Received: October 17, 2018

International Journal of
Intelligent Engineering & Systems

http://www.inass.org/

Classification of Ontological Violence Content Detection through Audio Features and Supervised Learning

Yakaiah Potharaju[1]* **Manjunathachari Kamsali[2]** **Chennakesava Reddy Kesavari[3]**

[1]Department of Electronics and Communication Engineering,
Rayalaseema University, Kurnool, Andhra Pradesh 518002, India
[2]Department of Electronics and Communication Engineering,
Gitam University, SangaReddy, Telangana, India
[3]Department of Electronics and Communication Engineering,
Jawaharlal Nehru Technological University, Telangana, India
* Corresponding author's Email: potharaju.yakaiah@gmail.com

Abstract: Violence detection is one of the important aspects, which can be used in different applications. Based on the data format, the violence can be defined in many ways. This paper focused to develop an automatic violence detection framework from audio type data. To do this, a new and efficient set of features are extracted from the audio signals, which provides more discrimination between different types of violence types in audio signals. Considering both spatial and Mel frequency characteristics of audio signals, totally 12 statistical functionals are accomplished to define every signal. Furthermore, the violence is defined in an ontological fashion, such that the all possible violence types which signify the violent behavior are detected. Extensive simulations are carried out over the proposed detection framework by considering the audio signals extracted from different video clips ripped from different movies. The performance is analyzed through the Receiver Operating Characteristics like, Accuracy, Precision, Recall, and False Positive Rate and the obtained results verify the performance enhancement and show a better performance than the conventional approaches.

Keywords: Violence detection, Audio, Ontology, SVM, Accuracy, False positive rate.

1. Introduction

In the recent years, the problem of recognizing the human actions from an audio-visual data has become submissive though the utilization of computer vision techniques [1]. Despite the benefits of this application, the violence detection has been comparatively less studied. However, the violence detection has an immediate applicability in the surveillance and monitoring applications. The main applications behind the deployment of video surveillance systems in the schools, psychiatric care facilities, prisons etc., are to alert the authorities from the dangerous situations. However, the human operators engaged for this accomplishment are overwhelmed with the huge number of camera feeds and also the manual operations are slow, resulting in a strong demand for automated alert system. Simultaneously, there is an increasing demand for automated processing systems that processes the huge sized videos to the websites.

Due to the vital importance of the Violence detection in the provision of security alerts in the video-surveillance system both at scientific level and application level, it has gained a greater importance in the current day's research. The violence detection is different in some particularities from the generic human action recognition. Due to these all reasons, the research interest in violence detection has been steadily growing. Further the psychological research on media violence [2], has upheld its negative effects on the children in their emotional behavior, attitude etc., and highlighting these aspects makes to develop new violence

detection tools which filters out such type of content form the data. Some of the research work is accomplished over the violence detection considering that the fist fighting, kicking, gunshots and explosions as violence creating object [3]. Due to the recent crime knowledge in the human mind, they are trying to create the violence in so many ways, but the up to developed automatic violence detection are not able to tackle these problems and there is a necessity to develop a novel automatic violent detection system. Further, there is no fixed definition for violence, detection of violence needs to consider multiple aspects like sounds, scenes, shots, screams etc. To detect the violence is such situations, the detection system needs to more effective in the learning process and it needs to be trained in that manner only. Ontology is one of the hierarchical strategy through which an objective can be defined in a hierarchical fashion, this paper accomplished ontology for violence representation.

Since an audio-visual data is composed of both audio and visual features, the violence detection system also considers both the audio and visual features to define the violence. For example, the screams are considered as violence in the audio data and the fighting is considered as violence in visual data. Focusing only on the audio features, this paper develops a new violence detection system. Furthermore, this paper also defines the violence in an ontological fashion, such that the all possible audio models which define the violence are considered for detection. Based on the both the spatial and Mel frequency characteristics of an audio signal, a new set of features are extracted which makes the detection system more effective and computationally less expensive. Followed by feature extraction this paper also considered Support Vector Machine algorithm for classification of violent signals.

Remaining paper is structured as follows: the details of literature survey are described in section 2. The illustration about the proposed violence detection framework is described in section 3. Experimental evaluations are illustrated in section 4 and the provision of conclusion is described in section 5.

2. Literature survey

Various approaches are developed to perform violence detection through auditory models. In [4], an energy entropy based acoustic scene classification is proposed by Geiger et.al. Abrupt changes in the audio signal can be detected through energy entropy, which, in general, may characterize the violence content. Though this energy entropy features can detect the violent content more successfully, some non-violent sounds like thunders are also detected as violence. Further, one more method is developed in [5] to segment and classify the audio signals based on the signals entropy. Next, the authors of [6] developed a new method based on Machine Learning and signal processing techniques to identify the vocal and non-vocal regions of the songs. The characteristics of vocal and non-vocal segments were obtained by using Artificial Neural Networks (ANN).

By extracting some important time domain and frequency domain features from an audio signal, a new violence detection method is developed by Giannakopoulos et al, [7] for audio signals. Further, different statistical functions are accomplished over the obtained feature set and they are trained through the support vector machine (SVM) algorithm. This SVM algorithm decides the segment belongs to violent or not. Extending to the method developed in [7], a new method is proposed in [8] to classify the audio segments recorded form movies, aiming to detect the violent data, to protect the social sensitive groups, e.g. children. At the end, the feature extraction technique proposed in [8] evaluated totally twelve features. Bayesian Network (BN) is accomplished in the one-versus-all fashion to classify the audio segment into totally six classes and three out of them are violent classes and remaining are non-violent.

As the MFCC's are more prominent in the evaluation of statistical properties of an audio signal, an Acoustic Event Classification (AEC) is proposed by Choez and Antolin [9] for audio events classification under both noisy and clean environments. However, a simple MFCC is not effective for violence classification. To overcome this issue, a multi-classifier system was developed by Zhang et al., [10] by considering the multi-classifier systems such as Random Forest with MCS, and Bagging with Adaboost. Though the multi-classifier is effective the violnec detection, the ontology adds an extra complexity. Furthermore, the feature extraction technique is very simple and not able to discriminate the violence classes perfectly.

Further, a new method is proposed by Thaweesak [11] to find the depression of speaker based on his/her speech. In this approach, the full-band and further sub-band entropies of eight evenly separated frequency bands of 625 Hz estimated from the female voiced segments were computationally extracted and consequently used to form the parameter models for between group classifications. Further a machine learning classifier is

accomplished over the extracted feature set to perform classification. A Gaussian Mixture Model (GMM) based audio signal detection and classification mechanism is proposed by M. Balede et al, in [12]. Here the sound spectrum is modeled with mixture model and then a dictionary is formulated. Further the classification is done through the estimation of likelihoods and the best match is used as a result.

Since there exist different environmental sounds, a new method is proposed by M. Loughlin et al [13] for environmental sounds detection based on the time-frequency audio features. Considering the semantics of different audio signals, different features are proposed to recognize different sounds form the audio signals. These features include both the auditory image front end features and spectrogram front end features. Environmental Sounds like rain sounds, chirpings of birds and insects which are having a typical flat and broad spectrum, which was similar to the spectrum of noise, based on the signatures in the time domain [15]. One more work is carried out by S. Sameh, Z. Lachiri [14] to analyze the environmental sounds through the Spatio-Temporal features analysis. Mainly they focused on the differentiation between the environmental sounds in urban environments and polyphonic music. The used Log-Gabor filters-based feature is to supplement the MFCC features to yield higher classification accuracy for environmental sounds. However the involvement of Log-Gabor filters yields an extra computational burden over the system due to its typical accomplishment.

S. Saman et al, [16] used auditory models to classify the violent scene based on ensemble learning. This approach extracted the Zero Crossing Rate (ZCR) of the audio signals as a feature and applied Random Forest algorithm for classification purpose. However a single ZCR can't provide much discrimination between different audio classes. Furthermore, it becomes too tough for the Ontological based classification due to the hierarchical strategy. Considering the low level multiple features of audio signals, Vu Lam et al, [17] tried to detect the violence scenes. One more approach is developed by Marta et.al., [18] by considering the multiple features namely, Mel-Frequency Cepstral Coefficients (MFCCs), Pitch [19], Harmonic Noise rate (HNR), Short time energy [20], ZCR etc., to detect the violence in real time environments. Further three different classifiers: a Least Squares Linear Detector (LSLD), a simplified version of Least Squares Quadratic Detector (LSQD) and a Neural Network based Detector with 5 hidden neurons are accomplished for classification.

Further, the violence can be defined through the emotions of speech uttering persons also. A person having anger emotion can be defined as violence and the remaining emotions can be defined as non-violence [21, 22]. For examples, the sadness, happy, joy, neutral etc., all are considered to be non-violent and only the anger is defined as violence. In earlier there are so many approaches are developed to detect the emotion from the audio signals [23]. With the help of MFCCs of speech signals, S. Demircan and H. Kahramanl [24] developed an emotion recognition system with unsupervised learning. This approach used K-Nearest neighbor algorithm to classify the emotions. However, the unsupervised learning algorithm constitutes an extra complexity for recognition system. Some more approaches are proposed based on the supervised learning [25] like Neural Networks [26] Recurrent Neural Networks [27] to identify the emotions more accurately from speech signals.

Sara et al, [25] introduced an optimized model of brain emotional learning (BEL) that merges the Adaptive Neuro-Fuzzy Inference System (ANFIS) and Multilayer Perceptron (MLP) for speech emotion recognition. The main limitation of the MLP algorithm is that, because of the way it is trained, it cannot guarantee that the minimum it stops at during training is the global minima. Another limitation of the MLP algorithm is that the number of Hidden Neurons must be set by the user, setting this value too low may result in the MLP model under fitting while setting this value too high may result in the MLP model over fitting. Recently sparse coding framework is developed by Diana et al, [28] to recognize the emotions from the speech signal. The sparse coding framework is adopted as a means to automatically represent features from audio and propose a hierarchical sparse coding (HSC) scheme.

3. Proposed violence detection framework

The proposed violence detection framework is accomplished under two stages, namely, feature extraction and classification. In the feature extraction phase, the required set of features is extracted from the audio signals and in the classification phase, the obtained feature set was processed for classification with a supervised learning algorithm. In this paper, the binary support vector machine classifier was used for classification purpose.

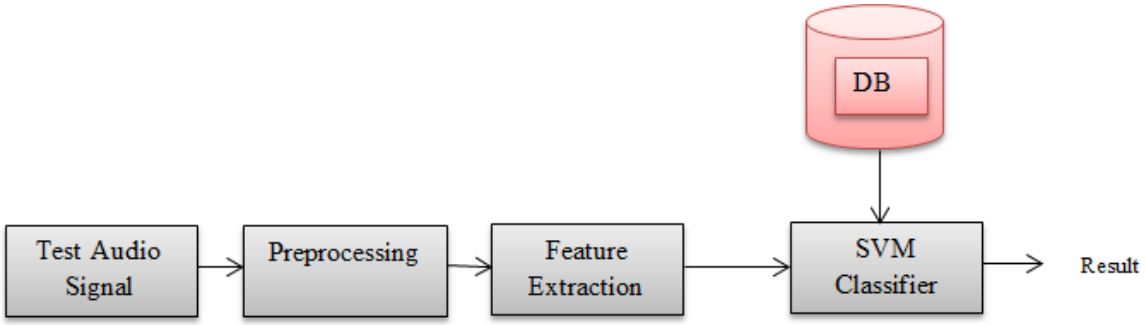

Figure. 1 Generalized architecture of audio violence detection system

Due to the binary classification nature of SVM classifier, it is accomplished over multiple instances to classify the audio signals into more deep class. Initially, in this section, the ontological structure of violence considered in this paper is illustrated and then the audio features considered for extraction are explained. Finally the architecture of Multi-class SVM is represented. The overall architecture of proposed violence detection system is represented in Fig. 1.

3.1 Audio semantics for violence

Due to various definitions of violence in various perspectives, a simple aspect cannot give a conclusion about the violence. In the case of audio type data, the sounds like screams and huge abrupt sounds can be considered as violence. Hence this paper accomplished to define the violence with respect to an ontological framework. Further this paper only focused on audio data, the violence is defined with respect to audio signals only.

Additional clues, relative to violence, increasing the accuracy of violence detection exist in the auditory modality. Contrary to visual semantics in audio semantics every class that is defined in the ontology is also extracted from the audio classification algorithms. Based on the auditory aspects, the broad class of violence is defined as music and sound. The further definitions of violence based on sounds are defined as person related sounds, weapon related sounds, environmental sounds and fight related sounds. Next the person related sound depends on the scream and speech of person. The weapon related sound depends on the weapon used to create violence like gun, sword, bottle and some other objects. The violence definition according to the environmental sounds is defined based on the abruptness and smoothness. Thus the audio classes of interest are Screams, Speech, Gunshot, Sharp Environmental Sound, Smooth Environmental Sound and Fights (beatings). The detailed hierarchy of audio semantics for violence is described in Fig. 2.

3.2 Audio features

According to the proposed feature extraction methodology, totally, twelve features are extracted from every audio segment. To do this, initially every audio signals has broken into some non-overlapping segments based on the time elapsed. Since 12 different features are extracted from every segment, 12 features are obtained for every segment and the entire information present in that segment is represented through these features only. Subsequently, a standard statistic like average value or standard deviation is measured for every audio segment for the obtained features and forming a 12-D vector representing the entire information in a single vector. The set of features and the statistical functions considered in this approach are depicted in Table 1.

For an extracted audio clip segmented form the audio sequence, every audio clip is depicted through different low-level features, namely, “12 MFCCs, Root Mean Square Frame Energy (RMSFE), pitch, Harmonic Noise Ratio (HNR) and Zero Cross Rate (ZCR)”. Further over the obtained 12 features, 12 statistical functions namely, skewness, standard deviation, kurtosis, four extremes (i. e., minimum and maximum value, relative position, and ranges) as well as two linear regression coefficients with their mean square error (MSE) are applied to the low-level features and their deltas. In this manner for an audio sequence, totally a2 features are extracted and the trained to the classifier. The details of low-level features and the statistical functionals those are applied over the low-level features are described in the Table 1.

3.3 Multi-class SVM

Since the consideration of hierarchical definitions for violence in the audio signals, the

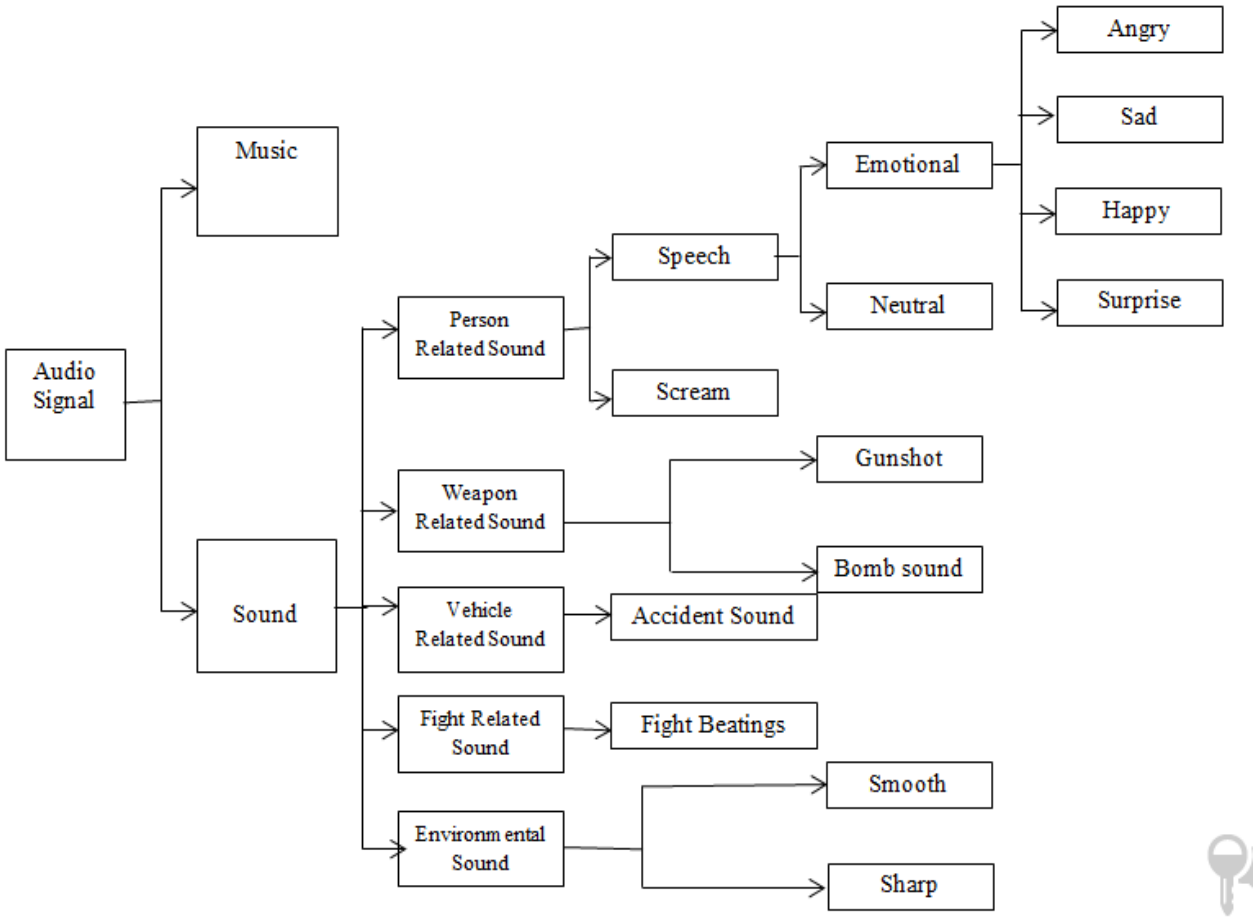

Figure. 2 Hierarchy of violence ontology obtained from the audio modality

classification also needs to be accomplished in the hierarchy fashion. Furthermore, due to the robustness and efficiency of SVM in the detection of required events, this paper also considered the SVM algorithm for classification purpose. However the SVM is a binary classifier and it only classifies only two classes at a time. Hence to realize the proposed ontological violence detection through SVM classifier, the SVM algorithm is also accomplished in hierarchical fashion at multiple instances. One-versus-one, one-versus-all and binary tree are the three most famous classification strategies when the SVM is accomplished for classification. Considering the computational burden and additional number of SVM classifiers, this method approached to binary tree SVM classifier. The number of SVM classifiers required in the case of one-versus-one and on-versus-all are observed as *k(k-1)/2 and* '*k*' respectively and the count is only 'k-1' in the case of binary tree SVM classifier. Thus the proposed classification model simply followed the binary tree SVM to classify the ontological violence events from the audio signal.

In the initial classification phase, the signal is classified into music and sound classes. In this context, the music is considered as no-violence and sound is considered as violence. Further the sound

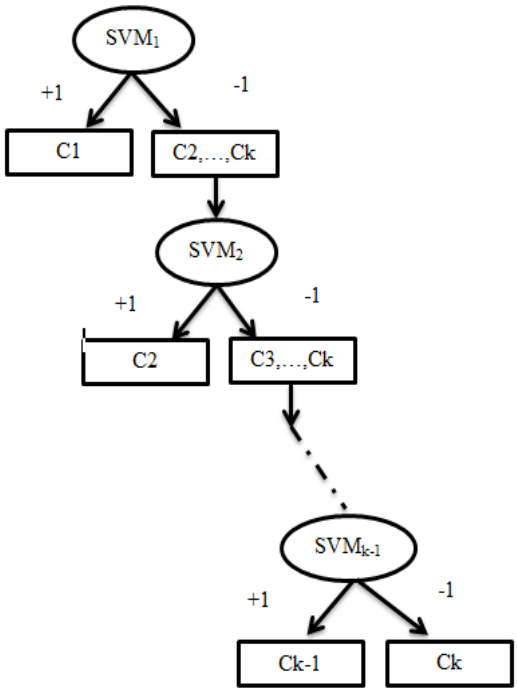

Figure. 3 Violence detection hierarchy based on binary tree multi class SVM

Table 1. Acoustic features and statistical functionals

Raw Features	Statistical functionals
Pitch	Mean, Standard deviation (SD), Kurtosis
Root Mean Square Frame Energy (RMSFE)	Skewness, minimum and maximum value
Zero Cross rate (ZCR)	relative position, Ranges
Harmonic to noise ratio (HNR)	Two linear regression coefficients with their
12 Mel-frequency Cepstral coefficients (MFCC)	Mean Square Error (MSE) of regression coefficients

signal is classified into five types of sounds, namely, Person related sound, weapon related sound, vehicle related sound, fight related sound, and environmental related sound. Next the person related sound is further classified as speech and scream. Here the scream is considered as violence and further the speech is classified as emotional and neutral. Under the emotional, it is further classified as Angry, Happy, Sad and Surprise.

Among these four classes, the angry is only considered as violence and the remaining are non-violent. Next, under the weapons related sound, two types of weapons sounds namely, Gunshot and bomb sound are considered as violence. In the case of vehicle related sound, only accident sound is considered as violent and the remaining is non-violent. In the case of fight related sound, the audio signal representing the beatings are considered as violent. Finally the environmental sounds are also considered as violence depends on the smoothness. If any sharp change in the environmental sound is observed in the audio signal, then it is considered as violent otherwise non-violent. In this manner, the given audio test signal is classified as violent or non-violent and it is manually verified to check the performance. A simplified architecture of MC-SVM is represented in Fig. 3.

After the extraction of sufficient set of features from the audio signals, they are processed for training. The training process leads to the extraction process followed by training through SVM algorithm. To obtain an optimal solution, the decision function according to the SVM is defined as

$$f(t) = sgn\left(\sum_{i=1}^{p}(\alpha_i - \hat{\alpha}_i)K(t_i, t_j) + b\right) \quad (1)$$

Where

$\alpha_i \; and \; \hat{\alpha}_i$ = coefficients of the Lagrange multiplier for the i^{th} sample of audio feature

$K(t_i, t_j)$ = kernel function and

b= an arbitrary constant.

Among the all possible kernel functions of SVM, classifier, the most popular and effective kernel, RBF kernel is only used in this approach at the classification phase to perform decision making. Mathematically the RBF kernel is represented as

$$K(t_i, t_j) = exp\left(\frac{-\|t_i - t_j\|^2}{\sigma^2}\right), \sigma \in R \quad (2)$$

According to the functional theory, as long as the function $K(t_i, t_j)$ satisfies Mercer's condition, it can be denoted as a positive definite kernel.

4. Simulation results

To evaluate the developed approach, extensive simulations are conducted over a new dataset acquired form Indian movies. Further the performance is measured through the performance metrics, including, Accuracy, Recall, Precision, F-Measure, and False Positive Rate (FPR).

4.1 Simulation setup

For training and evaluation purposes, totally 2000 videos with different scenes are ripped from 10 different films. Further form every video the respective audio signal is extracted. On an average, the duration of every video is 3 min. After extracting the audio signals from every video, they are manually annotated into the respective violent/non-violent scenes. Particularly, the audio signals those are annotated by humans are only used as ground truth in the evaluation of performance of developed detection system. Basically the violent signals are extracted from the audio signals like gunshots, screams, explosions, fight sounds whereas the non-violent signals are extracted from the speech signals and music signals. According to the ontological strategy depicted in the above section, different types of audio signals are extracted from different movies. Furthermore, the same class of violence with different emotions is also extracted to further check the detection mechanism performance. For an audio signal extracted from the movie, it is divided initially into the non-overlapping frames through hamming window of time span 400 msec. Since the Indian movies consist of different background sounds by which the violence can be defined in

Table 2. Experimental data considered for simulation

Movie Name	Class of Audio signals Extracted
Aashqui 2	Music, Sad
Raaz 2	Music, Fear, Scream
Queen	Happy, Neutral
Gangster	Gunshot, Music, Happy
Black Friday	Bomb Sound, scream
Murder 2	Scream, Angry
Newton	Surprise, sad
1924 Evil returns	Environmental Sounds
Singham Returns	Fight Beatings, Accident Sounds
Don	Accident Sounds

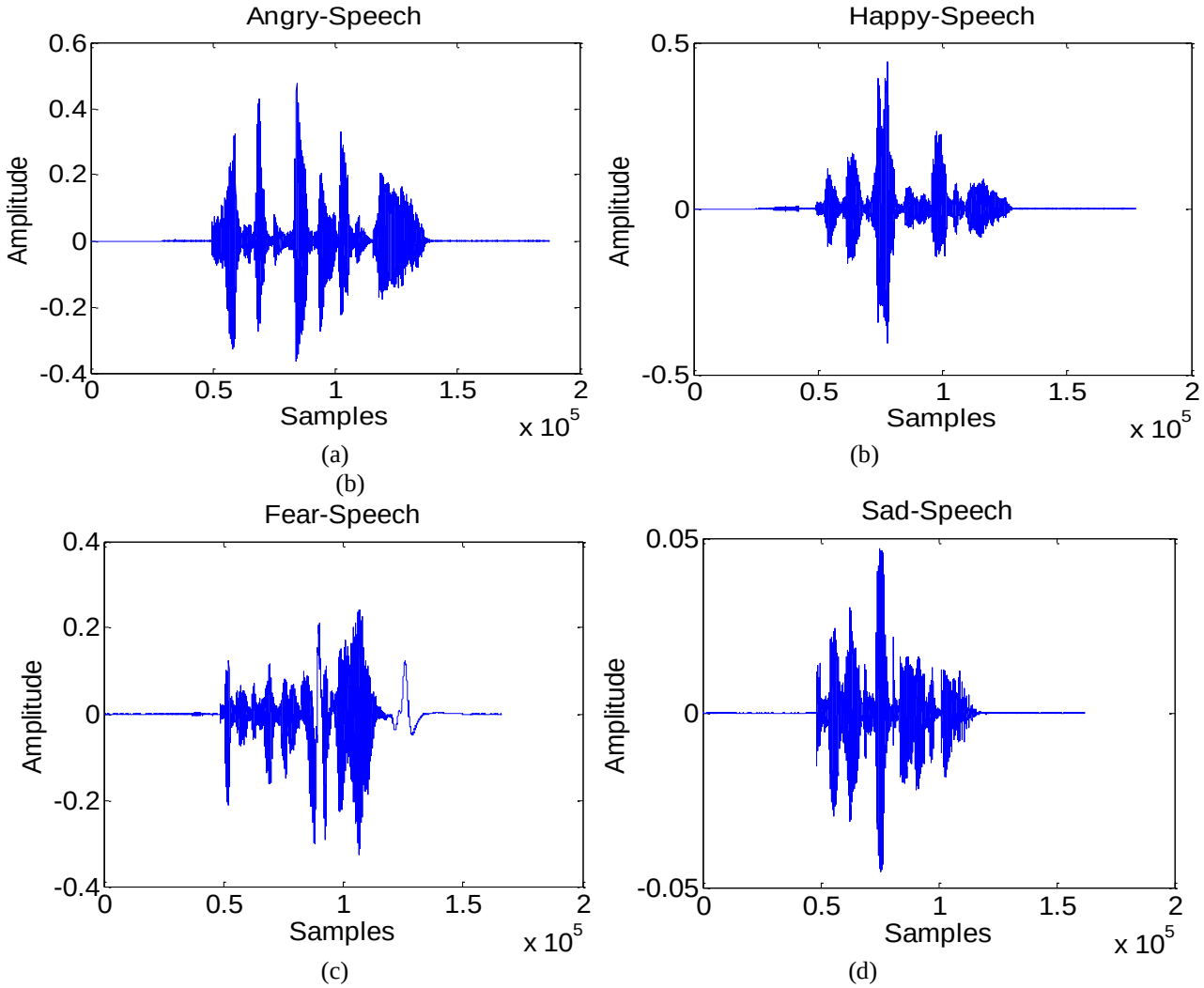

Figure. 4 Sample audio signals tetsed: (a) Angry, (b) Happy, (c) Fear, and (d) Sadness

different aspects, those are only considered for simulation experiments. The movies considered to extract different sounds are show in Table 2. The Sample audio signals representing the five emotions, namely Angry, Happy, Fear, Sad and Surprise are depicted in the Fig. 4.

4.2 Results

In this paper, the Accuracy, Precision, Recall, FPR and F-Score are considered to evaluate the performance of proposed approach. The basis for these metric evaluations is confusion matrix and it is represented in Table 3.

Table 3. Sample confusion matrix

		Predicted	
		Violent	Violent
Actual	Violent	TP	FN
	Non-Violent	FP	TN

Table 4. Performance metrics

Metrics	**Sara. et. al [25]**	**Vu Lam. et. al [17]**	**Proposed**
Precision (%)	72.3963	75.6637	77.8853
Recall (%)	62.8936	64.0217	66.7441
F-Measure (%)	63.9978	67.4571	71.8902
Accuracy (%)	75.9817	76.3319	78.2238
False positive Rate (%)	0.3634	0.3419	0.3006

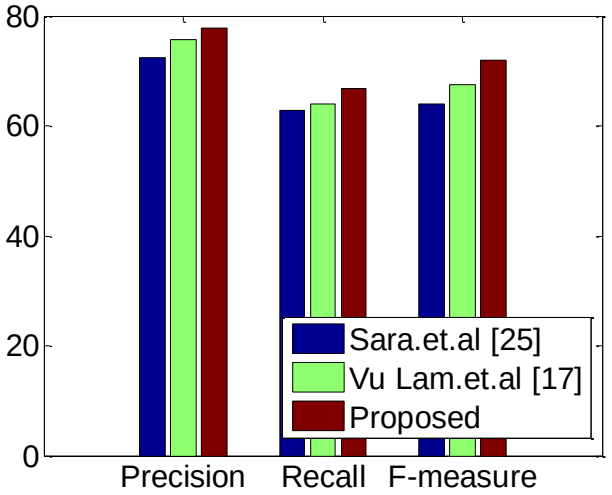

Figure. 5 Performance comparison through Precision, Recall and F-Measure

Based on the obtained TP, TN, FP and FN values from the confusion matrix, performance metrics are evaluated and the respective mathematical representation is given as;

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (3)$$

$$Precision = \frac{TP}{TP+FP} \quad (4)$$

$$Recall = \frac{TP}{TP+FN} \quad (5)$$

$$False\ Positive\ Rate = \frac{FP}{TN+FP} \quad (6)$$

$$F - Score = \frac{2\times Recall\times Precision}{Recall+Precision} \quad (7)$$

Where
TP= True Positives, TN = True Negatives, FP= False Positives and
FN = False Negatives.

After testing different audio signals through the proposed framework, the obtained performance metrics, Accuracy, Precision, Recall, F-Measure and False Positive rate are tabulated in Table 4. Furthermore, the Table 4 also describes the comparison between the proposed and conventional approaches in the detection of violence.

Table 4 describes the details of obtained performance metrics for the proposed approach after accomplishing it over the experimental audio data. From the Table 4, one can be perceived that the proposed mechanism attained an improved accuracy, precision, recall and F-measure. This improvement is due to the best features extraction which provides more discrimination between the audio signals; thereby the classifier can classify the audio signals more accurately. Further it is also perceived that the proposed mechanism attained a less FPR due to the accomplishment of Multi-class SVM. The conventional approaches focused only on a single objective, for instance, the Method proposed by Sara et.al., [25] mainly focused over the emotion recognition and the method prosed by Vu Lam et.al., [17] only on violence. Whereas the proposed approach tried to accomplish in an ontological fashion, the detection is more typical and the system needs more knowledge about the characteristics of violence at every stage.

Fig. 5 describes the details of obtained precision, recall and F-measure for the proposed as well as for conventional approaches. The values plotted in the Fig. 5 are the average values obtained after detecting the violent and non-violent audio signals. Since the proposed approach focused only on the features which are more discriminative, the test signals are classified more precisely by which the precision followed by recall and F-measure have increased. The main reason behind this achievement is the consideration of multiple features and also a more effective supervised learning algorithm, MC-SVM. A simple Binary SVM won't work effectively in such case. Furthermore like the conventional approaches, only a few set of features are not able to provide much discrimination between the violent

and non-violent classes at every ontology step. The classifier needs more discrimination to derive the support vector for all classes thereby the classification will be effective. This process results in an increased precision followed by detection rate (recall) compared to the conventional approaches.

Figs. 6 and 7 describe the details of obtained average FPR and Accuracy details of proposed as well as conventional approaches. Since the proposed set of features extracted for every audio signal are more informative and covers different classes in a single aspect, within less feature count, the proposed mechanism attained an improved accuracy followed by decrement in the false positive rate. For a given test audio signals, the proposed mechanism almost detects as it is by which the false positives are getting reduced. It can also be observed from the above figures, compared to the conventional approaches the proposed approach has achieved an improved accuracy and reduced FPR.

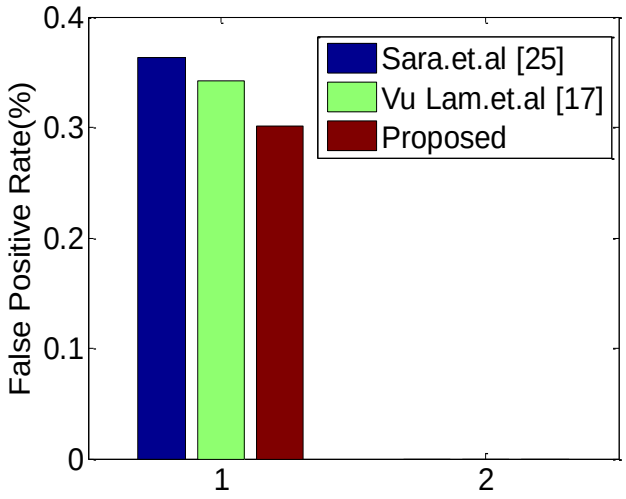

Figure. 6 Performance comparison through false positive rate

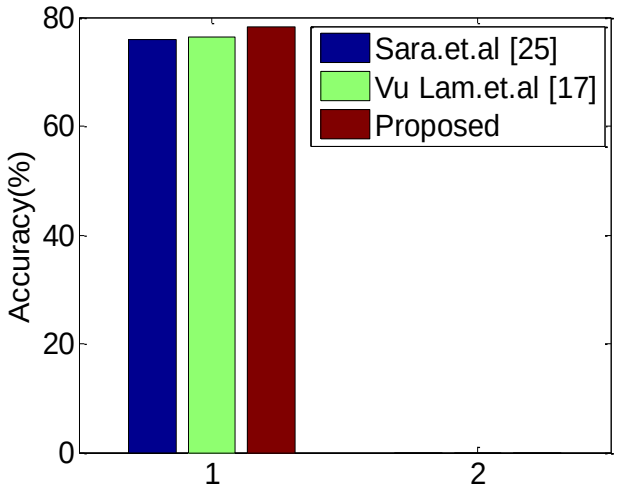

Figure. 7 Performance comparison through accuracy

The reduced FPR of proposed mechanism reveals its significance in the detection of violence and non-violence as it is. Next, this paper considered the violence in the ontological fashion, the detection of all violence at every step is very typical and it can be solved only with extraction of multiple features for every segment of speech signal. Since the developed framework extracted totally twelve features, the ontological based violence is achieved in a better way and it is proved with the above figures.

5. Conclusion

Due to the rapid growth in the technology and its effect on the human lives, detection of every aspect thorough the machine has become more prominent. Consequently, this paper also focused to develop an automatic violence detection system through machine learning algorithms through the auditory data. Since the violence has a number of definitions, defining the violence in only few orientations does not make the system robust in the detection. Hence this paper considered to define the violence in an ontological fashion and tried to cover all the possible definitions of violence. For this purpose, the proposed system considered Indian Bollywood Movies as test set and ripped 2000 audio signals form 10 different movies. For every audio signal, a set of features are measured and processed for the detection system. After simulation, the performance is measured through ROC metrics and observed that the proposed system have achieved an improved performance in the detection of all types of violence.

On an average the proposed approach obtained an improvement in the accuracy is observed as 2.2421% and 1.8919% from the conventional approaches, Sara et al., [25] and Vu Lam et al., [17] respectively. Next, the reduction in the false positive rate is observed as 0.0628% and 0.0413% from Sara et al., [25] and Vu Lam et al., [17] respectively.

This paper considered only audio features to classify the movie scene whether it is violent or non-violent. Along with audio semantics, if video semantics are also considered for violence detection, then it will be more effective and this work can be further extended in the future in that way.

References

[1] R. Poppe, “A survey on vision-based human action recognition”, *Image and Vision Computing*, Vol. 28, No.6, pp.976 – 990, 2010.

[2] M. Ángel Vidal, M. Clemente, and P. Espinosa, “Types of media violence and degree of

acceptance in under-18s", *Aggressive Behavior*, Vol.29, No.5, pp. 381–392, 2003.

[3] K. D. Browne and C. H. Giachritsis, "The influence of violent media on children and adolescents: A public-health approach", *The Lancet*, Vol.365, No.9460, pp. 702–710, 2005.

[4] J. T. Geiger, B. Schuller, and G. Rigoll, G, "Large-Scale Audio Feature Extraction and SVM for Acoustic Scene Classification", In: *Proc. of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp.1-4, 2013.

[5] M. Arumugam and M. Kaliappan, "An efficient Approach for Segmentation, Feature Extraction and Classification of Audio Signals", *Circuits and Systems*, Vol.7, No.5, pp.255-279, 2016.

[6] Y. S. Murthy and S. G. Koolagudi, "Classification of Vocal and Non-Vocal Regions from Audio Songs Using Spectral Features and Pitch Variations", In: *Proc. of IEEE 28th Canadian Conference on Electrical and Computer Engineering*, pp.1271-1276, 2015.

[7] T. Giannakopoulos, A. Pikrakis, and S. Theodoridis, "Violence content classification using audio features", In: *Proc. of the 4th Hellenic Conference on Artificial Intelligence*, pp. 502–507, 2006.

[8] T. Giannakopoulos, A. Pikrakis, and S. Theodoridis, "A Multi-Class Audio Classification Method With Respect To Violent Content In Movies Using Bayesian Networks", In: *Proc. of IEEE International Workshop on Multimedia Signal Processing*, pp. 90–93, 2007.

[9] J. L. Choez, and A. Gallardo-Antolín, "Feature Extraction Based on the High-Pass Filtering of Audio Signals for Acoustic Event Classification", *Computer Speech & Language*, Vol.30, No.1, pp.32-42, 2015.

[10] Y. Zhang, D. J. Lv, and H. S.Wang, "The application of multiple classifier system for environmental audio classification", *Applied Mechanics and Materials*, Vol. 462-463, No.11, pp. 225–229, 2014.

[11] Y. Thaweesak, "Spectral Entropy in Speech for Classification of Depressed Speakers", In: *Proc. of International Conference on Signal-Image Technology & Internet-Based Systems*, pp.1-6, 2016.

[12] M. Baelde, C. Biernacki, and R. Greff. "A mixture model-based real-time audio sources classification method", In: *Proc. of the 42nd IEEE International Conference on Acoustics, Speech and Signal Processing*, pp.52-59, 2017.

[13] Mc Loughlin, H. M. Zhang, Z. P. Xie, Y. Song, and W. Xiao, "Robust Sound Event Classification using Deep Neural Networks", *IEEE Transactions on Audio, Speech, and Language Processing*, Vol.23, No.3, pp.540–552, 2015.

[14] S. Sameh and Z. Lachiri, "Using Spectro-Temporal Features for Environmental Sounds Recognition", *American Journal of Circuits, Systems and Signal Processing*, Vol. 1, No. 3, 2015, pp. 60-68.

[15] S. Chachada and C. C. J. Kuo, "Environmental sound recognition: A survey", In: *Proc. of International Conference on Signal and Information Processing*, pp.45-50, 2013.

[16] S. Sarman and M. Sert, "Audio based Violet Scene Classification using ensemble learning", In: *Proc. of International Symposium on Digital Forensic and Security*, pp.1-6, 2018.

[17] V. Lam, S. Phan, and D. DInh, "Evaluation of Multiple features for violent scenes detection", *Journal of Multimedia Tools and Applications*, Vo.76, No.5, pp.7041-7065, 2017.

[18] M. B. Duran, R. G. Pita, and H. S. Hevia, "Acoustic Detection of Violence in Real and Fictional Environments", In: *Proc. of International Conference on Pattern Recognition Applications and Methods*, pp.456-462, 2017.

[19] R. Gil Pita, B. Lopez Garrido, and M. Rosa Zurera, "Tailored MFCCs for sound environment classification in hearing aids", *Advanced Computer and Communication Engineering Technology, Lecture Notes in electrical engineering*, Vol.315, No.3, pp. 1037–1048, 2015

[20] M. Jalil, F. A. Butt, and A. Malik, "Short-time energy, magnitude, zero crossing rate and autocorrelation measurement for discriminating voiced and unvoiced segments of speech signals", In: *Proc. of International Conference on Technological Advances in Electrical, Electronics and Computer Engineering*, pp. 208–212, 2013.

[21] B. Schuller and A. Batliner, "Computational Paralinguistics: Emotion, Affect and Personality in Speech and Language Processing", New York, NY, USA: Wiley, Nov. 2013.

[22] M. A. Quiros Ramirez, and T. Onisawa, "Considering cross-cultural context in the automatic recognition of emotions", *International Journal of Machine Learning Cybernetics*, Vol. 6, No.1, pp. 119–127, 2015.

[23] T. S. Gunawan, M. F. Algifhari, M. A. Morshidi, and M. KArtiwi, "A review on emotion recognition algorithms using speech analysis", *Indonesian Journal of Electrical Engineering and Informatics*, Vol. 6, No. 1, pp. 12-20, 2018.

[24] S. Demircan and H. Kahramanl, "Feature Extraction from speech data for Emotion recognition*", Journal of Advances in Computer Networks*, Vol.2, No.1, pp.28-30, 2014.

[25] M. Sara, S. Saeed, and A. Rabiee, "Speech emotion Recognition Based on a Modified Brain Emotional Learning Model", *Biologically Inspired Cognitive Architectures*, Vol.19, No.1, pp.32–38, 2017.

[26] P. Sathit, "Improvement of Speech Emotion Recognition with Neural Network Classifier by Using Speech Spectrogram", In: *Proc. of International Conference on Systems, Signals and Image Processing*, pp.1-8, 2015

[27] W. Lim, D. Jang, and T. Lee "Speech Emotion Recognition using Convolutional and Recurrent Neural Networks", In: *Proc. of Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, pp.1-8, 2017.

[28] D. Torres Boza, "Hierarchical sparse coding framework for speech emotion recognition", *Journal of Speech Communication,* Vol. 99, No.5, pp. 80-89, 2018.