speech can be considered as stationary). In this example, a 16 ms frame is used followed by a Fast Fourier Transform (FFT) to calculate the short-range spectrum. Subsequently, sixteen overlying Mel-scale triangular filters are computed. A logarithm transformation is applied to the filter bank outputs followed by Discrete Cosine Transform which consequently generates 12 cepstrum coefficients. To detect the spectrum relationships between neighboring frames, first and second derivatives of the 12 attributes plus three terms of energy are added to end up with 39 MFCC attributes [11,14].

The following we describe the signal processing used in this article:

1. An automatic segmentation of the sound which divides the signal into several quasi-stationary consecutive zones (Fig. 1). The algorithm used is the "Forward-Backward Divergence" [15].
2. Next, an activity detection algorithm that removes silence and low-level signal noise that is not considered essential sound activities (Fig. 2).
3. A merging step, to bring together the areas of successive activities.
4. MFCCs Extraction (Fig. 3)

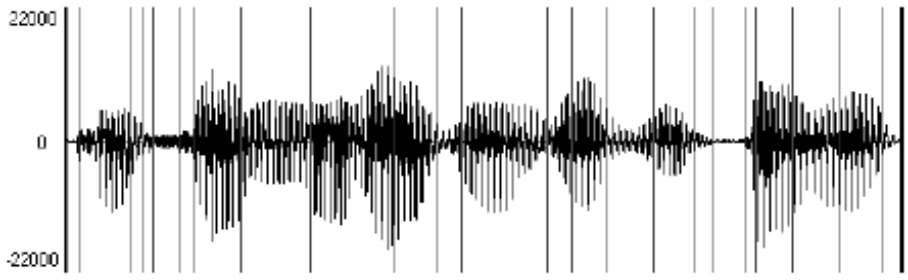


Fig. 1. Results of segmentation of a 1 second of speech

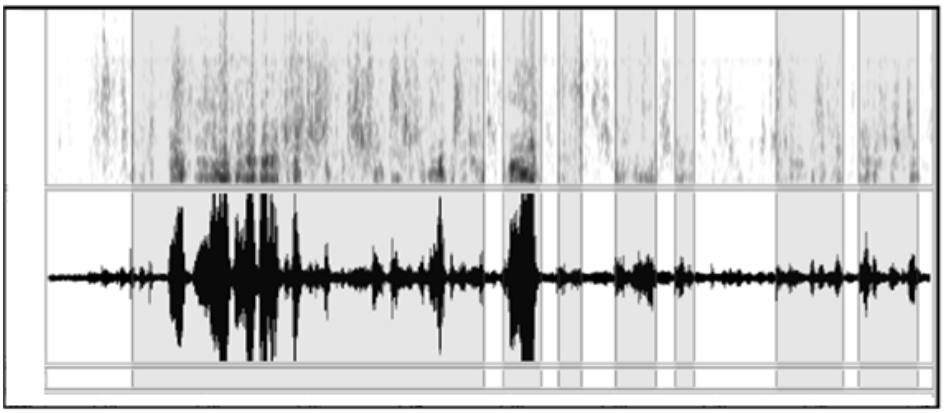


Fig. 2. Detecting audio signal activity zones (in gray)

The steps for calculating the MFCC coefficients after the step of pre-emphasis:

1. A Hamming windowing is done to limit the effects of the Gibbs phenomenon and to keep continuity of the signal.
2. To transform the signal time domain into frequency, we use Fast Fourier Transform FFT method.
3. Mel triangular filters spaced according to the Mel scale in order to get smooth magnitude spectrum.
4. Take the Log of the frequency.
5. Transformed into Discrete Cosine (DCT).
6. Add the log energy and perform delta operation (first and second derivatives).

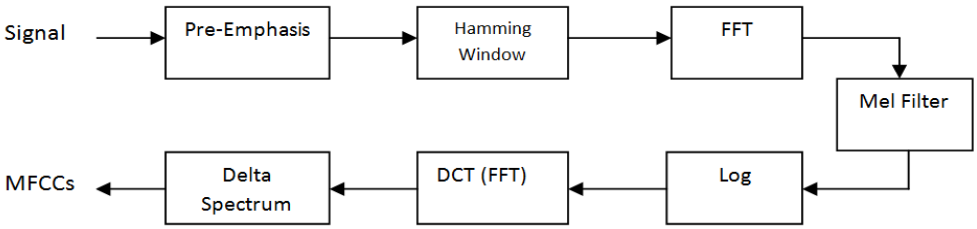


Fig. 3. MFCC feature extraction

1.2 The proposed approach

As noted before, the main issue with supervised classification is the availability of labeled datasets and that many external and internal factors affects the content of the speech signal making the use of a trained classifier from previous datasets inefficient and the need for an automated classification in real applications a necessity.

Fig. 4 describes the proposed approach for audio event detection. The audio signal is analyzed to get the 39 MFCC coefficients, then the KPCA is applied to reduce the dimension from 39 to few components (2 or 3) that capture the sufficient information to detect event in the audio signal. The KPCA components are input to the clustering algorithm; in our experiment we use the K-means algorithm for its simplicity.

In the following, we present the kernel PCA which will be used for the first time in the context of audio event detection. We will also present the kernel k-means algorithm for clustering. The results of the proposed method are compared to some state -of-art supervised methods. Therefore in the next section we will introduce the theoretical foundation of these supervised methods.

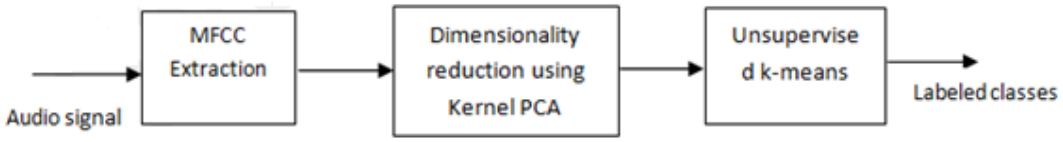


Fig. 4. Steps of the proposed detection system

1.3 Kernel principal components analysis

Kernel PCA is firstly introduced by Shölkopf and al. in [16]. It is considered an effective way to extract nonlinear features from a dataset. Kernel PCA can improve the features of the input data which make easier the separation of clusters. In [17] authors used the scatter plot of a few components of KPCA to determine the number of clusters in the data. By using a nonlinear kernel function i.e. Gaussian or polynomial, KPCA implicitly performs the familiar PCA in a high-dimensional space which is non-linearly related to the input space. Consequently, kernel PCA produces features that capture the nonlinear structure in the data better than linear PCA.

It the present article, Kernel PCA is used to reduce the dimensionality of the 39 MFCC features in order to get the new components that are essential for clustering. This will reduce the complexity of clustering methods and provide better results. As noted before the main problem with other approach in the literature of audio event detection is the data labeling step. In [9] the data were manually cross-labeled by two different audio experts. Authors in [10] used two parallel GMM classifiers to discriminate, respectively, between screams and noise, and between gunshots and noise. Each binary classifier is trained separately with the samples of the respective classes using the Figueiredo and Jain algorithm.

1.4 Kernel K-means

The classical K-means algorithm fails to give good results when the clusters are non-linearly separable. Therefore the kernel K-means has been proposed as an alternative to K-means; Kernel K-means uses a nonlinear transformation from the input space to a feature space using the kernel function. Consequently, the algorithm is able to separate non-linear clusters in the input space [18].

The algorithm minimizes the squared distances between the data points and the corresponding centers $c_k^{\emptyset}$ in the feature space:

$$MSE^{\emptyset} = \sum_1^K \sum_{x_i \in C_k} \left\| \emptyset(x_i) - m_k^{\emptyset} \right\|^2 \quad (1)$$

The Euclidian distance in the feature space is calculated using only the kernel function K:

$$\left\| \emptyset(x_i) - m_k^{\emptyset} \right\|^2 = K(x_i, x_i) - \frac{2}{|m_k|} \sum_{a \epsilon m_k} K(a, x_i) + \frac{1}{|m_k|^2} \sum_{a \epsilon m_k} \sum_{b \epsilon m_k} K(a, b) \quad (2)$$

Nevertheless, two parameters have to be chosen a priori, the parameter of the Gaussian kernel and K the number of clusters. In the following, we present the Kernel Davies & Bouldin internal index as a way to choose the optimal values of the two parameters.

1.5 Kernel Davies & Bouldin index

The non-linear version of the Davies & Bouldin (DB) validity index was first introduced by Nasser et al. [19]. The index is used to evaluate the quality of clustering algorithms by plotting the values of the index against the number of clusters; a minimal value indicates the optimal number of clusters within the dataset. The Kernel DB index algorithm is the following:

$$\text{kernel DB} = \max_{k \neq i} \frac{S_k^{\emptyset} + S_i^{\emptyset}}{d_{ik}^{\emptyset}} \quad (3)$$

The interclass dispersion S_k^{ϕ} in the feature space is defined by:

$$S_k^{\phi} = \frac{1}{|m_k|} \sum_{x \in m_k} \left\| \phi(x) - m_k^{\phi} \right\|^2 \tag{4}$$

The Euclidean distance d_{ik}^{ϕ} between clusters' centers m_i^{ϕ} and m_k^{ϕ} in the feature space is defined by:

$$d_{ik}^{\phi} = \left\| m_i^{\phi} - m_k^{\phi} \right\|^2 \tag{5}$$

For a Gaussian kernel interclass dispersions S_k^{ϕ} and d_{ik} distances between centers of clusters are defined by the following equations:

$$S_k^{\phi} = 1 - \frac{1}{|m_k|^2} \sum_{x \in C_k} \sum_{a \in C_k} k(a, x) \tag{6}$$

$$d_{ik}^{\phi} = \frac{1}{|m_i|^2} \sum_{a \in m_i} \sum_{c \in m_i} k(a,c) - \frac{2}{|m_i||m_k|} \sum_{a \in m_i} \sum_{b \in m_k} k(a,b) + \frac{1}{|m_k|^2} \sum_{b \in m_j} \sum_{d \in m_j} k(b,d) \tag{7}$$

1.6 The classifiers

Multi-layer Perceptron: A multi-layer perceptron (MLP) is a supervised classifier that uses backpropagation to learn to classify a dataset. The architecture of the network includes three layers, in addition to input and output layer, one or more hidden layers are included in between. Each layer is composed of several neurons or unit which performs a specific weighted transformation of its inputs providing the outputs of the previous layer. The training process of the network consists of adjusting the weight of each unit to arrive at the optimal solution. At the end of the process, the output layer is capable to choose which class the input belongs to. The MLP has been used in the domain of speech detection; in [20] a MLP is trained using modulation spectral features, compared to MFCC features. The speech features are input to the trained MLP to estimate phoneme posterior probabilities which they are merged into one speech class to derive speech/non-speech decisions.

SVM classifier: Support Vector Machine (SVM) is a supervised classifier that defines a hyperplane which divides the dataset into parts. Several algorithms are used to train the classifiers of which the sequential minimal optimization algorithm (SMO) has been shown to be an effective method. The SVM classifier has been widely used in the area of audio event and speech detection [11,13].

Random Forest: Random forests are a type of ensemble learning method for supervised classification. The Random forests classifier creates a set of decision trees from a randomly selected subset of the training set. The class of an object is decided by pooling together votes from different decision tree to come up with a final decision. Thambi et al. used the random forest to improve the performance of speech/non-speech detection [21].

Bayesian network: Bayesian networks are a type of graphical model that uses probability relationships to model conditional dependence. Giannakopoulos and al. used a multi-class classification algorithm for audio segments recorded from movies to detect violent content. A Bayesian network is used to classify the audio segments into six classes. Experiments showed good result as a multi-class classification scheme as well as a binary classifier for the problem of violent non-violent audio content [22].

Deep Neural Network (DNN): A DNN is a feed-forward neural network; it is deep because it is formed of several hidden layers of elementary units. Each layer uses the outputs of the previous layer as input allowing the network to operate as a chain with complicated non-linear transformations of the input layer. The units of each layer represent unknown features which explain the data allowing for various level of abstraction and enabling the output layer to discriminate more efficiently the dataset. DNN has been successfully used in the field of automatic speech recognition [9,23].

Logistic regression: Logistic regression classifier is a simple supervised algorithm that uses logistic function whose coefficients are estimated while training using the training dataset; it then predicts the probability of test samples for each target category. When used for multi-class classification problem it assigns a category with the highest probability to this test instance. Logistic regression is widely used in various fields including natural language processing, audio event detection [24].

Decision tree: A Decision tree is a flowchart-like tree structure where each internal node denotes a test on an attribute, each branch represents an outcome of the test and each leaf node holds a class label. In our experiment we used the C4.5 algorithm developed by Ross Quinlan to generate a decision tree. Decision trees have been used in the field of signal processing and automatic audio information retrieval and classification [25].

K-means clustering: K-means is an unsupervised clustering algorithm that partitions the dataset into k clusters where k is specified a priori [26]. By minimizing the distance between the center of the class and the class members, the K-means algorithm classifies objects in groups such that objects within the same cluster are as similar as possible whereas objects from different clusters are as dissimilar as possible. Although the simplicity of the algorithm, results are sensitive to initialization and the number of clusters should be specified in advance. Regrettably, no theoretical method exists to find the optimal number of clusters. A practical approach is to compare the outcomes of multiple runs with different k and choose the best one based on a predefined criterion [26]. Several approaches have been proposed to choose the optimal value of K, in [19] we proposed to use the kernel Davies & Bouldin index to determine the optimal k value. In [27] we compare several internal validity indices to determine the optimal value of K clusters.

EM clustering: Expectation–Maximization or EM algorithm derived from the GMM model. It assumes that each group of data has a specific distribution i.e. Gaussian. EM algorithm tries to find the parameters of the distributions by solving an optimization algorithm using two steps; The E-step which computes the probability that each data point belongs to a particular cluster evaluated using the current estimate for the parameters and the M–step which maximizes the probability of data point within the cluster by computing a weighted sum of data point within each cluster.

2 Experiments

The audio signals used in our experiments are recorded in public transportation by 4 microphones. The audio signal contains three classes: speech, noise and spray bomb. The total number of MFCC coefficients extracted is 39 formed by the first 12 MFCC coefficients plus energy, supplemented by their first and second derivatives. For a window of 16ms with an overlap of 8ms, an observation vector is obtained every 8 ms.

2.1 First audio signal with two classes

This dataset composed of speech and noise that we seek to classify. In this first signal, we try to analyze the content of the signal and the effectiveness of the unsupervised clustering using different framing windows. Fig. 5 shows the plots of the first two MFCC features using 8ms and 40 ms windows. Note that obviously when the width of the window increases, the sample data will decrease. In Fig. 6, we show the effect of the sampling window used to extract the MFCC attributes on clustering accuracy using K-means.

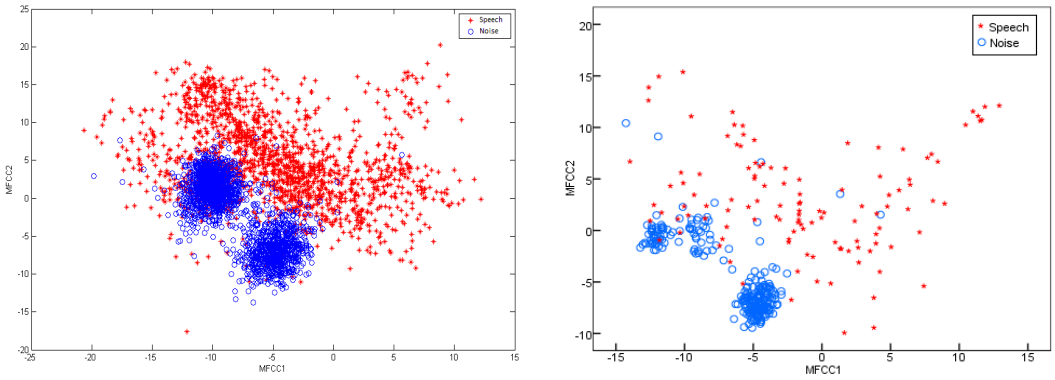
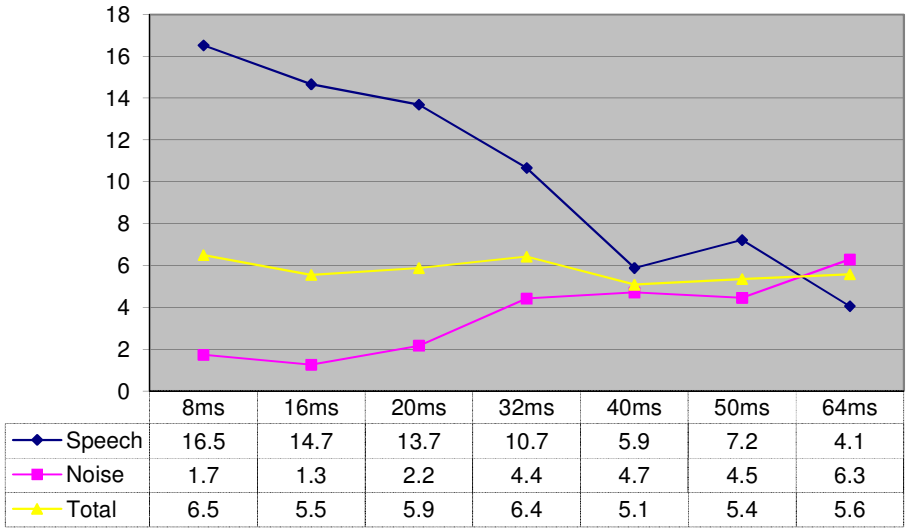


Fig. 5. Plot of the first and second MFCC with 8ms (left) and 40ms (right)

Fig. 6 shows the K-means classification error rates for the different sampling periods. Although this rate is minimal for MFCC coefficients sampled at 40ms, for the second audio signal we will use 16ms as it gives good results and the sampled data is much larger.



	8ms	16ms	20ms	32ms	40ms	50ms	64ms
Speech	16.5	14.7	13.7	10.7	5.9	7.2	4.1
Noise	1.7	1.3	2.2	4.4	4.7	4.5	6.3
Total	6.5	5.5	5.9	6.4	5.1	5.4	5.6

Fig. 6. K-means classification error

Table 1 gives the confusion matrix of the best result for a sampling period of 40ms. We note that 10.7% of the points are misclassified.

Table 1. Confusion matrix of K-means using window of 40ms

Results ->	**Speech**	**Noise**
Speech	94,1%	5,9%
Noise	4,8%	95,2%

2.2 Second audio signal with 3 classes

The signal we are using here has three classes, Speech, Noise and Spray Bomb. Fig. 7 shows the data plot on the 1st and 2nd MFCC features where we can distinguish the three classes: Speech (blue), Noise (red) and Spray bomb (green).

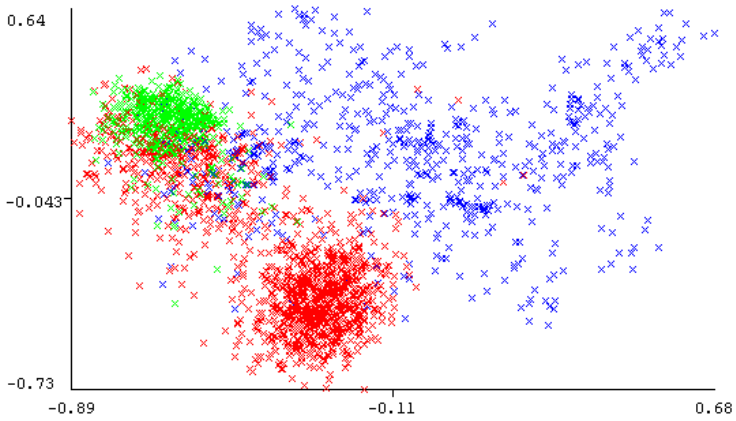


Fig. 7. Scatter plot on the 1st and 2nd MFCC

Fig. 8 depicts the variances of the 39 MFCC features. It shows that the first 12 MFCC features capture most of information.

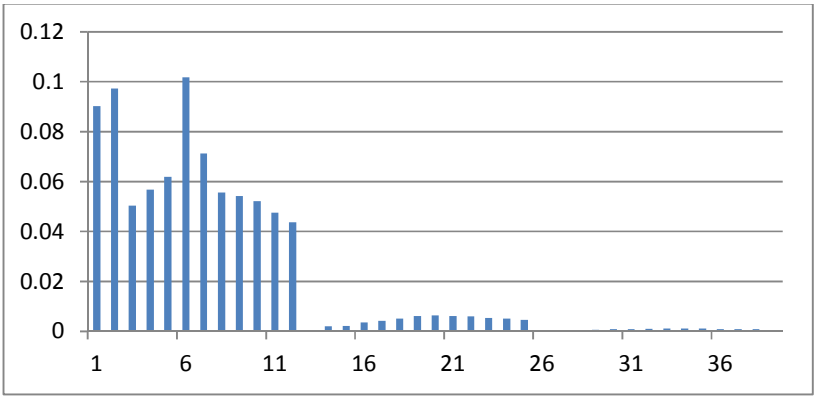


Fig. 8. Variance of MFCC features

3 Results of the Classification

The purpose of an automated detection system is to trigger an alarm once an unusual event happens in a public place.

For the experiments we used the open source software Weka from the University of Waikato New Zealand for all classification methods except for DNN we used the open source R programming using the H2O package, an open source machine learning platform that offers parallelized implementations of many supervised and unsupervised machine learning algorithms such as Deep Neural Networks (Deep Learning), K-Means, PCA and others.
For reliability measures, we run the algorithms several times with different seed values in order to test the sensitivity of the algorithms to initial starts and guarantee consistent results. We present in the following the confusion matrix of the result of each algorithm in the tables below:

1- Multi-layer perceptron (Supervised classification)

Table 2 showed that 1.32% of Speech and 0.16% of noise trigger incorrectly the alarm, whereas 7.39% of spray bomb missed the alarm.

Table 2. 1 hidden layer with 20 neurons, the Incorrectly Classification rate is 3.93%

Results->	Speech	Noise	Spray bomb
Speech	573(94.71%)	24(3.97%)	8(1.32%)
Noise	24(1.89%)	1247(97.96%)	2(0.16%)
Spray bomb	11(2.39%)	23(5%)	426(92.61%)

2- SVM with RBF Function (Supervised classification)

Table 3 showed that 0.5% of Speech and 0% of noise are classified as false alarm, whereas 9.78% of spray bomb missed the alarm.

Table 3. The number of support vectors used 734, the incorrectly Classification rate is 9.70 %

Results->	Speech	Noise	Spray bomb
Speech	428(70.74%)	174(28.76%)	3(0.5%)
Noise	5(0.39%)	1268(99.61%)	0(0%)
Spray bomb	0(0%)	45(9.78%)	415(90.22%)

3- Random Forest (supervised classification)

Table 4 showed that 0.17% of Speech and 0.08% of noise are classified as false alarm, whereas 7.17% of spray bomb missed the alarm.

Table 4. The incorrectly classification rate is 3.20%

Results->	Speech	Noise	Spray bomb
Speech	587(97.02%)	17(2.81%)	1(0.17%)
Noise	23(1.81%)	1249(98.19%)	1(0.08%)
Spray bomb	22(4.78%)	11(2.39%)	427(92.83%)

4- Bayesian Network Classifier (supervised classification)

Table 5 showed that 0.83% of Speech and 0% of noise are classified as false alarm, whereas 11.73% of spray bomb missed the alarm.

Table 5. The incorrectly classification rate is 6.75%

Results->	Speech	Noise	Spray bomb
Speech	555(91.74%)	45(7.44%)	5(0.83%)
Noise	54(4.24%)	1219(95.76%)	0(0.0%)
Spray bomb	29(6.3%)	25(5.43%)	406(88.26%)

5- DNN (supervised classification)

Table 6 showed that 0.47% of Speech and 4.13% of noise are classified as false alarm, whereas 0% of spray bomb missed the alarm.

Table 6. The architecture of the network is composed of 3 hidden layer with 100, 50, 10 neurons respectively. The incorrectly classification rate is 1.11%

Results->	Speech	Noise	Spray bomb
Speech	1267(99.53%)	0(0%)	6(0.47%)
Noise	1(0.22%)	440(95.65%)	19(4.13%)
Spray bomb	0(0%)	0(0%)	605(100%)

6- Logistic regression

Table 7 showed that 1.65% of Speech and 0.39% of Noises are classified as false alarm, whereas 7.2% of spray bomb missed the alarm.

Table 7. Logistic regression results with 4.53% of incorrectly classified instances

Results->	Speech	Noise	Spray bomb
Speech	559(92.4%)	36(5.95%)	10(1.65%)
Noise	22(1.73%)	1246(97.88%)	5(0.39%)
Spray bomb	5(1.09%)	28(6.09%)	427(92.83%)

7- Decision tree

Table 8 showed that 2.98% of Speech and 0.55% of Noises are classified as false alarm, whereas 5.2% of spray bomb missed the alarm.

Table 8. Decision tree results with 5.26% of incorrectly classified instances

Results->	Speech	Noise	Spray bomb
Speech	555(91.74%)	32(5.29%)	18(2.98%)
Noise	42(3.30%)	1224(96.15%)	7(0.55%)
Spray bomb	16(3.48%)	8(1.74%)	436(94.78%)

8- K-means confusion matrix (Unsupervised classification)

Table 9 showed that 24.63% of Speech and 34.72% of Noises are classified as false alarm, whereas 1.09% of spray bomb missed the alarm.

Table 9. The incorrectly Classification rate is 28.18% with a rate of false alarm of 31.5% and 1.09% missing alarm

Results->	Speech	Noise	Spray bomb
Speech	406(67.11%)	50(8.26%)	149(24.63%)
Noise	13(1.02%)	818(64.26%)	442(34.72%)
Spray bomb	0(0%)	5(1.09%)	455(98.91%)

9- EM confusion matrix (Unsupervised classification)

Table 10 showed that 0% of Speech and 0% of Noises are classified as false alarm, whereas 13.3% of spray bomb missed the alarm.

Table 10. The incorrectly classification rate is 23.09%

Results->	Speech	Noise	Spray bomb
Speech	605(100%)	0(0%)	0(0%)
Noise	479(37.63%)	794(62.37%)	0(0%)
Spray bomb	60(13.04%)	1(0.22%)	399(86.74%)

Comparing supervised classification methods, the DNN algorithm gave the best result with 0% missed alarm and 1.4% fault alarm. In the other hand, K-means is able to differentiate spray bomb with a rate of 98.91% thus 1.09% missing alarm although a high rate of fault alarm.

Classification using K-means on reduced data and kernel k-means

Now we will apply KPCA for dimensionality reduction technique to the 39 MFCC. The variances of the first 39 KPCA components are presented in Fig. 9. It shows that the first 13 components capture 87.8% of the

information about the 39 components. For the calculation of the kernel function, we used a Gaussian kernel with width σ equals 11.26 found by using the approximated Parzen window given by [28]:

$$\sigma_{AMISE} = \hat{\sigma}.\left[\frac{4}{(2D+1)N}\right]^{\frac{1}{D+4}} \tag{8}$$

$\hat{\sigma}^2$ is the trace of the covariance matrix of the dataset.

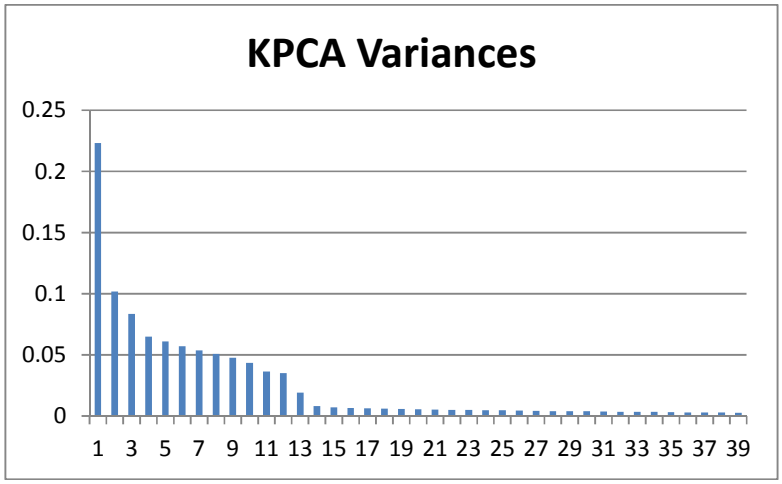


Fig. 9. Variances of KPCA components

The K-means method is applied on the 39 and 2 KPCA components. The numerical value of K is three since we want to differentiate between normal speech, noise and an unusual event like shouting, spray bomb or screaming. Nevertheless, to check the efficiency of the Kernel DB index, we calculated the index for several values of σ and K decided a priori where $K \in \{2,\cdots;8\}$ and $\sigma \in [3.32\,;\,59.68]$. The values 3.32 and 59.68 are the minimal and maximal distances between points in the initial space, we then used 60 values of σ equally spaced in that interval.

Kernel K-means is computed 10 times for each pair of parameters (K,σ) to avoid the problem of initialization of the cluster centers. Note that, only the minimal values of kernel DB are kept. Fig. 10 "The effect of parameters on the clustering results" shows that the best value of k and sigma are 3 and 20 respectively.

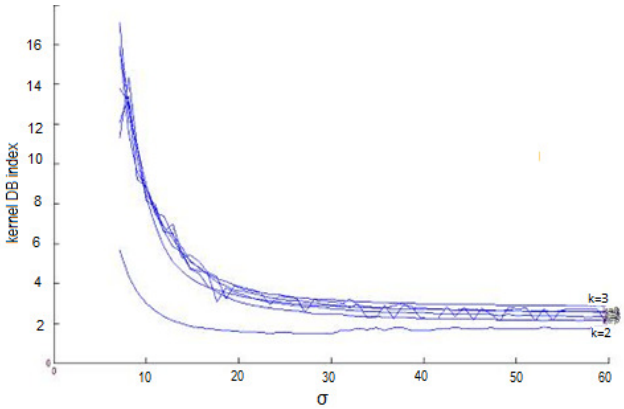


Fig. 10. Effect of parameters on the clustering results

Table 11. Kmeans + 39 KPCA, the incorrectly classification rate is 24.41%

Results->	Speech	Noise	Spray bomb
Speech	445(37.96%)	827(64.96%)	1(0.08%)
Noise	529(87.44%)	73(12.07%)	3(0.5%)
Spray bomb	43(9.35%)	6(1.3%)	411(89.35%)

Table 12 showed that 4.3% of Speech and 4.63% of Noises are classified as false alarm, whereas 7.8% of spray bomb missed the alarm.

Table 12. Kmeans + 2 KPCA, the incorrectly classification rate is 12.44%

Results->	Speech	Noise	Spray bomb
Speech	424(70.08%)	155(25.62%)	26(4.3%)
Noise	15(1.18%)	1199(94.19%)	59(4.63%)
Spray bomb	0(0.0%)	36(7.83%)	424(92.17%)

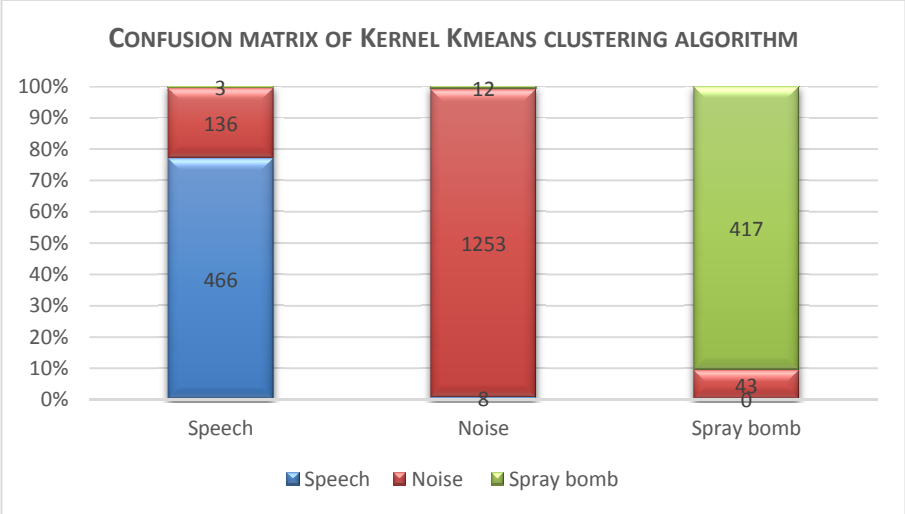


Fig. 11. Confusion matrix of Kernel Kmeans clustering algorithm

Tables 11 and 12 show the confusion matrices obtained by K-means method applied to the 39 and 2 KPCA components respectively.

The MFCC features after dimensionality reduction are classified into three classes: spray bomb, speech, and noise. Noise is often confused with speech class but infrequently affect with the bomb class.

Table 11 showed that 0.08% of Speech and 0.5% of Noises are classified as false alarm, whereas 10.7% of spray bomb missed the alarm.

Now for comparison purpose, we applied kernel k-means to the MFCC features. The confusion matrix is given in Table 13. 0.5% of speech is categorized as bomb which triggers a false alarm and 0.94% of noise are categorized a spray bomb triggering a false alarm. Whereas, 9.35% of spray bomb are classified as Noise.

Table 13. The incorrectly classification rate is 8.63%

Results->	Speech	Noise	Spray bomb
Speech	466(77.02%)	136(22.48%)	3(0.5%)
Noise	8 (0.63%)	1253(98.43%)	12(0.94%)
Spray bomb	0(0%)	43(9.35%)	417(90.65%)

3.1 Evaluation of the different methods

Table 14 summarizes the result of the classification methods presented in this paper. Although supervised classification methods give better results than unsupervised classification methods, whereas our proposed approach along with the Kernel k-means provides competitive results compared to other supervised methods. The percentage of false alarm (False Positive rate) was relatively comparable to the supervised classification. The obtained results recommend the adoption of the unsupervised classification in real-time detection system.

Table 14. Summary of classification results

	Classifier	True positive rate	False positive rate (Fault alarm)	True negative rate	False negative rate (Missed alarm)	Incorrectly classified (%)
Supervised	MLP	92.6%	0.5%	99.5%	7.4%	3.93
	SVM	90.2%	0.2%	99.8%	9.8%	9.7
	R.F	92.8%	0.1%	99.9%	7.2%	3.2
	Bayes	88.3%	0.3%	99.7%	11.7%	6.75
	DNN	**100.0%**	**1.4%**	**98.6%**	**0.0%**	**1.11**
	Logistic regression	92.8%	0.8%	99.2%	7.2%	4.53
	Decision tree	94.8%	1.3%	98.7%	5.2%	5.26
Unsupervised	Kmeans	98.9%	31.5%	68.5%	1.1%	28.18
	EM	86.7%	0.0%	100.0%	13.3%	23.09
	kmeans+ 39 KPCA	89.3%	0.2%	99.8%	10.7%	24.41
	Kmeans + 2 KPCA	92.2%	4.5%	95.5%	7.8%	12.44
	Kernel Kmeans	90.7%	0.8%	99.2%	9.3%	8.63

4 Conclusion

We proposed in this paper a new approach for automatic detection system based on unsupervised classification and presented a comparative study of several supervised and unsupervised classification methods for audio event detection. The motivation of this work is to secure the public transportation and places by triggering an alarm whenever an unusual audio event happens in those places. The MFCC feature extraction is used to transform the audio signal into discrete features using 39 features, then kernel PCA is used to reduce the 39 dimensions into fewer dimensions which capture 87.8% of the information. After that, the k-means clustering algorithm is then applied to those KPCA components and compared to Kernel K-means. K-means on two KPCA components gave good results for triggering a true alarm as well as for the fault alarm where the percentage of misclassification is 12.44%. The percentage of misclassification is 8.63% for kernel k-means. Note that results for KPCA are quite similar to kernel K-means. In the other hand, DNN network gave the best results compared to other supervised methods presented in this article. We conclude that our proposed method is reliable for audio event detection in real applications.

5 Future Work

The proposed approach of using unsupervised classification and KPCA will be subject to more investigation on the future using real dataset from publicly database and hopefully with data recorded in the Lebanese roads for security purposes. Our goal is to build an unsupervised application for the early detection of abnormal events in public places in the hope of reducing violence and crime.

Competing Interests

Author has declared that no competing interests exist.

References

[1] Petridis S, Stafylakis T, Ma P, Cai F, Tzimiropoulos G, Pantic M. End-to-end audiovisual speech recognition. CoRR, abs/1802.06424; 2018.

[2] Cristani M, Bicego M, Murino V. Audio-visual event recognition in surveillance video sequences. IEEE Trans. Multimedia. 2007;9(2):257-67.

[3] Afouras T, Chung J, Senior A, Vinyals O, Zisserman A. Deep audio-visual speech recognition, arXiv:1809.02108v2 [cs.CV]; 2018.

[4] Schindler A, Boyer M, Lindley A, Schreiber D, Philipp T. Large scale audio-visual video analytics platform for forensic investigations of terroristic attacks. Multimedia Modeling. Springer; 2019.
Available:https://doi.org/10.1007/978-3-030-05716-9_9

[5] Rabiner LR, Juang B. Fundamentals on speech recognition. New Jersey: Prentice Hall; 1996.

[6] Almaadeed N, Asim M, Al-Maadeed S, Bouridane A, Beghdadi A. Automatic detection and classification of audio events for road surveillance applications. 2018;18(6):1858.
DOI: 10.3390/s18061858

[7] Roneel V. Sharan, Tom J. Moir. An overview of applications and advancements in automatic sound recognition. Neurocomputing. 2016;22-34.
Available:https://doi.org/10.1016/j.neucom.2016.03.020

[8] Vacher M, Istrate D, Besacier L, Serignat JF, Castelli E. Sound detection and classification for medical telesurvey. ACTA Press, Calgary. 2nd Conference on Biomedical Engineering. Innsbrück, Austria. 2004;395-398.

[9] Pierre Laffitte P, Yun Wang, David Sodoyer A, Laurent Girin. Assessing the performances of different neural network architectures for the detection of screams and shouts in public transportation, expert systems with applications. Elsevier. 2019;117:29–41.
Available:https://doi.org/10.1016/j.eswa.2018.08.052 0957-4174/

[10] Valenzise G, Gerosa L, Tagliasacchi M, Antonacci F, Sarti A. Scream and gunshot detection and localization for audio-surveillance systems. In Proceeding the IEEE AVSS. 2007;21-26.
DOI: 10.1109/AVSS.2007.4425280

[11] Rouas JL, Louradour J, Ambellouis S. Audio events detection in public transport vehicule. IEEE. 2006;733-738.

[12] Ntalampiras S, Potamitis I, Fakotakis N. An adaptive framework for acoustic monitoring of potential hazards. EURASIPJ. Audio Speech Music Process. 2009;13:1-13.

[13] Pasquale Foggia, Nicolai Petkov, Alessia Saggese, Nicola Strisciuglio, Mario Vento. IAPR Fellow "Reliable detection of audio events in highly noisy environments". Pattern Recognition Letters. Elsevier. 2015; 65:22-28.
Available:http://dx.doi.org/10.1016/j.patrec.2015.06.026

[14] Nasser A, Hamad D, Jean-Luc Rouas J, Ambellouis S. The use of kernel methods for audio events detection. IEEE; 2008.
DOI: 10.1109/ICTTA.2008.4529996

[15] André-Obrecht R. A new statistical approach for automatic speech segmentation. IEEE Transactions on Acoustics, Speech and Signal Processing. 1988;36(1):29-40.

[16] Shölkopf B, Smola AJ. Learning with kernels: Support vector machines, regularization, optimization and beyond. The MIT Press, Cambridge, Massachusetts, London, England; 2002.

[17] Nasser A, Hamad D, Nasr C. Kernel PCA as a visualization tools for clusters identifications. In Lecture Notes in Computer Science. Springer, Berlin, Heidelberg. 2006;4132.
Available:https://doi.org/10.1007/11840930_33

[18] MacQueen JB. Some methods for classification and analysis of multivariate observations, proceedings of 5th Berkeley symposium on mathematical statistics and probability. Berkeley, University of California Press. 1967;1:281-297.

[19] Nasser A, Hébert PA, Hamad D. Clustering evaluation in feature space. In: de Sá JM, Alexandre LA, Duch W, Mandic D, (Eds) Artificial Neural Networks – ICANN 2007. ICANN 2007. Lecture Notes in Computer Science. Springer, Berlin, Heidelberg. 2007;4669.
Available:https://doi.org/10.1007/978-3-540-74695-9_33

[20] Ganapathy S, Rajan P, Hermansky H. Multi-layer perception based speech activity detection for speaker verification, Published in: 2011 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA). IEEE; 2011.
DOI: 10.1109/ASPAA.2011.6082323

[21] Sincy V. Thambi, Sreekumar KT, Santhosh Kumar C, Reghu Raj PC. Random forest algorithm for improving the performance of speech/non-speech detection, Published in: First International Conference on Computational Systems and Communications (ICCSC); 2014.
DOI: 10.1109/COMPSC.2014.7032615

[22] Giannakopoulo T, Pikrakis A, Theodoridis S. A multi-class audio classification method with respect to violent content in movies using Bayesian networks. IEEE 9th Workshop on Multimedia Signal Processing; 2007.
DOI: 10.1109/MMSP.2007.4412825

[23] Baby D, Gemmeke JF, Virtanen T, Van Hamme V. Exemplar-based speech enhancement for deep neural network based automatic speech recognition, 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); 2015.
DOI: 10.1109/ICASSP.2015.7178819

[24] Bahuleyan H. Music genre classification using machine learning techniques.
Available:https://arxiv.org/pdf/1804.01149.pdf

[25] Lin KZ, Pwint M. Structuring sport video through audio event classification. In: PCM 2010, Part I, LNCS 6297. Springer. 2010;481–492.

[26] Shawe-Taylor J, Cristianini N. Kernel methods for patter analysis. Cambridge University Press; 2004.

[27] Nasser A. Investigating k-means and kernel k-means algorithms with internal validity indices for cluster identification. Journal of Advances in Mathematics and Computer Science; 2019.
Available:https://doi.org/10.9734/JAMCS/2019/45837

[28] Jenssen R. An information theoretic approach to machine learning. University of Tromso, Thesis; 2005.

Peer-review history:
The peer review history for this paper can be accessed here (Please copy paste the total link in your browser address bar)
http://www.sdiarticle4.com/review-history/54824