VIVA-Tech International Journal for Research and Innovation
ISSN(Online): 2581-7280

Volume 1, Issue 4 (2021)
Article No. X
PP XX-XX

VIVA Institute of Technology
9thNational Conference on Role of Engineers in Nation Building – 2021 (NCRENB-2021)

VIVA-TECH IJRI
ANNUAL RESEARCH JOURNAL

VIVA-TECH INTERNATIONAL JOURNAL FOR RESEARCH AND INNOVATION
ANNUAL RESEARCH JOURNAL
ISSN(ONLINE): 2581-7280

A Deep Learning Model For Crime Surveillance In Phone Calls.

Abhishek Khaire[1], Pranav Mahadeokar[2], Praveen Kumar Prajapati[3], Prof. Sunita Naik[4]

[1](Department of Computer Engineering, Mumbai University, India)
Email: 18301062abhishek@viva-technology.org
[2](Department of Computer Engineering, Mumbai University, India)
Email: 18312069pranav@viva-technology.org
[3](Department of Computer Engineering, Mumbai University, India)
Email: 18301072praveenkumar@viva-technology.org
[4](Department of Computer Engineering, Mumbai University, India)
Email: sunitanaik@viva-technology.org

Abstract: *Public surveillance Systems are creating ground-breaking impact to secure lives, ensuring public safety with the interventions that will curb crime and improve well-being of communities, law enforcement agencies are employing and deploying tools like CCTV for video surveillance in banks, residential societies shopping malls,markets,roads almost everywhere to detect where and who was responsible for events like robbery, rough driving, insolence, murder etc. Other surveillance system tools like phone tapping is used to plot an event to catch a suspected person who is threatening an innocent or conspiring a violent activity over a phone call which is being done voluntarily by authorities . Now analyzing both the scenarios in the case of CCTV the crime(robbery,murder) has already occured and in the case of phone tapping there must be information in advance about the suspected person who is going to commit a crime, Now to overcome this issue a system is proposed to know in advance who is suspected and what conspiracy is being done over phone calls as well to detect it automatically and report it to law enforcemet authorities.*

Keywords - *Deep learning ,Phone calls, Audio analysis ,Surveillance Systems,Threats*

I. INTRODUCTION

Mass Surveillance practice began in world war 1 over 100 years ago with domestic spying of foreign officials residing in different countries conducting surveillance, scandals and transferring of scurrilous information have been key weapons of rival countries in WW1, WW2, cold wars and even now. Back then it was the fight of countries, but now it is fight against criminals, terrorists, sleeper cells residing within the country. Situations and scenarios like bank robbery, terrorist attacks, communal riots are unpredictable and adversely involve a large group of innocent victims .Offences like kidnapping ,rapes, threats can happen with anyone, Now here mass surveillance comes into picture with common community policing strategies. This helps law agencies to tackle the issue by tracking suspects, integrating camera systems, deploying officers, initiating neighbourhood crime watch and Mapping evolving crime patterns. Majority of planning, threats of such crimes occur over phone calls. To detect this a surveillance systems is proposed incorporating Deep learning, to train the system with harsh words and threatening high pitch tone, enabling to detect mishap before its occurrence. This will prevent potential crime from happening and save lives. Surveillance data has mostly concentrated on video analysis, but here focus on audio analysis is done for the same. Audio analysis can take us closer to semantics than video analysis could and also is computationally more efficient.

II. RELATED WORKS

Philip Duncal et.al[1], describes about information retrieval, minimizing the development cycle, identifying feature spaces, signal cleaning algorithms and pattern recognition methods, classification and segmentation of soundtracks present similar challenges. These terminologies have substantial overlaps. They are essentially pattern recognition tasks at different levels. Much of the literature deals with the development of these algorithms in a combined manner, while fine-tuning algorithms to carry out different tasks. Some of the methods used for recognition employ classification algorithms including support vector machine (SVM) as a binary classifier , Hidden Markov Machine (HMM) which works on the bases of prediction using previous data values, Artificial Neural Network (ANN) which is trained in a supervised learning regime to map features/descriptors onto predetermined classes and decision trees. These are the most commonly used types of classifiers, and may be employed individually and in combinations thereof.Support Vector Machine (SVM) and Hidden Markov Model (HMM) were used to classify broadcast news audio into 6 classes: silence, pure speech, music, environmental sound, speech over music, and speech over environmental sound classes

Eduard frank, et.al [2],the architecture is an adaptation of an image processing CNN, programmed in Python using Keras model-level library and TensorFlow backend. The concept lays the basis of the classification of emotions based on voice parameters is briefly presented. According to the obtained results, the model achieves the mean accuracy of 71.33% enlisting six emotions (fear, sadness, happiness,anger,disgust,surprise), which is comparable with performances reported in scientific literature. A person's speech can be altered by various changes in the autonomic nervous system and affective technologies can process this information to recognize emotion. As an example, speech produced in a state of fear, anger, or joy becomes loud and fast, with a higher and wider range in pitch, whereas emotions such as sadness or tiredness generate slow and low-pitched speech. Some emotions have been found to be more easily computationally identified, such as anger and happiness

Regunathan Radhakrishnan, et.al[3],this is a surveillance system for event detection cannot completely rely on a supervised audio classification framework. In this paper, authors have proposed a hybrid solution that consists of two parts; one that performs unsupervised audio analysis and another that performs analysis using an audio classification framework obtained from off-line analysis and training. The proposed system detects new kinds of suspicious audio events which are outliers against a background of usual activity. It adapts via Gaussian Mixture Model(GMM) to model background sounds and updates the model incrementally as latest audio data arrives. New types of suspicious events can be detected as deviants from this usual background model. The results on data are promising surveillance systems based on an audio classification framework would only be able to detect known kinds of suspicious activity like banging sounds and screaming. However, it is also important to detect suspicious events that the system has not seen before. Towards that end, the author proposes a hybrid solution that consists of two parts; one that performs unsupervised audio analysis and another that performs analysis using an audio classification framework obtained from off-line analysis and training.

Arpit Shah, et.al[4], paper discusses a system in which an audio file of conversation between two people from the EXOTEL (cloud telephony platform) dataset was taken as the basic input which was then separated into different segments using VAD (voice activity detection). During VAD, It involves removing the non-speech data like noise, pause-in between the conversation to separate the segments into different chunks with voiced and unvoiced data. Since audio may involve different numbers of speakers, used agglomerative clustering. It does segmenting of data into many small pieces. Each piece is assigned to a cluster. Audio is segmented depending on the pauses that are present in the conversation, but if there is an overlap between two people's conversation then audio separation does not take place properly which in turn affects the clustering process and furthermore miss-classifies the sentiments of the chunks as well. So instead of using a sliding video and only relying on the pauses in the conversation, future scope would be to build an unsupervised model (Auto-encoders) that performs VAD detection and further cleans out the segments with non-homogeneous data.

Cynthia Van Hee et.al [5], the Cyberspace is one of humanity's great inventions that bring great benefits but also exposes us to cyber threats. Cyberbullying is common to many on social platforms. In this paper authors have proposed a framework to detect cyberbullying messages in the form of text data using deep neural networks and word embedding. They stack together the state-of-the-art Bert and Glove embedding to improve the performance of the classifier. As a result, the itl outperforms the traditional machine learning methods such as Logistic Regression and SVM . To perform specific word detection ,they did Initial Word Embedding.

Word embedding is the process of representing each word as a real value vector. The embedding layer of there models processes a fixed length sequence of words. In this study three methods are used for initializing word embeddings: random, GloVe and SSWE. Using words embeddings during training improve the model to learn task specific word embeddings. Task specific word embeddings can differentiate the style of cyberbullying among different online platforms as well as topics.

Hao Hu, et.al [6], the GMM supervector based SVM is applied to this field with spectral features. A GMM is trained for each emotional utterance, and the corresponding GMM supervector is used as the input feature for SVM. Results on emotional speech database demonstrate that GMM supervector based SVM outperforms standard GMM on speech emotion Recognition. Since the gender-dependent emotion recognition system is preferred, they analysed the confusion between different emotions in condition of separate-gender subject System proposes solution via Confusion matrix of GMM supervector based SVM for female subject and male subject individually constituting 5 emotions anger; fear; happiness, neutral, sadness

Pavol Harár Radim ,Malay kishore Dutta et.at [7], in this paper the author has developed a method for Speech Emotion Recognition (SER) using Deep Neural Network (DNN) architecture with convolutional, pooling and fully connected layers. They used 3 class sets (angry, neutral, sad) of German Corpus containing 271 labelled recordings with a total length of 783 seconds. Audio files were split into 20ms segments without overlap. Voice Activity Detection (VAD) algorithm used to eliminate blank segments and divided all data into training (80%)validation (10%) and testing(10%) sets. DNN is optimized using Stochastic Gradient Descent. As input authors used raw data without and feature selection. The trained model achieves test accuracy of 96.97%. The model works well, the model does not have any pre-given context, still it gives better results. According to the author, providing context can improve the efficiency to the next level. This model can be used to find out whether the person speaking is angry or happy. So that it can be sure that what is the intention of any person detected speaking wrong suspected words.

Xavier Anguera Miro and Nicholas Evans et.al [8],Speaker diarization is the task of determining "who spoke when?" in an audio or video recording that contains an unknown amount of speech and also an unknown number of speakers ,The analysis of speaker diarization performance as reported through the NIST Rich Transcription evaluations. System incorporates HMM- UBM models and also Speech activity detection (SAD) which involves the labeling of speech and nonspeech segments. SAD can have a significant impact on speaker diarization performance , Author discusses importance of overlap detection incase where simultaneously multiple speakers are talking.

III. METHODOLOGY

The proposed system is divided into three stages: The stage one is called speaker diarization. Speaker diarization is the process of identifying caller1 and caller2 in a call and separating them into segments. Speaker diarization works in three steps, first is voice activity detection(VAD), next is speaker change detection, and lastly speaker clustering the results to get caller1 and caller2 segments. We have trained a deep neural network model to identify the sentiment of the conversation. RAVDESS speech dataset is used for classification. Acoustic features such as MFCC, STFT, Contrast, Mel Spectrum, Chroma and Tonnetz are extracted from the audio clips of the dataset. We stack all the features to get a feature vector of 193 dimensions. The extracted features are passed to a DNN composed of 3 hidden layers with the number of perceptrons to be 128, 128, 128 with RELU activation for all 3 hidden layers, 2 dropout layers of 0.2. Last layer being the softmax layer and adadelta optimizer. The model trained in the Training phase is used to test the sentiment of the chunks generated after Speaker diarization. Every chunk will yield sentiment which can help us understand the sentiment of the customer throughout the conversation. Next stage is to identify threat keywords from the source audio. For this we use google speech API to convert audio to text. Then we use the NLTK library to clean the text. Cleaning involves removing punctuations, stopwords, unnecessary blank space. This gives us a vector of cleaned words used in the conversation. We compare this cleaned word vector with another vector of most commonly used threat keywords. Thus if a keyword from cleaned words vector is present in the threat keywords vector, we identify it as a threat keyword in the source audio. If the sentiments in the audio call are associated with threat and threat keywords are identified in the conversation then it can be classified as a threat call. Fig 1. shows the system flow diagram.The source audio file is first passed through a speaker diarization system. After speaker diarization each segment is passed through DNN to find the sentiment of each segment. Next the source audio file is transcribed into a text file. This text file is then sent to a TextSentimentAnalyser which checks if there are any threat keywords or not and if there are which threat keywords are present in the conversation. In the source audio if there has been any sentiment

associated with threat such as angry, fearful, etc. and at the same time if there were threat keywords in the same source audio, the conversation will be classified as threat call and an alert will be sent to authorities.

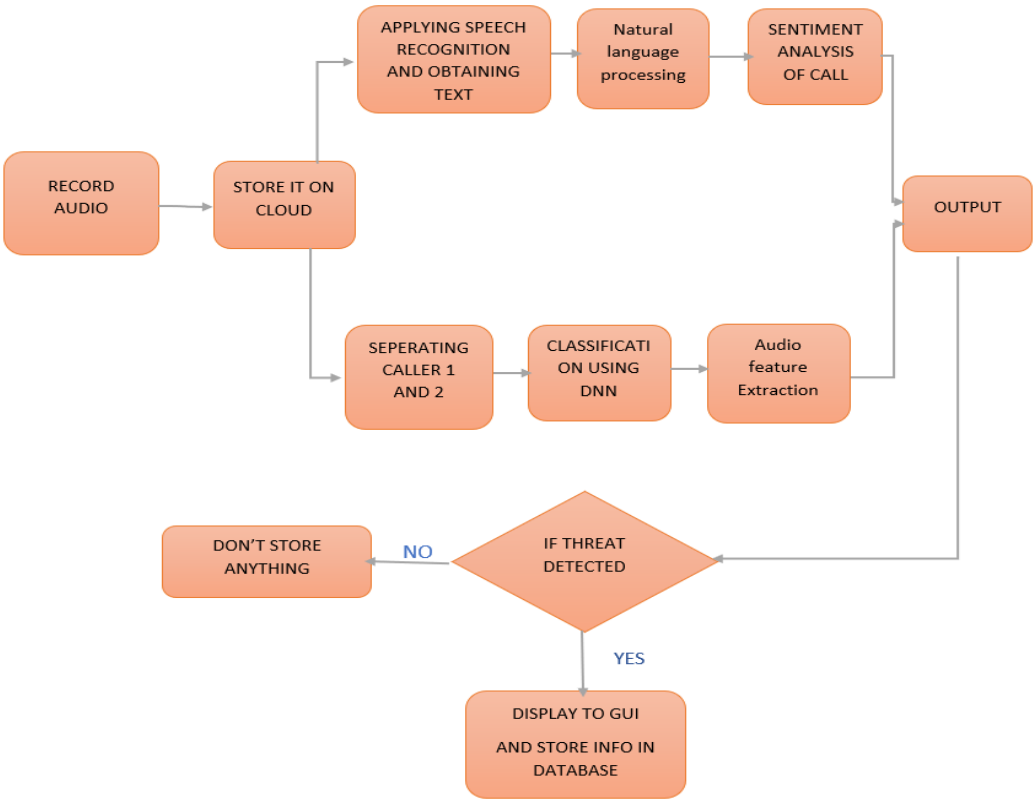

Figure 1: system flow diagram.

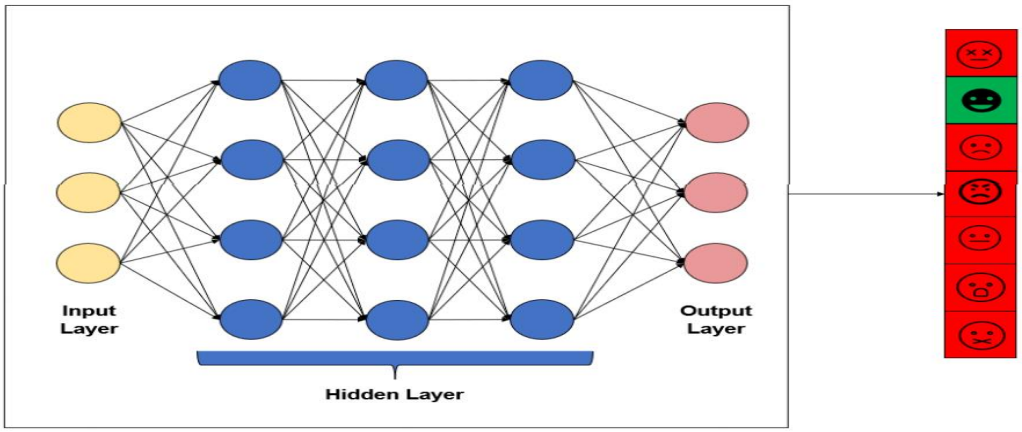

Figure 2: sentiment classification

IV. FIGURES AND TABLES

Table 1: Result analysis table

Model	Training	Test
CNN	90%	91%
XGBOOST	67.45%	45%
DNN	97%	93%

Table 1 shows the accuracy on training and test data of sentiment analysis using different models on 8 emotions(Clam, Happy, Sad, Angry, Fearful, Surprised, Neutral, Disgust)

Table 2: Precision, Recall and F1-scores for DNN model

Emotion	Precision	Recall	F1-Score
Angry	0.93	0.97	0.95
Calm	0.94	0.95	0.94
Disgusting	0.97	0.95	0.96
Fearful	0.88	1.0	0.93
Happy	0.89	0.92	0.91
Sad	0.96	0.86	0.91
Surprise	0.92	0.87	0.89

Table 2 shows the Precision, Recall and F1-scores for each emotion using the DNN model.

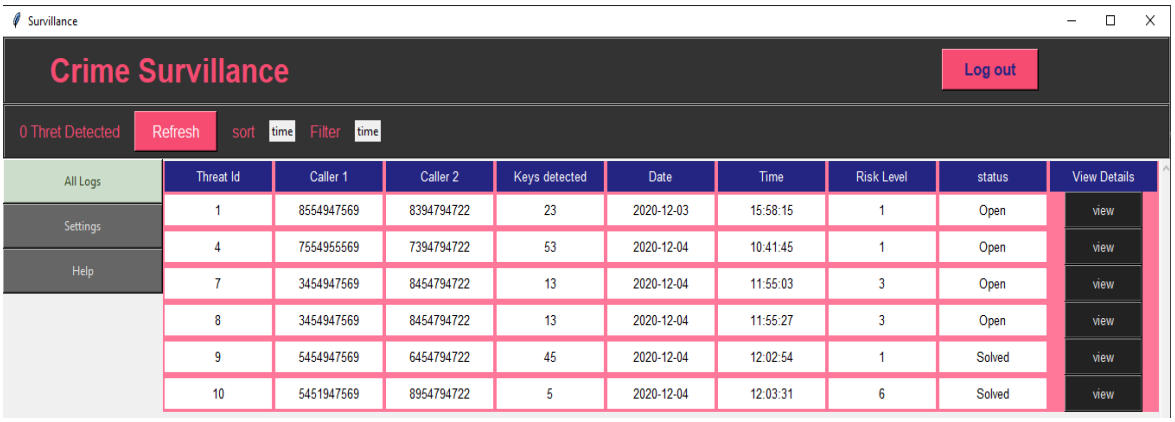

Figure 3: GUI Surveillance system.

Fig. 3 shows the GUI of the updated main page of the application, the page lists detected threats along with threat id, caller 1, caller 2, priority, date etc. this is connected to the database.

V. CONCLUSION

The proposed system will detect suspicious threat activity on phone calls, the system is capable of extracting tones of a person like anger, shouting, abusing, harassment, blackmailing and also when people use keywords like bomb, blast, murder, killing in english and hindi etc. A detailed report is generated indicating threat or not threat and reports to police authorities enabling them to detect mishap or crime before its occurrence. This project can be used by law enforcement agencies to trace criminals and prevent major incidents. This will also eliminate the need to manually listen to calls for threats. Giving more importance to privacy, the system will be very confidential, secure, encrypted and will not save any details if threats are not detected. The system will prevent potential crime from happening and save lives.

Acknowledgements

We would like to express a deep sense of gratitude towards our guide Prof.Sunita Naik, Computer Engineering Department for her constant encouragement and valuable suggestions. The work that we are able to present is possible because of her timely guidance. We would like to pay gratitude to the panel of examiners Prof. Sunita Naik, Prof. Dnyaneshwar Bhabad, & Prof. Vinit Raut for their time, effort they put to evaluate our work and their valuable suggestions time to time. We would like to thank Project Head of the Computer Engineering Department, Prof. Janhavi Sangoi for her support and coordination. We would like to thank the Head of the Computer Engineering Department, Prof. Ashwini Save for her support and coordination. We are also grateful to the teaching and the non-teaching staff of the Computer Engineering Department who lend their helping hands in providing continuous support.

REFERENCES

[1] Philip Duncal, Duraid Mohammed " Audio information extraction from arbitrary sound recordings" Audio Engineering Society (AES) ,2014 , 22nd International Congress on Sound and Vibration (ICSV22) and AES 136th Convention
[2] Eduard frantic ,Monica Dascalu , "Voice Based Emotion Recognition with Convolutional Neural Networks for Companion Robots" ,(ROMANIAN JOURNAL OF INFORMATION SCIENCE AND TECHNOLOGY)Volume 20, Number 3, 2017, 222–240.
[3] Regunathan Radhakrishnan, Ajay Divakaran, "Audio Analysis for surveillance applications for elevators" Mitsubishi Electric Research Labs 2005 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics.
[4] Arpit Shah,Shivani , Firodiya "Audio Sentiment Analysis after a Single-Channel Multiple Source Separation" Indiana University Bloomington,2019.
[5] Cynthia Van Hee , Gilles Jacobs "Automatic Detection of Cyberbullying in Social Media Text" AMiCA (Automatic Monitoring of Cyberspace Applications), 2018, rXiv:1801.05617v1 [cs.CL]
[6] Hao Hu, Ming-XingXu, and Wei Wu, "GMM Supervector Based SVM With Spectral Features For Speech Emotion Recognition" The international Conference on Acoustics, Speech, & Signal Processing (ICASSP),IEEE ,2007,1-4244-0728-1/07
[7] Radim Burger and Malay Kishore Dutta "Speech Emotion Recognition with Deep Learning" 4th International Conference on Signal Processing and Integrated Networks (SPIN) , IEEE ,2017, 978-1-5090-2797-2/17/
[8] Xavier Anguera Miro and Nicholas Evans "Speaker Diarization: A Review of Recent Research" IEEE Transactions On Audio, Speech, And Language Processing, IEEE Vol. 20, No. 2 ,February 2012 ,1558-7916/
[9] James Albert Cornel, Carl Christian Pablo,"Cyberbullying Detection for Online Games Chat Logs using Deep Learning," IEEE , 2019, 978-1-7281-3044
[10] Alexandru Lucian Georgescu , Horia Cucu "Automatic Annotation of Speech Corpora using Complementary GMM and DNN Acoustic Models," IEEE/ACM Transactions on Audio, Speech, and Language Processing, IEEE, 2018 , 978-1-5386-4695
[11] Fayek, H. M., M. Lech, and L. Cavedon. "Towards real-time speech emotion recognition using deep neural networks." Signal Processing and Communication Systems (ICSPCS), IEEE, 2015, 978-1-4673-8118
[12] Eesung Kimand ,Jong Won Shin, "DNN-based emotion recognition based on bottleneck acoustic features and lexical features", ICASSP ,IEEE, 2019, 978-1-5386-4658/
[13] Starlet Ben Alex and Ben P. Bab,"**Utterance** Syllable Level Prosodic Features for Automatic Emotion Recognition", Recent Advances in Intelligent Computational Systems (RAICS), IEEE,2018