Human Scream Detection and Analysis to ControlCrime Rate using Machine Learning

Sk.Kowsar
B.Tech Student,
Department of ECE,
RVR&JC College Of Engineering
Guntur,India
kowsarshaik24@gmail.com

V.A.Abhinaya
B.Tech Student,
Department of ECE,
RVR&JC College Of Engineering
Guntur,India
aarthiabhinaya17@gmail.com

N.Gowthami
B.Tech Student,
Department of ECE,
RVR&JC College Of Engineering
Guntur,India
gowthaminettyam@gmail.com

V.Manohar
B.Tech Student,
Department of ECE,
RVR&JC College Of Engineering
Guntur,India
veeravallimanohar152@gmail.com

***Abstract*—Public safety is significantly hampered by delayed police response due to a lack of accurate and timely information about crimes. Human scream detection using audio classification offers a promising solution. This work presents a novel three-phase scream detection system leveraging a K-Nearest Neighbors (KNN) classifier and a Multilayer Perceptron (MLP) model. The system first separates human distress sounds from background noise using MFCC features and KNN. Subsequently, it differentiates screams from shouts within the distress category using another KNN classifier. Finally, the classified screams trigger emergency notifications sent to the police station via the Twilio library. Our proposed system offers a robust and layered approach to scream detection, potentially enhancing response times and improving public safety.**

***Index Terms*—MFCCs, KNN, Multilayer perceptron model, Scream Detection.**

I. Introduction

Smartphones are no longer just communication devices their increasing processing power and advanced sensors have opened doors to new applications. One promising area is their potential role in personal safety. This study investigates the feasibility of using smartphones to detect screams and shouts, vocal cues that can signal distress. By leveraging a smartphone's built-in microphone and machine learning algorithms, the goal is to create a system that can distinguish these critical sounds from everyday background noise.

This research investigates the potential applications of scream and shout detection in various fields, including home security, healthcare for the elderly, and other critical areas. To achieve this, we built a system that can distinguish these vocalizations from background noise. The dataset for this model is collected from online platforms like kaggle, github and capturing real-life screams and sounds recorded during daily activities. For accurate identification, a three-stage approach is employed. First, a K-Nearest Neighbour (KNN) classifier with Mel-frequency cepstral coefficients (MFCC) as features separates screams, shouts, and speech from noise. Subsequently, the remaining speech sounds are differentiated from screams and shouts in the second phase. Finally, the third stage focuses on isolating screams from shouts within the identified non-speech category. To further enhance accuracy, the system leverages a Multilayer Perceptron model.

After passing the recorded audio to Multilayer Perceptron model, If the scream is detected the a Notification with your mobile number is being sent to police station using Twilio library. As the notification is reached to the police , then they can track the user location and save their lives.

II. Literature survey

Scream and shout detection has gained significant research interest in recent years due to its wide range of real-world applications. Prior studies have primarily focused on either feature extraction techniques or acoustic modeling for scream detection.

Existing research on scream and shout localization includes studies like [2], where researchers used microphones and Time Difference of Arrival to differentiate screams and gunshots from background noise. They achieved a 93% accuracy with a 5% false alarm rate using two parallel Gaussian Mixture Models (GMMs). Similarly, [3] explored scream, explosion, and gunshot identification in a subway environment.

Several studies have addressed broader event classification tasks that may include screams or cries. For example, [4] proposed a model to classify events like glass breaking, water flowing, screams, and cries, although their work focused solely on non-distress data. In [5], researchers used ten audio recordings encompassing screams, door slams, speech, cries, claps, laughter, knocks, explosions, phone rings, and glass breaking, all recorded in noisy environments.

In [6] they employed analytical features followed by SVM classification for event identification, while [7] utilized MFCC, MPEG-7 features, and Hidden Markov Models (HMMs) for gunshot and scream classification. A parallel GMM classifier network achieved a reported 90% accuracy with an 8% false alarm rate in [8].

Shout detection methodologies have also been explored in studies like [9] and [10], considering the presence of phonetic structure in shouted speech compared to pure screams.

Our work shares similarities with [11], which employed two-stage supervised learning for human cry and scream detection.

Our approach distinguishes itself by utilizing a combination of a pure KNN-based classifier and a Multilayer Perceptron Model for scream and shout detection from speech and noise. While [1] explored T2-statistics and Bayesian Information Criteria, and [12] used SVM with Gaussian Mixtures for scream detection, our three-stage KNN classification employs a novel filtration technique for noise removal in phase 1, speech detection in phase 2, and finally, scream detection from shouts in phase 3. Furthermore, our model incorporates real-life scream recordings from volunteers for enhanced generalizability.

III. Proposed Methodology

A. Data Collection

The given Fig. 1 describes the two phases of data collection. In the first phase, collect the shout, scream, and speech sounds embedded with environmental noises (like factory noises, vehicle noises, etc.). These sounds are collected from the internet and movies and some online platforms like GitHub. In the second phase of data collection, three volunteers (3 males) recorded distress sounds (scream, shout, etc.) using the Smartphone's microphone as they approached their day-by-day schedules. The volunteers were not given any directions concerning dealing with the phone's mic while it was recording. They used the phone's mic as they did in their day-to-day lives. All audio datasets are encoded at 16 bits per sample.

The recordings are done in a controlled way. The objective is to record sound samples with high constancy. The phone's mic is the only source used for making these recordings. All the time of recording, the phone is placed in the hands of volunteers. A total of three volunteers (only 3 males) collected data during the subsequent stage.

B. Testing and Training sets

All gathered sounds split into as long as five to ten second-long audio samples. They scream and shout sounds are split into 5s. Whereas audio recordings of speech and noise are a little bit lengthier, the normal length of a shout recording in the misery data gathered in the main stage was around 3s.

Scream, shout, and cry audio sets obtained from the main stage and the subsequent stage are utilized to make the training set and the testing set. These samples in the two sets are totally unrelated. The test set comprises of the apparent multitude of tests acquired from the second phase of data assortment.

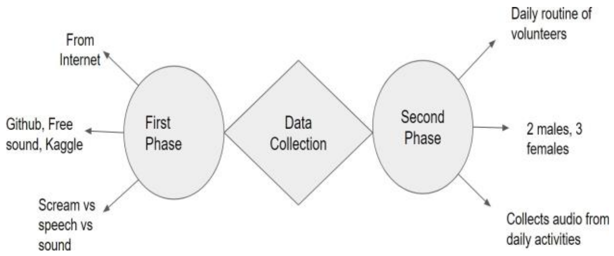
From Internet
Github, Free sound, Kaggle
Scream vs speech vs sound
First Phase
Data Collection
Second Phase
Daily routine of volunteers
2 males, 3 females
Collects audio from daily activities

Fig. 1. Different phases of collecting audio.

C. K-Nearest Neighbour

This system aims to reduce crime by identifying screams within real-time audio streams. It leverages K-Nearest Neighbors (KNN) classification, a powerful technique for categorizing audio events like screams and shouts. Unlike Support Vector Machines (SVMs), KNN avoids the complexity of kernel functions, making it a simpler approach.

The audio samples in the training, validation, and testing sets are converted to Mel-frequency cepstral coefficient (MFCC) as feature vectors before they can be classified by the three-phase method. Feature extraction is carried out in the following manner. For the last 3 decades, Mel-Frequency Cepstral Coefficients (MFCC) is broadly utilized in numerous speech identification frameworks. The major feature used that one audio differs from the second audio is Mel-Frequency Cepstral Coefficients (MFCC). Initially for the MFCC extraction, the continuous speech goes for frame blocking and windowing. Frame Blocking divides the signal into short frames or segments of 20–30 ms. Every frame is then multiplied by the hamming window to retain the permanency of initial and final points in the frame. Then this transforms to Fast Fourier Transform to get the magnitude frequency of each frame. Then it continues to the Mel-scale filter bank and then the Mel-frequency spectrum continues to triangular bandpass filters just to compute the logarithmic function. To get the final MFCC they pass energy (Ek) through Discrete Cosine Transform(DCT).

The equation is given by:

$$C_m = S_k = \frac{1}{N} \cos \frac{m \times (k - 0.5) \times \pi}{N} \times E_k$$

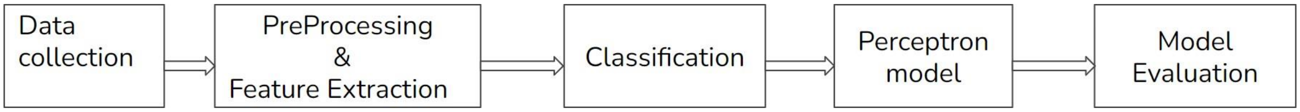

Fig. 2. Block diagram of the proposed methodology for Classification of Scream

where m = 1, 2, . . . , L (1)
N number of triangular bandpass filters and
L number of MFCC’s

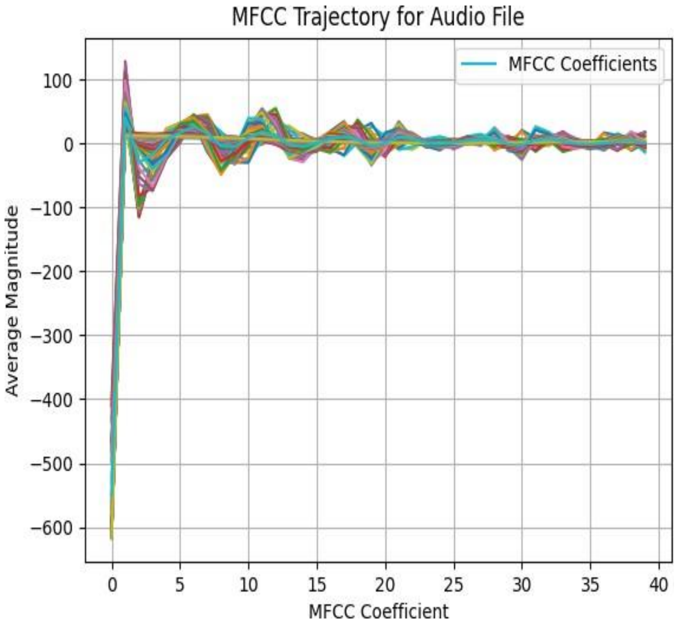

Fig. 3. Mel-Frequency Cepstral Coefficients at different echos .

Fig 2. outlines the Mel-Frequency cepstral coefiicients at different echos and later we mean all Mfccs into a single featured vector which is represented in fig.3

To train the KNN model, we utilize a pre-labeled dataset containing audio clips categorized as noise, screams, and shouts. The KNN algorithm learns from these examples by storing their corresponding MFCC features and class labels. This essentially creates a reference library for the model to use during the classification stage.

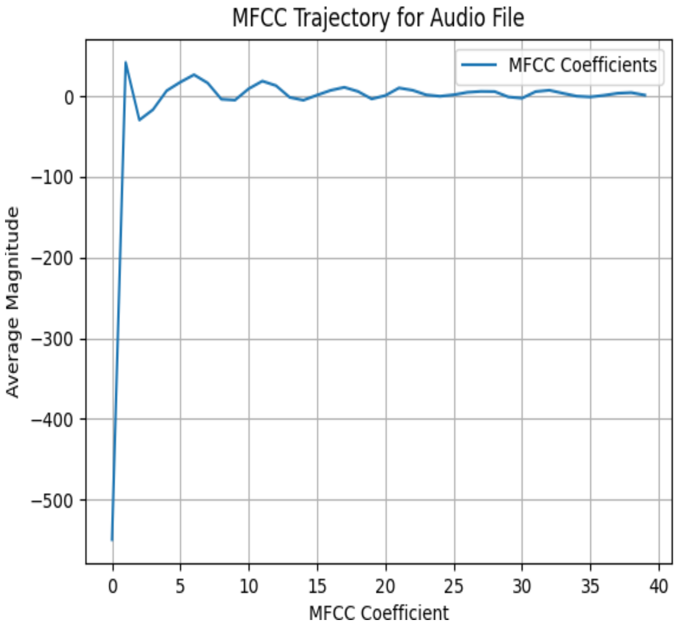

Fig. 4. Single featured Vector of the Mfcc’s.

When a new audio clip arrives, the system first extracts its MFCC features. The KNN algorithm then identifies the k nearest neighbors within the training data. These neighbors represent the k audio clips in the training set that exhibit the most similar MFCC features to the new clip. In simpler terms, the KNN algorithm finds the closest matches to the new audio clip based on its sound characteristics.

The final step involves predicting the class label (noise, scream, or shout) for the new audio clip. The KNN algorithm assigns the majority class label among the k nearest neighbors to the new clip. If, for instance, most of the k closest neighbors are classified as screams, then the new clip is also classified as a scream. This approach leverages the ”wisdom of the crowd,” where the most frequent class label among the similar sounding clips is considered the most likely classification for the new audio sample.

KNN offers several advantages. Firstly, it excels in real-time classification due to its relatively simple implementation. Additionally, the model can adapt to new data without requiring a complete retraining process. However, KNN also has limitations. Because it stores the entire training dataset for comparison, it can be memory-intensive for very large datasets. Furthermore, choosing the optimal value for k, the number of nearest neighbors considered, is crucial for achieving accurate classification. Finding the right balance for k ensures the model considers the most relevant neighbors for effective audio clip categorization.

D. Multilayer Perceptron Model

The sample audio sets come in this model and the input size is fixed for this model which is defined at the training time all the sounds gets loaded in the data frame and since sounds can be of different length this data frame gets arranged concerning the smallest sound because every deep learning-based model needs input layer to be of fixed size hence after adjusting data frame all sounds will be of same dimensions. Hence, define the input layer input size. After the input layer has five more layers overall there are six layers (including the input layer). All these layers are dense layers that are interconnected to each other. Except for the output layer, each layer has the activation function 'ReLU' because this is a widely used activation function. In the last layer, an activation function 'sigmoid' is used because it exists between (0 and 1). Therefore, it is used for those models where it is required to predict whether an event will occur or not. Since the probability of everything exists between 0 and 1 hence sigmoid function was the better choice. The Sigmoid Function curve looks like an S-shape. In the first layer, consider 12 perceptrons, in the second layer consider 8 perceptrons, in the third layer consider 10 perceptrons, in the fourth layer consider 5 perceptrons, in the fifth layer consider 3 perceptrons, and in the last layer consider 1 perceptron. This type of configuration is used in this model because only this configuration is providing the best learning rate and accuracy till now. The current accuracy of this model is 89.333% with a 5% false rate.

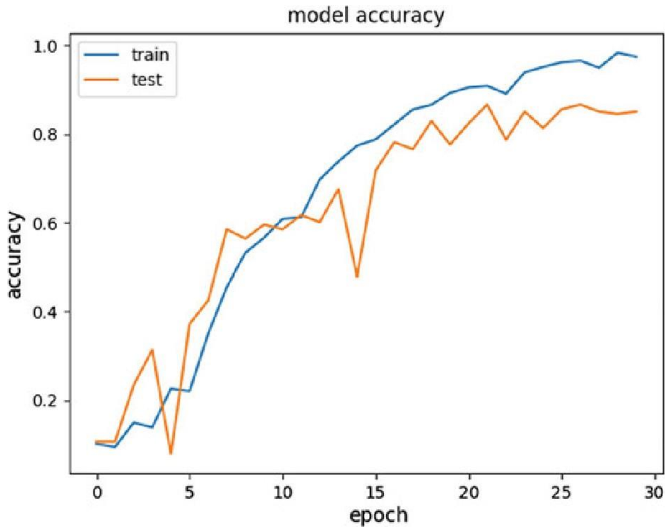

Fig. 5. Accuracy of testing data in multilayer perceptron model of different epochs.

For accuracy calculation, load all the datasets and all rows of the dataset are shuffled and used 400 audio datasets for audio files. All the types of sounds that are a scream, shout, speech, and noise is used for 33% of the dataset as test data and rest of other are using for training purpose. Using the evaluate method, the python library is used to calculate the accuracy of trained files which results in 93%, and later the same method is used for calculating the accuracy of test data which results in 82%. Figure 6 shows the flowchart for this model.

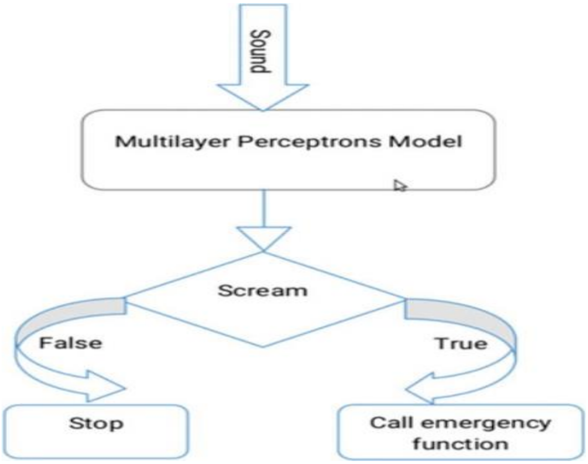

Fig. 6. Approach using in multilayer perceptron model.

Result Analysis

This work investigates a human scream detection system that combines a three-stage learning model with a final Multilayer Perceptron model. This approach effectively detects screams in diverse real-world scenarios with background noise, including human conversations, various machinery sounds (both indoors and outdoors), and human gatherings. To ensure the model's ability to handle such complexities, we meticulously crafted a training dataset. This dataset was built through a two-phase collection process, incorporating a wide range of audio samples. The model's effectiveness was further validated using audio sets recorded by volunteers during their daily activities, simulating real-world scream capture with smartphone microphones. The combined model achieved a remarkable accuracy of 90%.

We specifically chose the Multilayer Perceptron model due to its superior learning capabilities compared to K-Nearest Neighbors (K-NN) and logistic regression models. While both K-NN and logistic regression initially showed improvement during training epochs, their accuracy plateaued after a certain point. In contrast, the Multilayer Perceptron model demonstrated continuous learning throughout the training process. This sustained learning ability makes it the optimal choice for achieving the best possible performance in scream detection.

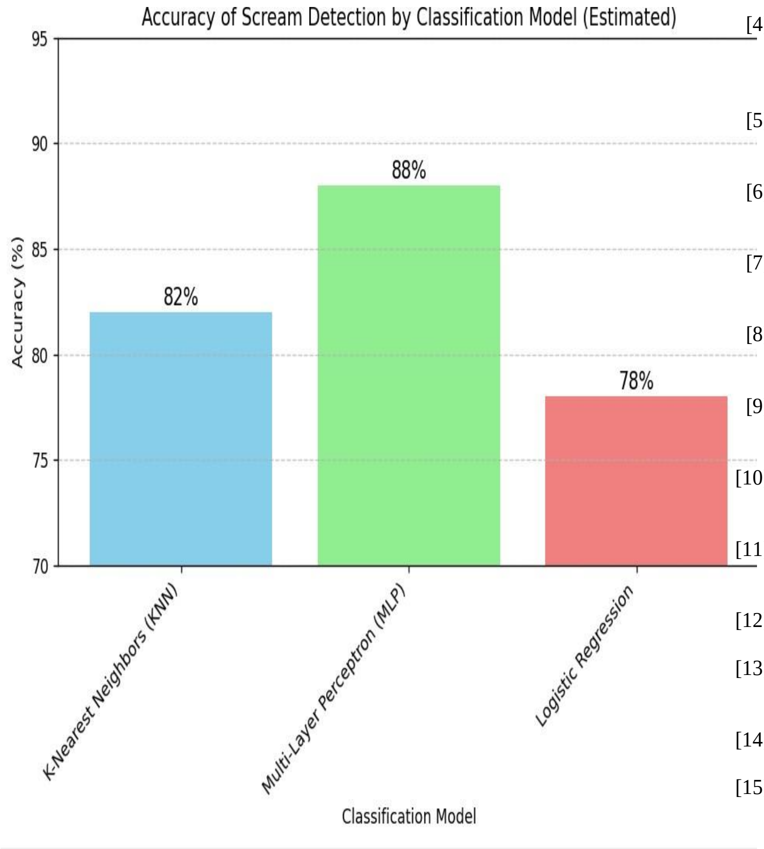

Fig. 7. Accuracy of different models.

CONCLUSION

This research demonstrates the potential of combining a three-stage KNN-based classifier with a Multilayer Perceptron model for scream and shout detection in noisy environments. This approach effectively distinguishes screams and shouts from speech and background noise. The inclusion of real-world scream recordings from volunteers during training further enhances the model's generalizability and robustness. These findings suggest that combining supervised learning techniques with deep learning architectures can be a promising approach for developing accurate and reliable audio classification systems.

REFERENCES

[1] M.K. Nandwana, A. Ziaei, J.H. Hansen, Robust unsupervised detection of human screams in noisy acoustic environments, in 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (IEEE, 2015), pp. 161–165

[2] G. Valenzise, L. Gerosa, M. Tagliasacchi, F. Antonacci, A. Sarti, Scream and gunshot detection and localization for audio-surveillance systems, in 2007 IEEE Conference on Advanced Video and Signal Based Surveillance (IEEE, 2007), pp. 21–26

[3] S. Ntalampiras, I. Potamitis, N. Fakotakis, On acoustic surveillance of hazardous situations, in IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 165–168 (2009)

[4] M.A. Sehili, B. Lecouteux, M. Vacher, F. Portet, D. Istrate, B. Dorizzi, J. Boudy, Sound envi- ronment analysis in smart home, in International Joint Conference on Ambient Intelligence (Springer, Berlin, Heidelberg, 2012), pp. 208–223

[5] H.D. Tran, H. Li, Sound event recognition with probabilistic distance SVMs. IEEE Trans. Audio Speech Lang. Process. 19(6), 1556–1568 (2010)

[6] W. Huang, T.K. Chiew, H. Li, T.S. Kok, J. Biswas, Scream detection for home applications, in 2010 5th IEEE Conference on Industrial Electronics and Applications (IEEE, 2010), pp. 2115– 2120

[7] S. Ntalampiras, I. Potamitis, N. Fakotakis, On acoustic surveillance of hazardous situations, in 2009 IEEE International Conference on Acoustics, Speech and Signal Processing (IEEE, 2009), pp. 165–168

[8] L. Gerosa, G. Valenzise, M. Tagliasacchi, F. Antonacci, A. Sarti, Scream and gunshot detection in noisy environments, in 2007 15th European Signal Processing Conference (IEEE, 2007), pp. 1216–1220

[9] J. Pohjalainen, P. Alku, T. Kinnunen, Shout detection in noise, in 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (IEEE, 2011), pp. 4968– 4971

[10] K. Mittal, B. Yegnanarayana, Production features for detection of shouted speech, in 2013 IEEE 10th Consumer Communications and Networking Conference (CCNC) (IEEE, 2013), pp. 106–111

[11] A. Sharma, S. Kaul, Two-stage supervised learning-based method to detect screams and cries in urban environments. IEEE/ACM Trans. Audio Speech Lang. Process. 24(2), 290–299 (2015)

[12] S. Chung, Y. Chung, Scream sound detection based on SVM and GMM, in RTET-17, 2017

[13] John H. L. Hansen, Mahesh Kumar Nandwana, Navid Shokouhi; Analysis of human scream and its impact on text-independent speaker verification. J. Acoust. Soc. Am. 1 April 2017; 141 (4): 2957–2967

[14] Engelberg JWM, Schwartz JW, Gouzoules H. 2019. Do human screams permit individual recognition? PeerJ 7:e7087

[15] J.F.P. Kooij, M.C. Liem, J.D. Krijnders, T.C. Andringa, D.M. Gavrila, Multi-modal human aggression detection, Computer Vision and Image Understanding, Volume 144,2016,Pages 106-120, ISSN 1077-3142,

[16] O'Donovan R, Sezgin E, Bambach S, Butter E, Lin S Detecting Screams From Home Audio Recordings to Identify Tantrums:Exploratory Study Using Transfer Machine Learning JMIR Form Res 2020;4(6):e18279

[17] R. Mathur, T. Chintala and D. Rajeswari, "Identification of Illicit Activities& Scream Detection using Computer Vision & Deep Learning," 2022 6th International Conference on Intelligent Computing and Control Systems (ICICCS), Madurai, India, 2022, pp. 1243-1250

[18] Q. Nguyen, S. -S. Yun and J. Choi, "Detection of audio-based emergency situations using perception sensor network," 2016 13th International Conference on Ubiquitous Robots and Ambient Intelligence (URAI), Xi'an, China, 2016, pp. 763-766,

[19] Schwartz, J. W., & Gouzoules, H. (2019). Decoding human screams: perception of emotional arousal from pitch and duration. Behaviour, 156(13-14), 1283-1307.

[20] Disa Anna Sauter, Martin Eimer; Rapid Detection of Emotion from Human Vocalizations. J Cogn Neurosci 2010; 22 (3): 474–481.