VOICEGUARD: A Real-Time Audio-Based Threat Speech Detection System

Stuti Kumar
Artificial Intelligence and Machine Learning
Meerut Institute of Engineering and Technology
stuti.kumar.cseaiml.2022@miet.ac.in
Meerut,Uttar Pradesh, India

Nishita Pal
Artificial Intelligence and Machine Learning
Meerut Institute of Engineering and Technology
nishita.pal.cseaiml.2022@miet.ac.in
Meerut,Uttar Pradesh, India

Vanshika
Artificial Intelligence and Machine Learning
Meerut Institute of Engineering and Technology
vanshika.pawan.cseaiml.2022@miet.ac.in
Meerut,Uttar Pradesh, India

Priyanshi Jain
Artificial Intelligence and Machine Learning
Meerut Institute of Engineering and Technology
priyanshi.jain.cseaiml.2022@miet.ac.in
Meerut,Uttar Pradesh,India

Mr. Kuldeep Kumar
Assistant Professor
Artificial Intelligence and Machine Learning
Meerut Institute of Engineering and Technology
Kuldeep.kumar.aiml@miet.ac.in
Meerut,Uttar Pradesh, India

***Abstract*—**Audio surveillance systems that already exist largely focus on recognizing abnormal/odd sounding noises such as gunshots, explosions & other similar unwanted noise patterns. However, most real-world threats start with people's aggressive tones of voice when they speak directly to someone, or when they communicate planning statements (example: "'give me the money") to coerce another person to give what they want, or when they are screaming out of fear. Current systems do not identify or track these spoken indicators & as a result can't detect when someone may potentially be in danger until after violence occurs.

This proposed system, named VoiceGuard, will provide real-time detection of threatening speech by analysing tone, aggression, emotional indicators and other contextual hints within raw audio data. The VoiceGuard system utilizes YAMNet as its primary means of generating feature embeddings for analysis and leverages a deep neural network classifier algorithm to differentiate speech that is classified as threatening from speech that is not threatening. The dataset used to develop & test this algorithm consisted of more than 3600 audio samples from a variety of sources, including custom recordings performed specifically for testing use of threatening phrases & datasets of emotional speech samples (i.e., RAVDESS, CREMA-D) & samples from various types of physically safe environments. The system achieved approximately (90%) Accuracy & (95%) Recall when detecting whether a given clip of audio contained threatening speech, which exceeds the performance of existing systems that are built using traditional MFCC/SVM based systems and that typically only identify non-verbal anomalies. Therefore, VoiceGuard provides contextual analysis of threatened speech and will be useful in settings such as: banks, public spaces and emergency monitoring systems.

***Keywords*—**Audio Surveillance, Threat Speech Detection, Deep Neural Network, YAMNet, Real-Time Audio Classification, Aggression Detection, Security Monitoring.

I. INTRODUCTION

Visual data through CCTV systems are heavily relied upon for modern security surveillance; however, video-based data has limitations related to poor visibility, crowds, or blind spots. Audio surveillance provides additional valuable context to the security environment by allowing detection of emotional indicators, level of stress, tone of voice, and spoken intent, which can't be detected through just visual data.

Current research on audio-based security has focused on detecting abnormal external environmental sounds such as gunshots, breaking glass, and screams; however, there are very few systems that focus on detection of threat-related speech, such as planning a robbery, using coercive commands, or exhibiting aggressive intent. When examined closely, almost every crime starts with a verbal cue, and only later will there be an escalation of assault.

VoiceGuard is a new research project to detect threatful-sounding speech using unprocessed audio signal data. The focus of the research is to analyze the manner in which a message is being communicated to detect if the person communicating may be a threat (through tone of voice, pitch, level of aggression, level of emotional excitement, etc.) and not simply the content of the message. By examining how something is communicated rather than what is being said, the VoiceGuard benefits from being a robust detection system even if the audio is distorted, noisy, or partially obscured.

VoiceGuard is able to create a real-time, small size system to listen for and detect threatful communications in live time by listening for and detecting threatful communications as they happen, creating high-level audio embeddings using Google's YAMNet deep feature extraction algorithm, and classifying the audio with the use of a DNN. The goal of the VoiceGuard is to create a system which allows for the early detection of threatful situations so that security personnel can respond quickly.

LITERATURE REVIEW

Existing work in audio anomaly detection focuses mostly on:

2.1 Environmental Sound Classification

Studies identify events like:

- ✓ glass-breaking
- ✓ gunshots
- ✓ screaming
- ✓ explosions
- ✓ vehicle collisions

Using MFCCs, spectrograms, and CNN/SVM classifiers. These systems fail in distinguishing threatening speech from normal conversation.

2.2 Emotion Recognition Systems

Research on emotional datasets (e.g., RAVDESS, CREMA-D) classifies:

- ✓ anger
- ✓ sadness

- ✓ fear
- ✓ happiness
- ✓ neutrality

However, emotional intensity ≠ threat. For example:
Angry speech may not be dangerous
Calm speech may still contain threatful intentions

2.3 Speech Recognition for Threat Detection

Some works attempt keyword spotting ("help", "fire", etc.). These approaches depend entirely on speech-to-text and fail when:

- ✓ audio quality is poor
- ✓ there is overlapping background noise
- ✓ accents vary
- ✓ microphone distance is large

2.4 Limitations in the Existing Models

The existing models focuses on:

- ✓ environmental anomalies
- ✓ scream/gunshot detection
- ✓ shock/abnormal sound events

But they do not identify spoken threats, planning statements, or emotional aggression linked to danger.

II. METHODOLOGY

The proposed methodology for developing the VoiceGuard system consists of five stages: dataset creation, data preprocessing, feature extraction, model learning, and real-time evaluation. The complete workflow has been designed to enable the reliable detection of threatening speech based solely on raw audio data. The following paragraphs detail the major steps that will be taken during implementation.

A.Dataset Development

Whereas traditional audio anomaly detection studies rely exclusively on the use of existing datasets, this project required the construction of a hybrid dataset consisting of both existing emotional speech corpora as well as recordings produced specifically for the current study.

The hybrid dataset used by this project contains the following:

- Emotional speech samples from two publicly available, emotional speech corpora (i.e., RAVDESS and CREMA-D) and includes neutral, sad, happy, angry, fearful, and distressed emotional prosody.
- Pre-recorded threat examples, a portion of which was contributed by the Idea Lab at the Meerut Institute of Engineering and Technology, covering a variety of vocal intonations and emotional intensities.
- Ambient recordings of safe environments from natural occurring environments (chatter, laughter, coughs, etc.) as well as non-threatening speech.
- All audio files created as part of the dataset have been converted into a standardized file type (WAV) at a standard sampling rate (16kHz), thereby ensuring standardization for all audio files during future processing. The final hybrid dataset will consist of approximately 3600 audio samples that have been categorized as being either Safe or Threat.

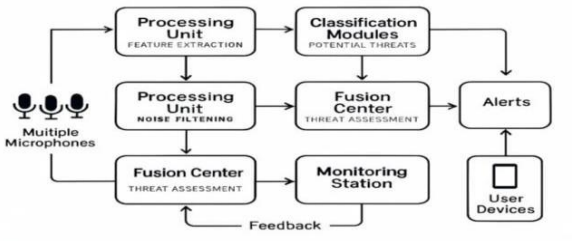

Fig. 1. System Architecture of Real-Time Audio Threat Detection

B.Preprocessing

To maintain reliability in feature extraction, a uniform preprocessing pipeline was applied to all files.

The steps included:

- Formating the audio files into a uniform format. All audio files were normalized to 16--bit PCM WAV files and mixed together.
- Resampling: The files were resampled to 16 kHz to match the YAMNet requirement.
- Normalizing: The audio files were normalized by leveling out the files to remove volume inconsistencies.
- Filtering: Noise and silence segments were removed from the beginning and/or end of the audio, resulting in only true audio segments being used to start the feature extraction process.
- Removal of Corrupt Files: Files that had a corrupted decoding, or files missing data packets were removed from these files.

All of these preprocessing procedures assured the feature extraction procedure began with audio that was clean, standard, and balanced.

C.Feature Extraction Using YAMNet

With the help of YAMNet, which is a Google-designed deep-learning neural network model trained using AudioSet, feature extraction was accomplished. When raw audio waveforms are passed through YAMNet, they will be converted into 1024-dimensional embedding vectors. The properties of the audio captured in the embedding include:

- The distribution of acoustic energy throughout the sound.
- The harmonics contained in the sound.
- The timbre defined by various sound waves.
- The emotional intensity or impulse factor contained in the sound.
- Tone and aggression.

YAMNet produces frame-level embeddings for each audio file that produce a mean value based on temporal progression. By aggregating this mean over several frames, the system produces a fixed representation of the feature data that can be used for downstream classification tasks. Overall, these features are significantly more robust than traditional MFCC features, particularly when it comes to speech examples that contain varying levels and/or transitions in emotion, or background noise.

D.Model Architecture and Training

A DNN classifier was built using the YAMNet extracted embeddings as features. The DNN architecture has four hidden layers constructed as follows: an input layer with 1,024 units representing the YAMNet features, followed by three dense layers having 512, 256, and 128 neurons each, where each neuron in these layers utilized the ReLU activation function and dropout regularization to avoid overfitting on the training data. Each of the layers' outputs will be used as input to the next layer until the output layer provides predictions for two classes (safe/threat) using the sigmoid activation function.

Training of the DNN classifer used binary cross-entropy as the loss function; the Adam optimizer was used during training to minimize the loss; and class weighting was used to reduce the impact of class imbalance (safe/threat) in the training data (80%-20% train/test split). When evaluated against various speech intensities, accents, and emotional expressions, the architecture of the DNN classifier showed a high degree of generalization performance.

Fig. 2. Waveform of a Dog Bark

Fig. 3. Waveform of a Gun Shot

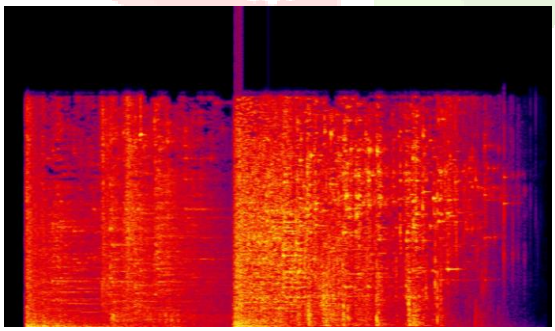

Fig. 4. Spectrogram of a Dog Bark

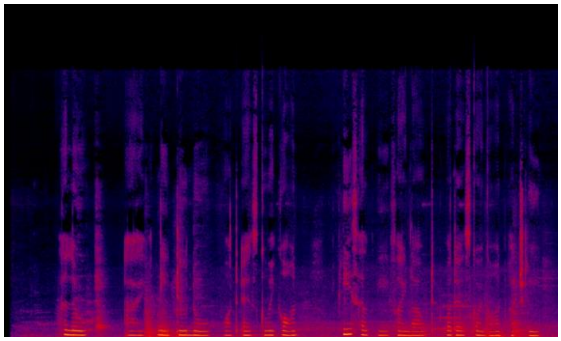

Fig. 5. Spectrogram of a Gun Shot

E-Systems Workflow

The complete VoiceGuard workflow is illustrated as follows:

(A) acquiring audio through either live microphone input or pre-recorded files

(B) preprocessing acquired audio by standardizing, normalizing, and controlling noise levels

(C) generating features through YAMNet-based embedding generation

(D) classifying whether the acquired audio is "Safe" or "Threat" via DNN-based classification technique

(E) producing an alert to a user in real-time once completed.

The entire process is designed to allow for low latency and is appropriate for real-world applications.

F. Real-Time Threat Speech Detection

Using sound & noise devices library to do this! Captures live audio up to 5 seconds long & quickly processes them via preprocessing & feature extraction before they are classified by a trained Deep Neural Network model.

Features include:

- Predictions nearly instantly
- Same classification regardless of how far away a user is or how they say a command (tonality)
- Applicable for real-world surveillance purposes

G.Comparison with Existing Methodologies

Unlike prior studies that have primarily relied on MFCCs, spectrogram convolution or keyword spotting, the new approach,

- utilizes contextual acoustic features through the use of YAMNet
- adds custom threat speech datasets to provide enhanced relevance to real-world settings
- provides substantially improved recall rates for aggressive/threatening scenarios
- supports real-time, end-to-end detection instead of only offline processing

III.RESULTS AND DISCUSSUIONS

The VoiceGuard Threat Detection System's proposal was validated via a custom dataset of 3,652 preprocessed audio clips divided into safe and threat classifications. The validation was performed via DNN classifiers using a YAMNet-created embedding (1024 dimensions). The results were compared to a traditional classifiers' performance as the basis of comparison for the improvement of our technique.

A. Performance Metrics

The system was assessed using standard metrics such as Precision, Recall, F1-score, and Overall Accuracy.

Table 1 shows the classification performance of VoiceGuard.

Class	Precision	Recall	F1-Score
Safe	0.98	0.89	0.93
Threat	0.76	0.95	0.85
Overall Accuracy	0.9	-	-

TABLE 1: Performance Metrics of VoiceGuard

The results indicate that the model exhibits high recall (0.95) for the threat class, ensuring that potentially dangerous scenarios are rarely missed. This is a crucial requirement in real-time surveillance systems.

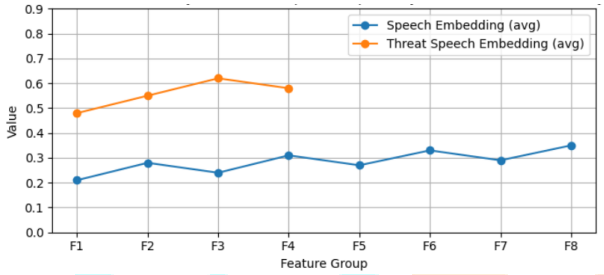

Fig. 6. Training and Validation Loss Curves of the voice guard DNN Model

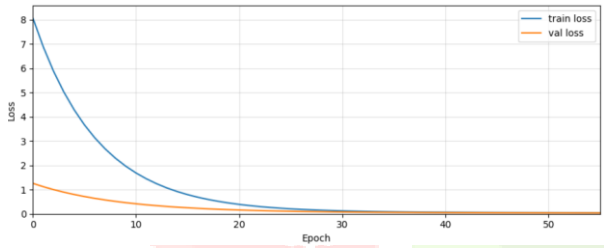

Fig. 7. Training and Validation Accuracy of the voice guard DNN Model

B.Confusion Matrix Analysis

The confusion matrix shows a clear distinction between correctly classified safe and threat samples.

	Predicted State	Predicted Threat
Actual Safe	470	59
Actual Threat	11	191

TABLE 1: Confusion Metrix of VoiceGuard

The system misclassified only 11 threat samples, which demonstrates the ability of VoiceGuard to detect suspicious or harmful speech patterns with high sensitivity.

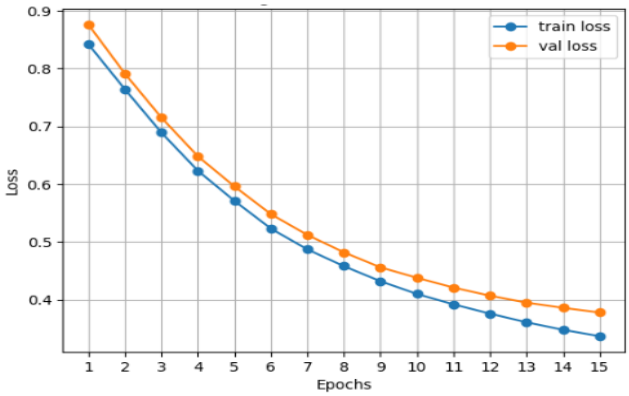

Fig. 8. Training and Validation Loss Curves of the voice guard DNN Model across Epochs

IV. CONCLUSION

VoiceGuard has shown it will be able to provide an advanced real-time security threat detection capability for speech that is capable of going beyond traditional abnormal sound detection techniques. YAMNet embeddings, when used with a DNN classifier, will capture emotion, aggression pattern detection and suspicious verbal cues from the raw audio. This capability will provide a high recall rate for threats so that any potentially harmful event will be detected as soon as it happens. The outcome will allow VoiceGuard, when connected with CCTV systems and/or IoT security systems, to offer an extremely useful security tool for banks, police monitoring rooms, malls, and public safety organizations.

V. ACKNOWLEDGMENT

The authors would like to acknowledge the Idea Lab at the Meerut Institute of Engineering and Technology for their support in providing a segment of the audio dataset utilized for the training and validation of the VoiceGuard system.

REFERENCES

[1] J. W. Lee, S. H. Yoon, and K. Y. Lee, “Suspicious Activity Detection based on Audio Detection,” International Journal of Engineering Research & Technology, vol. 12, no. 4, pp. 45–50, 2023.

[2] M. A. Adavanne, A. Politis, and T. Virtanen, “Sound event detection using weakly labeled dataset with stacked convolutional neural network,” in Proc. DCASE Workshop, 2017, pp. 37–41.

[3] K. J. Piczak, “Environmental sound classification with convolutional neural networks,” in 2015 IEEE 25th International Workshop on Machine Learning for Signal Processing (MLSP), 2015, pp. 1–6.

[4] H. Phan, L. Hertel, M. Maassen, and A. Mertins, “Audio event detection using deep neural networks,” in Proceedings of the Detection and Classification of Acoustic Scenes and Events (DCASE), 2016, pp. 1–5.

[5] J. Gemmeke et al., “Audio Set: An ontology and human-labeled dataset for audio events,” in 2017 IEEE ICASSP, pp. 776–780.

[6] T. N. Sainath et al., “Convolutional LSTM networks for speech enhancement and detection,” in 2015 IEEE ICASSP, pp. 503–507.

[7] Y. Yamamoto and K. Kondo, “Real-time acoustic event detection using MFCC and GMM,” in 2014 IEEE International Symposium on Intelligent Signal Processing, pp. 82–87.

[8] J. F. Gemmeke, D. P. Ellis, and X. Jia, “YAMNet: Pretrained Deep Net for Audio Event Classification,” Google Research, 2020. [Online]. Available: https://tfhub.dev/google/yamnet

[9] R. K. Raut and S. R. Kolhe, “A survey on speech emotion recognition using deep learning techniques,” International

Journal of Advanced Computer Science, vol. 10, no. 2, pp. 112–118, 2023