

Probabilistic Detection Methods for Acoustic Surveillance Using Audio Histograms

M. S. Shankar Reddy · Karan Nathwani ·
Rajesh M. Hegde

Received: 16 April 2014 / Revised: 5 November 2014 / Accepted: 6 November 2014 /
Published online: 2 December 2014
© Springer Science+Business Media New York 2014

Abstract Acoustic surveillance is gaining importance given the pervasive nature of multimedia sensors being deployed in all environments. In this paper, novel probabilistic detection methods using audio histograms are proposed for acoustic event detection in a multimedia surveillance environment. The proposed detection methods use audio histograms to classify events in a well-defined acoustic space. The proposed methods belong to the category of novelty detection methods, since audio data corresponding to the event is not used in the training process. These methods hence alleviate the problem of collecting large amount of audio data for training statistical models. These methods are also computationally efficient since a conventional audio feature set like the Mel frequency cepstral coefficients in tandem with audio histograms are used to perform acoustic event detection. Experiments on acoustic event detection are conducted on the SUSAS database available from Linguistic data consortium. The performance is measured in terms of false detection rate and true detection rate. Receiver operating characteristics curves are obtained for the proposed probabilistic detection methods to evaluate their performance. The proposed probabilistic detection methods perform significantly better than the acoustic event detection methods available in literature. A cell phone-based alert system for an assisted living environment is also discussed as a future scope of the proposed method. The performance evaluation is presented as number of successful cell phone transactions. The results are motivating enough for the system to be used in practice.

M. S. S. Reddy · K. Nathwani (✉) · R. M. Hegde
Indian Institute of Technology Kanpur, Kanpur, India
e-mail: nathwani@iitk.ac.in

M. S. S. Reddy
e-mail: s.shankar437@gmail.com

R. M. Hegde
e-mail: rhegde@iitk.ac.in

Keywords Probabilistic detection · Acoustic surveillance · Event detection · Audio histogram

1 Introduction

Automatic surveillance systems should be focus on detecting acoustic events. In present day research, most of the automatic surveillance systems use large number of cameras in an area where surveillance is required. It has been observed that installing cameras in all the places (where automatic surveillance is required) leads to privacy problems. For example, installing cameras in each room of a home or a bank leads to privacy issues. Therefore, in such scenarios there is a need for alternative surveillance technologies like acoustic surveillance which is based on speech signals. Speech signals are the evidences for a situation analysis. Most of the times, human beings express their emotions like happiness, sadness, anger, panic, shock, and surprising events in terms of different forms of speech. Some of the abnormal situations like people fighting and shouting generates distressed speech signals. Hence, most of the acoustic events in human presence can be detected from the speech signals. In this work, the event detection for acoustic surveillance with focus mainly on two broad categories of speech namely, neutral and distressed speech.

The surveillance of events in general can be categorized into acoustic, video, and multimedia surveillance. The video surveillance discussed in [22], can be used to detect and monitor the large scale video content in a real world event. In [1], the intelligent multimedia surveillance is designed to analyses multiple inputs such as video and audio data. These data are used to detect and track multiple events such as people, vehicles, and other objects. In [2], the audio-based event detection has been used for multi media surveillance application. The system designed in [2] utilizes various heterogeneous sensors including video and audio for detecting human's coughing in office environment [9]. Similar work on impulsive sound detection like gun shots [4] can be found. In this work, surveillance of events is limited to acoustic events such as detection of normal and distressed speech events. In this context, it is important to note that the detection of distressed speech events is critical for acoustic surveillance in an assisted living environment. Production of distressed speech signals can be considered as acoustic events in such an environment. These acoustic events generally occur in situations that need human intervention in the assisted living environment and need to be detected automatically. Widely used acoustic event detection methods in the literature are one-class support vector machine (SVM) [5,19], universal Gaussian mixture model (UGMM) [18], and Gaussian clustering [17]. These methods perform well when one uses concatenated feature vectors of high dimension. In acoustic surveillance under real-world conditions, it is not feasible to obtain large varieties of abnormal data [5,17]. This is the primary motivation for developing an audio detection method which does not use distressed data for training purpose.

In this paper, novel probabilistic detection methods using audio histograms are proposed for acoustic event detection in a multimedia surveillance environment. The proposed detection methods use audio histograms to classify events in a well-defined acoustic space. The proposed methods belong to the category of novelty detection

methods, since audio data corresponding to the event is not used in the training process. These methods hence alleviate the problem of collecting large amount of audio data for training statistical models. These methods are also computationally efficient since a conventional audio feature set like the Mel frequency cepstral coefficients in tandem with audio histograms are used to perform acoustic event detection. Experiments on acoustic event detection are conducted on the SUSAS database available from Linguistic data consortium. The performance is measured in terms of false detection rate and true detection rate. Receiver operating characteristics (ROC) curves are obtained for the proposed probabilistic detection methods to evaluate their performance. A cell phone-based alert system for an assisted living environment is also developed using the proposed methods.

The rest of the paper is organized as follows. Section 2 deals with the categorization of situations based on speech signals. Probabilistic methods for acoustic event detection using audio histograms is presented in Sect. 3. In Sect. 4, the performance evaluation of proposed method is discussed. A brief conclusion and future scope is described in Sect. 5.

2 Brief Review of Event Detection for Acoustic Surveillance

In this Section, a brief review of event detection for acoustic surveillance is discussed. In this paper, events correspond to detection of neutral and distressed speech. This Section also introduces the characteristics of neutral and distressed speech. The events resulting out of these types of speech signal can be classified into normal and distressed events which is explained as follows.

Normal events are those events which occur in a place, where the state of mind of the people around it is indifferent or unexcited. Normal events are almost always associated with people communicating in neutral speech which is occasionally loud. Distressed events are those events associated with behavior of the people, where their state of mind is either panic, frightened, shocked, or surprised. Distressed events are often associated with unusual or unforeseen happenings. Hence, the tone of speech in these events is distressed.

The variation in the characteristics of speech correspond to normal and distressed events is discussed in ensuing Section.

2.1 Characteristics of Neutral and Distressed Speech

In the context of acoustic surveillance, it is important to understand acoustic aspects of neutral and distressed speech.

Acoustics of speech deal with waveform's spectrum, amplitude, pitch, and other related physical properties.

Respiration is considered to be a sensitive indicator in certain emotional situations. In almost all stressful events, the respiration rate of human increases, this increases the glottal pressure during speech. The spectrographic analysis of neutral and distressed speech is illustrated in Fig. 1. Figure 1 shows that the distressed speech duration is less compared to neutral speech for the same speaker speaking the same word. This happens

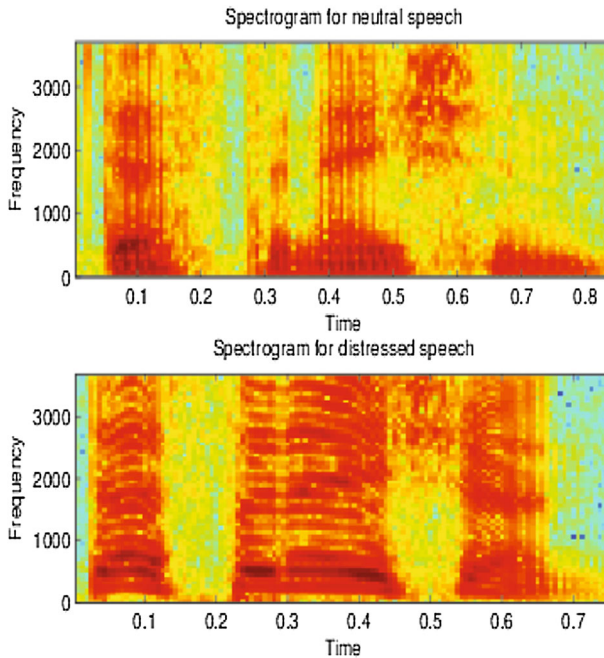


Fig. 1 Spectrogram analysis of neutral speech (*top*) and distressed speech (*bottom*) for a word “destination” spoken by same speaker

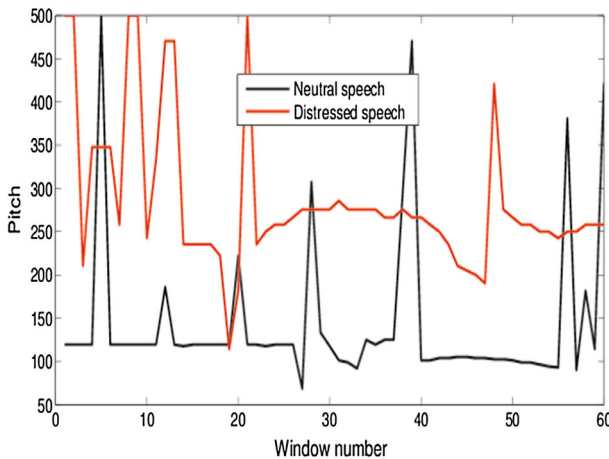


Fig. 2 Fundamental frequency (pitch) comparison for neutral and distressed speech of same word spoken by same speaker

due to the increase in respiration rate, which decreases the speech duration and effects the temporal pattern. These temporal patterns correspond to the articulation rate. The fundamental frequency (pitch) increases due to the increased glottal pressure during the voiced section [7]. This has been shown in Fig. 2. The dryness of the mouth observed during condition of excitement, fear, anger, etc., can also effect speech production

such as muscle activity of larynx and vocal cord situation. The velocity of volume through glottis is directly affected by muscle activity of the larynx and vibrating vocal cords. This in turn affects the fundamental frequency.

3 Probabilistic Methods for Acoustic Event Detection Using Audio Histograms

In this section, the proposed probabilistic method for acoustic event detection is designed using audio histogram. The motivation for using probabilistic approach for event detection is detailed in this Section. The Section also discuss the computation of audio histogram and distance measure used for acoustic event detection. The Section end with the algorithm used for detecting audio events by obtaining audio histogram.

3.1 Motivation

In acoustic event detection method [20], two GMMs are computed for discriminating screams from noise and gunshots from noise, respectively. Each classifier is trained using different features. Widely used methods in the literature, one-class support vector machine [5, 19], universal GMM [18], and Gaussian clustering [17], perform well when one uses large varieties of features concatenated together. The features can be Mel frequency cepstral coefficients (MFCCs), intonation, Teager energy operator (TEO)-based features, perceptual wavelet packet integration analysis-based features, and MPEG-7 audio protocol-based features [17]. Multi-class SVMs [15] can also be used for acoustic event detection for detecting an audio class among more than two classes. In above methods, large amount of training data is required. These methods fail to give good results when only one set of features is used.

So, a new acoustic event detection method is presented herein which uses only one set of features. An illustration of histograms of normal and distressed speech using universal GMM with 256 mixtures is shown in Fig. 3. It can be observed from Fig. 3 that the histogram of a normal speech signal differs considerably from the

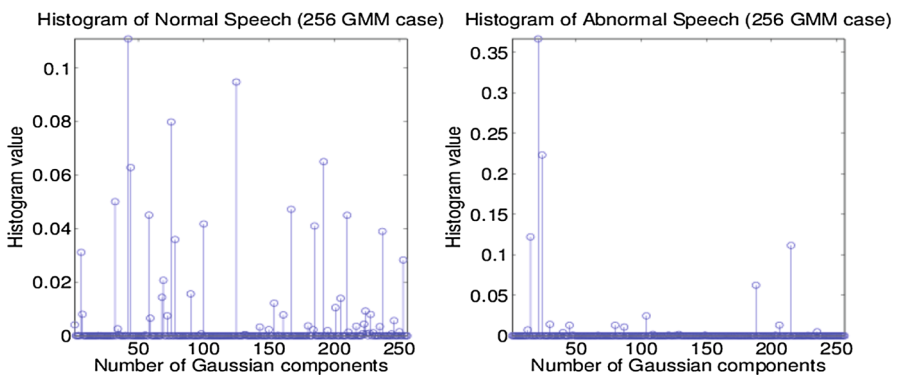


Fig. 3 Histograms of normal speech (*left*) and abnormal speech (*right*) using universal GMM with 256 modes. Both the speech signals are taken from testing data

histogram of distressed speech signal when UGMM of normal training data is used for computing the histograms. The normal speech signal has a larger number of significant histogram values than the distressed speech histogram. This fact can be utilized to detect distressed speech signal using audio histograms. The feature vectors used are the Mel frequency cepstral coefficients.

The computation of audio histograms used in probabilistic event detection is discussed in the ensuing Section.

3.2 Computation of Audio Histograms

The distribution of audio data can be characterized by histograms. In general, audio histograms are computed by splitting the data points (feature vectors) into equal-sized bins and counting number of data points in each bin [21]. The audio histograms are used in the applications of audio watermarking against cropping attacks [21, 23], environmental audio recognition [11], searching methods for audio [12], etc. In this work, an audio histogram method is used as a baseline for acoustic event detection of distressed speech signals.

In audio histogram method, the histograms for each of the training speech signal is first computed. In order to compute histograms, the feature vectors from all the training signals are first extracted. From these feature vectors, we build a universal Gaussian mixture model, shown in Eq. (1), using k-means algorithm [10] for initialization of the parameters and E-M algorithm [16] for re-estimation of GMM parameters.

$$p(x) = \sum_{i=1}^M f_i(x) = \sum_{i=1}^M w_i \mathcal{N}(x/\mu_i, \Sigma_i), \quad (1)$$

where M is the number of Gaussian components or number of modes in the GMM, x is a D -dimensional feature vector, w_i , $i = 1, \dots, M$, are the mixture weights which satisfies the condition $\sum_{i=1}^M w_i = 1$, $\mathcal{N}(x/\mu_i, \Sigma_i)$, $i = 1, \dots, M$, are the component Gaussian densities with mean vector μ_i and covariance matrix Σ_i .

Let x_k be the feature vector of k -th frame of a speech signal S_n containing N frames. The posterior probability of x_k is given by,

$$p(f_i/x_k) = \frac{f_i(x_k)}{\sum_{i=1}^M f_i(x_k)}, \quad (2)$$

where $f_i(x_k)$ is the likelihood of i -th Gaussian for x_k . The posterior probability is then calculated for all feature vectors of the speech signal S_n with respect to the i -th Gaussian component $f_i(x)$. The average of all these posterior probabilities gives the histogram value at i -th position in the histogram of S_n [11].

$$h_i(S_n) = \frac{1}{N} \sum_{k=1}^N p(f_i/x_k) \quad (3)$$

Similarly, calculating histogram values for all Gaussian components will give us the histogram representation of speech signal S_n which is shown in Eq. (4).

$$H(S_n) = [h_1(S_n) h_2(S_n) \dots h_M(S_n)] \quad (4)$$

The following Section present the distance measures used for acoustic event detection in this work.

3.3 Distance Measures Used for Acoustic Event Detection

Finding appropriate distance measure between two probability distributions is a major issue in signal-processing applications. In many cases, the probability distributions are not continuous. In this work, Jensen–Shannon divergence distance and total variation distance are used for calculating the distance between two histograms. These two measures do not need continuity condition.

3.3.1 Jensen–Shannon Divergence (JSD) Distance

Unlike Kullback divergences, the JSD does not require the absolutely continuous condition on the probability distributions [14]. Therefore, JSD is being used in cases, where the absolute continuous condition is not satisfied. The JSD is a symmetrized and smoothed version of Kullback–Leibler (K–L) divergence $D(X||Y)$ [6] and is defined as

$$JSD(X||Y) = \frac{1}{2}D(X||Z) + \frac{1}{2}D(Y||Z), \quad (5)$$

where Z is the average of two probability distributions X and Y and is given by $Z = \frac{1}{2}(X + Y)$. Histograms are equivalent to discrete probability distributions, the K–L divergence $D(X||Z)$ for discrete probability distributions $X = \{x_1, x_2, \dots, x_n\}$ and $Y = \{y_1, y_2, \dots, y_n\}$ is defined as

$$D(X||Y) = \sum_{i=1}^n x_i \log(x_i/y_i) \quad (6)$$

The JSD measure is square of a metric. In other words, \sqrt{JSD} is a metric. The measure is well characterized by non-negativity, finiteness, and boundedness properties [14]. JSD is bounded by $0 \leq JSD(X||Y) \leq \ln(2)$.

3.3.2 Total Variation Distance (TVD)

The total variation distance [13] between two discrete probability distributions, $X = \{x_1, x_2, \dots, x_n\}$ and $Y = \{y_1, y_2, \dots, y_n\}$, is given by

$$TVD(X||Y) = \frac{1}{2} \sum_{i=1}^n |x_i - y_i| \quad (7)$$

The algorithm illustrating the steps involved in acoustic event detection using audio histograms method are explained in ensuing Section.

3.4 Algorithm for Probabilistic Acoustic Event Detection Using Audio Histograms Method

Algorithm 1 Training procedure using audio histogram

1. Apply hamming window on each training speech signal to make the signal into frames.
2. Compute the MFCC vectors for each frame.
3. Build a universal M component GMM using all the MFCC vectors obtained from training speech signals with component Gaussian densities denoted by $\mathcal{N}(x/\mu_i, \Sigma_i)$. Here, μ_i and Σ_i correspond to mean vector and covariance matrix respectively for $i = 1, \dots, M$. The x is a D-dimensional feature vector.
4. Compute histograms for all training speech signals. The posterior probability is then calculated for all feature vectors of the speech signal S_n with respect to the i -th Gaussian component $f_i(x)$. The average of all these posterior probabilities gives the histogram value at i -th position in the histogram of S_n [11].

$$h_i(S_n) = \frac{1}{N} \sum_{k=1}^N p(f_i/x_k) \quad (8)$$

Similarly, calculating histogram values for all Gaussian components will give us the histogram representation of speech signal S_n which is shown in Equation (9).

$$H(S_n) = [h_1(S_n) \ h_2(S_n) \ \dots \ h_M(S_n)] \quad (9)$$

5. The universal GMM and the histograms of training speech signals represents the entire training data. This is because most of the feature vectors of normal speech signal are distributed among many Gaussian components. In other words, many Gaussian components in the universal GMM give significant likelihood values for normal speech signal than distressed speech signal as the GMM is trained with normal speech signals.
-

In this Section, an algorithmic description of the acoustic event detection method by audio histograms with JSD and TV distance measures is given. The Algorithm 1 presents with training procedure using audio histogram method.

On the other hand, the Algorithm 2 discuss the classification procedure used in the proposed method.

Algorithm 2 Classification using audio histogram

1. Apply hamming window on each training speech signal to make the signal into frames.
 2. Compute the MFCC vectors for each frame.
 3. Compute histogram for the test speech signal using the universal GMM of the training data. The histogram of a test speech signal is computed as explained in step 4 of Algorithm 1. The UGMM used herein is the universal GMM of training data.
 4. Calculate the pair-wise distance between test histogram and all training histograms one at a time. In this work, we use Jensen Shannon divergence and total variation distance as distance measures.
 5. Count how many distances are less than a predefined threshold on the distances.
 6. If this count is greater than a predefined threshold on the number of nearest neighbors, the test signal is detected as normal speech signal otherwise the test signal is detected as abnormal signal (distressed speech signal).
-

The testing procedure of AH method is shown through flowchart in Fig. 4. In flowchart, “distance_threshold” represents the threshold on the distance measure. The “neighbors_threshold” in Fig. 4 represents the threshold on nearest neighbors based on the distance.

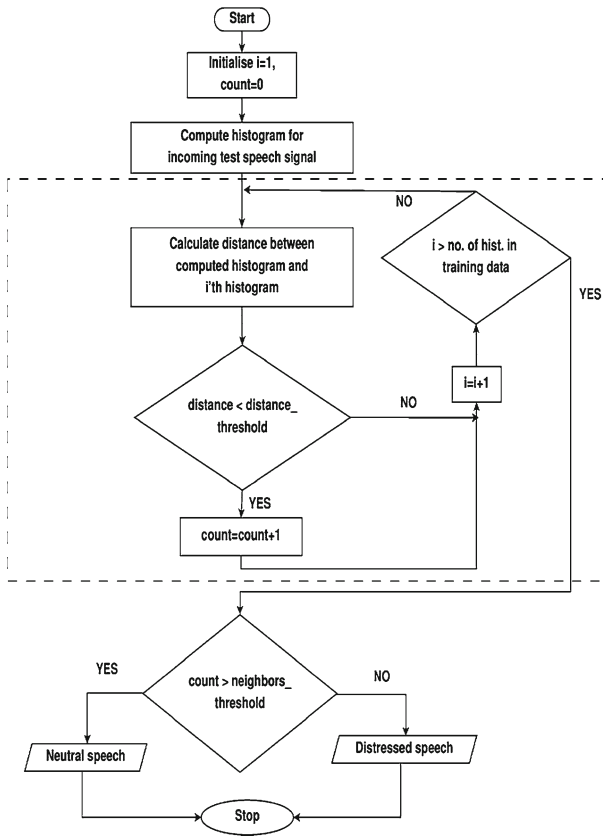


Fig. 4 Flowchart for acoustic event detection of a test speech signal using audio histogram method

4 Performance Evaluation

In this Section, the performance of the proposed method is evaluated by conducting acoustic event detection experiments using audio signals. The proposed method of acoustic event detection is also compared with one class SVM (OCSVM), universal GMM (UGMM), and Gaussian clustering (GC) methods. The database used for conducting experiments is SUSAS [8]. The experimental results are illustrated in terms of Receiver Operating Characteristic (ROC) curve.

4.1 Database Used in Performance Evaluation

One thousand normal (neutral and loud speech) examples in SUSAS [8] database have been used for training and 746 examples for testing. The testing data consist of 414 distressed (fear, anxiety, scream, and angry) and 332 normal and loud examples, which are not included in the training data. The sampling rate for both training and testing

speech signals is 8,000 samples/s. Eight male and three female speakers aged between 22 and 76 were employed in generating both training and test data.

4.2 Experimental Results

The performance of the proposed method is evaluated by conducting experiments on acoustic event detection using audio histogram and one class SVM. The experiments are presented in terms of receiver operating characteristic curve.

4.2.1 Receiver Operating Characteristic Curve

ROC curve is a plot of false detection rate (FDR) taken on x-axis vs true detection rate (TDR) taken on y-axis. The cost/benefit analysis is generally done using ROC curves. The ROC curves are also used for deciding thresholds in a particular model. In our case, FDR is defined as

$$\text{FDR} = \frac{\text{No. of neutral examples detected as distressed}}{\text{Total no. of neutral examples in the testing data}} \quad (10)$$

and TDR is defined as

$$\text{TDR} = \frac{\text{No. of distressed examples detected as distressed}}{\text{Total no. of distressed examples in the testing data}} \quad (11)$$

The detection results are obtained for four different acoustic event detection methods and compared them in a single ROC curve.

4.2.2 Experimental Results for Acoustic Event Detection Using Audio Histogram (AH) Method

As explained in Section 3.2, during training, the histograms for all the training speech signals are computed using the universal GMM of training data. In testing, the histograms for each testing speech signal are generated using the same universal GMM which is used for calculating histograms of the training speech signals. The distances are computed between testing histograms, taking one at a time, and for each of the training histogram. Now, a distance_threshold is applied for a testing signal and count the number of distances which are less than the distance_threshold. If this count is greater than a predefined threshold on number of neighbors, then the test signal is detected as a normal, otherwise detected as an distressed one.

In order to illustrate the detection results for AH method at a particular distance_threshold, the distressed speech signals are considered for testing. The objective is to observe how many examples are correctly detected as distressed. TDR is then calculated from the number of correctly detected examples using the Eq. (11). Similarly, the normal speech signals from the test data are considered for testing. Based on the number of normal examples detected as distressed, the FDR can be calculated

Table 1 Detection results for AH method using UGMM with 256 modes and 512 modes for JSD and TV distance measure

256 modes			512 modes		
Threshold	FDR	TDR	Threshold	FDR	TDR
<i>JSD</i>					
0.2842	11.750	98.79	0.3500	12.95	99.52
0.2876	10.84	98.79	0.3535	12.95	99.52
0.2911	9.94	98.55	0.3570	11.14	99.28
0.2946	9.64	98.31	0.3604	9.94	99.28
0.2980	8.73	97.83	0.3639	9.64	99.03
<i>TV</i>					
0.5700	12.65	98.79	0.6450	12.95	99.03
0.5750	11.14	98.31	0.6500	12.05	99.03
0.5800	9.94	97.58	0.6550	11.14	98.79
0.5850	8.73	96.86	0.6600	10.24	98.55
0.5900	7.22	95.17	0.6650	8.43	98.07

using the Eq. (10). By varying the distance_threshold, the different TDRs are obtained with corresponding FDRs.

The detection results for this method using Jensen–Shannon distance measure (JSD) and total variation distance (TVD) measure with different number of modes (or Gaussian components) in UGMM is shown in Table 1. The detection results shown in Table 1 are obtained at 256 modes and 512 modes for JSD and TV distance measure.

The threshold on number of nearest neighbors is kept constant in all the cases.

4.2.3 Experimental Results for Audio Event Detection Using One-Class SVM

The one-class SVM is generated using all the MFCC vectors of the training data. One-class SVM is labeled +1 in a small region, where most of the vectors with which it was trained are mapped and -1 elsewhere. The experiments on SVM are conducted using [3]. It is important to note that all the MFCC vector values are linearly scaled to $[-1, 1]$ before building the SVM.

The RBF kernel is used for building this one-class SVM. The best value of parameter ν is chosen empirically, and kept constant.

The one-class SVM represents the training data. In testing, the MFCC vectors for a test signal are computed. The MFCC vectors are predicted using the one-class SVM. If the number of vectors predicted as distressed (-1 in our case) is greater than the number of vectors predicted as normal ($+1$ in our case) then the test signal is classified as distressed audio and vice versa. TDR and FDR are calculated using the Eqs. (11) and (10) respectively. In these experiments, ν is varied from 0.01 to 0.07. This give reasonably good results. By varying γ of RBF kernel and building different one-class SVM for each value of γ , the different prediction results on test signals are obtained. Hence, the different TDR and FDR values are obtained.

Table 2 Detection results for one-class SVM method with parameter $\nu=0.05$

Gamma (γ)	$2^{6.5}$	$2^{6.75}$
FDR (%)	7.2289	12.3494
TDR (%)	90.3382	93.7198

The detection results for OCSVM method with parameter $\nu = 0.05$ value is shown in Table 2. From the Table 2, it can be seen that as γ is increased, the TDR results are improved.

4.2.4 Experimental Results for Audio Event Detection Using Universal GMM Method

The universal GMM is built with all the MFCC vectors of the training data. K-means algorithm is used for UGMM parameter initialization, estimation-maximization (EM) algorithm is used for re-estimation of UGMM parameters iteratively. This universal GMM represents the training data. In testing, the MFCC feature vectors from the testing speech signal are first calculated. Then, the likelihood of test signal is found. The likelihood of test signal is compared with a predefined threshold. If likelihood of test signal surpasses the threshold, the test signal is then detected as normal, else detected as distressed.

Applying this procedure on all distressed testing data gives TDR and normal testing data gives corresponding FDR. By varying the threshold on likelihood, the different TDRs and corresponding FDRs are obtained.

The detection results for UGMM are shown in Table 3. The results are presented in terms of FDR and TDR by varying likelihood threshold for 256 and 512 modes in UGMM.

4.2.5 Experimental Results for Audio Event Detection Using Gaussian Clustering Method

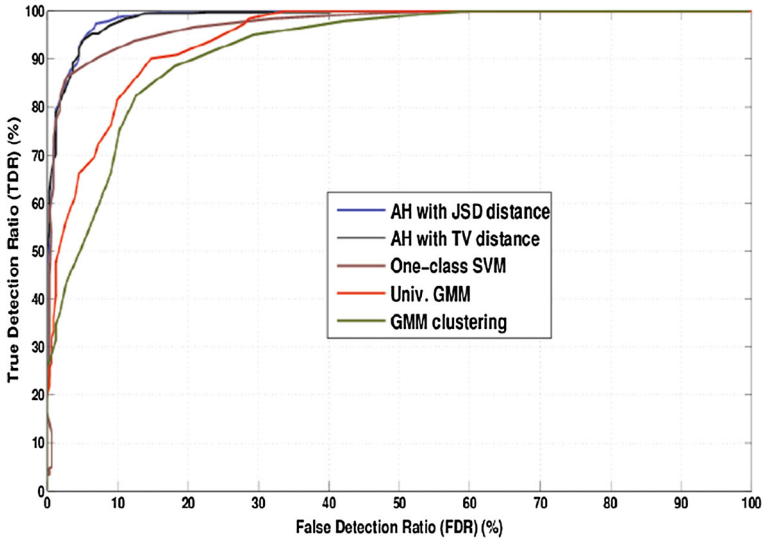
In this method, GMMs are built for each speech signal in the training data and a centric model is found. The centric model represents the complete training data. In testing, GMM with same number of modes used in training is built and the difference between

Table 3 Detection results for UGMM method with 256 modes and 512 modes in UGMM

256 modes			512 modes		
Threshold	FDR	TDR	Threshold	FDR	TDR
$10^{-6.7}$	7.23	72.22	$10^{-6.6}$	7.23	78.26
$10^{-6.6}$	9.04	76.09	$10^{-6.5}$	9.34	80.43
$10^{-6.5}$	9.94	81.40	$10^{-6.4}$	9.94	83.57
$10^{-6.4}$	11.45	84.06	$10^{-6.3}$	10.84	85.51

Table 4 Detection results for Gaussian clustering method with 2-modes

Threshold on distance	14.0543	15.7827	17.5110	19.2394	20.9677
FDR (%)	12.6506	10.2410	9.0361	6.9277	4.8193
TDR (%)	82.3671	75.1208	66.1836	58.2126	50.4831

**Fig. 5** ROC curve using 256 modes in universal GMM for AH method and UGMM method, $\nu = 0.05$ in one-class SVM and 2 mode GMM in GMM clustering method

centric model and test signal GMM is found. If the difference value is more than a predefined threshold then the test signal is detected as distressed otherwise detected as normal speech signal. Applying this procedure on all distressed testing examples gives TDR and normal testing examples give corresponding FDR. By varying the threshold on likelihood, the different TDRs and corresponding FDRs are obtained.

The detection results for this method are shown in Table 4. The different FDR and TDR values are obtained by varying threshold values. The results presented in Table 4 are obtained for 2-modes used in Gaussian clustering method.

4.2.6 Comparison of the Proposed Method with Other Probabilistic Methods

The performance of the proposed method are compared with the detection results of other three methods as mentioned above.

The comparison are illustrated through ROC curve by considering false detection rates on x-axis and true detection rates on y-axis. As described above, $\nu = 0.05$ in one-class SVM and 2 mode GMM in GMM clustering method produced better results compared to other parameter values. It is found experimentally that when 256 modes or 512 modes is used in UGMM, the AH method performs better than other three methods. This can be seen from the ROC curves illustrated in Figs. 5 and 6.

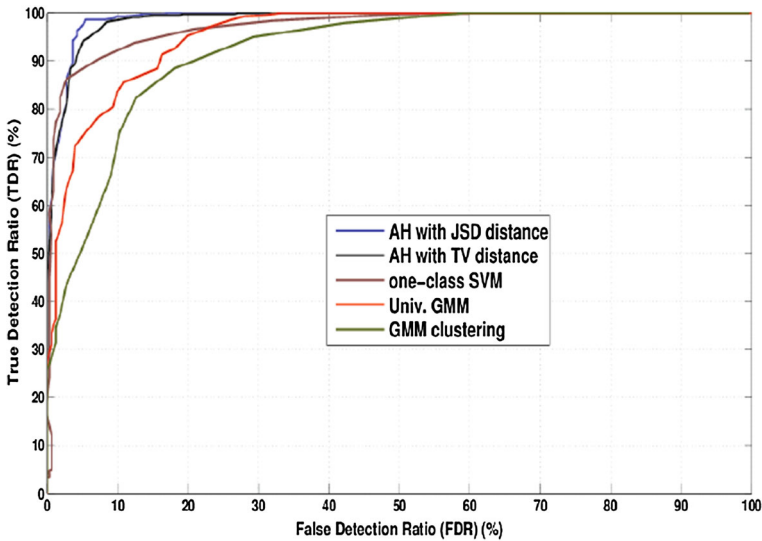


Fig. 6 ROC curve using 512 modes in universal GMM for AH method and UGMM method, $\nu = 0.05$ in one-class SVM and 2 mode GMM in GMM clustering method

Table 5 TDRs of different acoustic event detection methods using 256 mode and 512 mode UGMM at a constant FDR of 10 %

Method	256 Mode TDR (%)	512 Mode TDR (%)
AH-JSD	98.6	99.3
AH-TV	97.6	98.5
OCSVM	92.5	92.5
UGMM	81.4	83.6
GC	75.0	75.0

Applications like acoustic surveillance in assisted living, FDR equal to 10 % is allowable. With this FDR in comparison to TDRs of all the methods, MFCC feature vector show the best method used for acoustic event detection. From Table 5, it is clear that both audio histogram-based acoustic event detection methods, using JSD measure (AH-JSD) and using TV distance measure (AH-TV) outperform acoustic event detection method based on OCSVM, UGMM, and Gaussian clustering (GC).

5 Conclusion and Future Scope

In this paper, novel methods for acoustic surveillance using audio histograms are proposed. Experiments on acoustic event detection in a surveillance environment indicate that the proposed method performs significantly better than some of the widely used techniques in literature. The results are motivating enough to pursue further research on environmental audio detection in general using this method. Issues like combining

the MFCC with other feature vectors like intonation and Teager energy operator-based features, perceptual wavelet packet integration analysis-based features, and MPEG-7 audio protocol features are being addressed currently. Concatenating more than one set of features for improved detection performance is also an area of interest. An application of proposed probabilistic event detection method using audio histograms in acoustic surveillance for an assisted living environment can be pursued for future research. The brief overview of this application is presented herein.

Assisted living residences do not typically provide the level of continuous skilled nursing care found in nursing homes and hospitals. Hence, some sort of continuum care is needed in an assisted living residence. In this type of residence, continuum care can be provided using automatic surveillance. This application requires training of the system using proposed algorithm as explained in Algorithm 1 of Sect. 3.4. The system, after training, consists of UGMM of normal training speech signals and their histograms. One microphone is used to collect the audio data in the lab. The audio data are cut into intervals of one second each. The silences and noise audio signals are detected and deleted from the data. These speech signals are sent to the above-mentioned system, where the proposed algorithm is running.

The system computes histogram for each incoming speech signal and detects whether the speech signal is normal speech example or distressed speech example using the procedure given in Algorithm 2 of Sect. 3.4. If the speech signal is detected as a normal speech example, '0' is sent to the Asterisk server otherwise '1' is sent. On receiving '0' Asterisk does not take any action. When the Asterisk server receive '1', it initiates a phone call to a prerecorded phone number, given in its flow control PHP script, with a prerecorded voice message. The same experiment is done using an SMS gateway which sends a prerecorded SMS to a prerecorded phone number on receiving '1' (distressed event) from the system.

The problem with the proposed approach of event detection is that the certain non-speech events, like a door bang, or even animal cries may be detected as abnormal speech though they are not the distressed speech. In order to make the proposed algorithm useful in such cases, it is required to detect such non stationary events and remove from the data and then apply the proposed algorithm for event detection.

References

1. P.K. Atrey, A. Cavallaro, M.S. Kankanhalli, *Intelligent Multimedia Surveillance* (Springer, Berlin, 2013)
2. P.K. Atrey, M. Maddage, M.S. Kankanhalli, in Audio based event detection for multimedia surveillance, in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*, vol. 5, (IEEE, 2006), pp. V–V
3. C.-C. Chang, C.-J. Lin, LIBSVM: a library for support vector machines. *ACM Trans. Intell. Syst. Technol.* **2**, 27:1–27:27 (2011)
4. C. Clavel, T. Ehrette, G. Richard, Events detection for an audio-based surveillance system. in *Proceedings of the IEEE International Conference on Multimedia and Expo, (ICME 2005)*, (2005) pp. 1306–1309
5. M. Davy, F. Desobry, A. Gretton, C. Doncarli, An online support vector machine for abnormal events detection. *Signal Process.* **86**(8), 2009–2025 (2006). Special Section: Advances in Signal Processing-assisted Cross-layer Designs

6. B. Fuglede, F. Topsøe, Jensen–Shannon divergence and Hilbert space embedding. in *Proceedings of the International Symposium on Information Theory, (ISIT 2004)*, (2004) p. 31
7. J.H.L. Hansen, S. Patil, Speech under stress: analysis, modelling and recognition. in ed. by C. Miller, *Speaker Classification I Ser. Lecture Notes in Computer Science*, vol. 4343, (Springer, Berlin, 2007) pp. 108–137
8. J.H.L. Hansen, S.E. Bou-Ghazale, R. Sarikaya, B. Pellom, Getting started with SUSAS: a speech under simulated and actual stress database. *Eurospeech* **97**(4), 1743–1746 (1997)
9. A. Harma, M.F. McKinney, J. Skowronek, Automatic surveillance of the acoustic activity in our living environment. in *Proceedings of the IEEE International Conference on Multimedia and Expo, (ICME 2005)* (2005) 4 pp.
10. Z. Huang, Extensions to the k-means algorithm for clustering large data sets with categorical values. *Data Mining Knowl. Disc.* **2**(3), 283–304 (1998)
11. K. Hwang, S.-Y. Lee, Environmental audio scene and activity recognition through mobile-based crowd-sourcing. *IEEE Trans. Consum. Electron.* **58**(2), 700–705 (2012)
12. K. Kashino, T. Kurozumi, H. Murase, A quick search method for audio and video signals based on histogram pruning. *IEEE Trans. Multimedia* **5**(3), 348–357 (2003)
13. J. Kennedy, M.P. Quine, The total variation distance between the binomial and poisson distributions. *Ann. Probab.* **17**(1), 396–400 (1989)
14. J. Lin, Divergence measures based on the Shannon entropy. *IEEE Trans. Inform. Theory* **37**(1), 145–151 (1991)
15. L. Lu, F. Ge, Q. Zhao, Y. Yan, A SVM-based audio event detection system. in *Proceedings of the International Conference on Electrical and Control Engineering (ICECE 2010)* (2010) pp. 292–295
16. G. McLachlan, T. Krishnan, *The EM Algorithm and Extensions* (Wiley, Hoboken, 2007)
17. S. Ntalampiras, I. Potamitis, N. Fakotakis, Probabilistic novelty detection for acoustic surveillance under real-world conditions. *IEEE Trans. Multimedia* **13**(4), 713–719 (2011)
18. D. Reynolds, R.C. Rose, Robust text-independent speaker identification using Gaussian mixture speaker models. *IEEE Trans. Speech Audio Process.* **3**(1), 72–83 (1995)
19. B. Schlkopf, R. Williamson, A. Smola, J. Shawe-Taylor, J. Platt, Support vector method for novelty detection. *Adv. Neural Inform. Process. Syst.* **12**, 582–588 (1999)
20. G. Valenzise, L. Gerosa, M. Tagliasacchi, E. Antonacci, A. Sarti, Scream and gunshot detection and localization for audio-surveillance systems. in *Proceedings of the IEEE Conference on Advanced Video and Signal Based Surveillance, (AVSS 2007)* (2007) pp. 21–26
21. S. Xiang, J.J. Huang, Based audio watermarking against time-scale modification and cropping attacks. *IEEE Trans. Multimedia* **9**(7), 1357–1372 (2007)
22. L. Xie, A. Natsev, X. He, J. Kender, M. Hill, J.R. Smith, Tracking large-scale video remix in real-world events. *IEEE Trans. Multimedia* **15**(6), 1244–1254 (2013)
23. X. Zhang, X. Yin, Z. Yu, Robust audio watermarking algorithm based on histogram specification. in *Proceedings of the International Conference on Intelligent Information Hiding and Multimedia Signal Processing, IHHMSP '08* (2008) pp. 163–166