Rare Sound Detection Using GUI Based Python Software, ”echoNet”

Nitesh Kumar, Pratham Gandhi, Neeraj Kumar and Garvit Saini

March 18, 2024

Rare sound detection using GUI basedpython software, “echoNet”

Nitesh Kumar
Computer Science and Engineering
Parul University
Vadodara, India
200303126214@paruluniversity.ac.in

Pratham Gandhi
Computer Science and Engineering
Parul University
Vadodara, India
200303126204@paruluniversity.ac.in

Neeraj Kumar
Computer Science and Engineering
Parul University
Vadodara, India
200303126211@paruluniversity.ac.in

Garvit Saini
Computer Science and Engineering
Parul University
Vadodara, India
200303126210@paruluniversity.ac.in

Himadri Vegad
Assistant Professor
Parul University
Vadodara, India
himadri.vegad21143@paruluniversity.ac.in

***Abstract*— The exponential growth of digital technologies has led to a massive increase in the volume of multimedia data generated from various smart devices such as smartphones, cameras, and audio recording devices. This massive volume of data has made it difficult to extract useful information from the multimedia data.**

This study proposes an efficient technique for anomaly detection and classification of rare events in audio data. The proposed technique is based on a deep learning-based approach that uses a convolutional neural network (CNN) to extract high-level features from the audio data. The extracted features are then used to train a support vector machine (SVM) classifier that can accurately detect and classify rare events in the audio data. The proposed technique has several advantages over traditional anomaly detection techniques.

In conclusion, the proposed technique represents a significant step forward in the field of anomaly detection and classification in multimedia data. The technique has the potential to revolutionize the way in which we process and analyze multimedia data and could have a significant impact on a wide range of applications.

I. INTRODUCTION

Audio Event Detection (AED) uses audio analysis to identify the daily audio show has attracted a lot of attention lately. A lot Environmental and wildlife monitoring system, smart home and video event detection. AED is also used in autonomous driving to avoid incidents related to visual object detection Hard Some examples of this are finding and identifying sirens on the road Dangerous situations such as car accidents and driving. AED is also used noise and loudness detection in subway trains. different signal processing and machine learning methods are proposed for AED.

Approach first Mel-frequency cepstral coefficients (MFCCs) are often used as algorithmic features such as hidden Markov model, Gaussian mixture model, non-negative matrix. factorization, support vector machine (SVM) and random forest. Deep learning (DL) models have become very popular for the detection of noisy phenomena recent years. This DL model uses a framework approach or area-based method. In the frame level approach, they analyse each subframe within it vote to determine whether or not it belongs to an event. The architecture of this model often based on convolution and repeated layers.

Traction layer Iterative layers are used to extract high-level features before they are used for learning Long-term temporal dependence between One of the biggest downsides Frame-level methods cannot account for longer context dependencies in the voice. This is solved by the second group of DL-based SED models: region-based approach. Thisalgorithm processes the spectrogram of the audio signal as an image and use the object detection model in the field of computer vision for detection the sound part related to different events.

II. LITERATURE REVIEW

2.1 Convolutional neural networks:

Hybrid Deep Neural Network (DNN)-Hidden Markov Models (HMM) have been shown to significantly improve speech recognition performance over traditional Gaussian Mixture Models (GMM)-HMM. The performance improvement is partly due to the DNN’s ability to model complex correlations between the voice features. In this article, we show that the error rate can be further reduced by using a Convolutional Neural Network (CNN). Experimental results show that CNN reduces error rate by 6 per-cent to 10 per-cent compared to DNN on TIMIT phone recognition and voice search of large vocabulary speech recognition tasks.

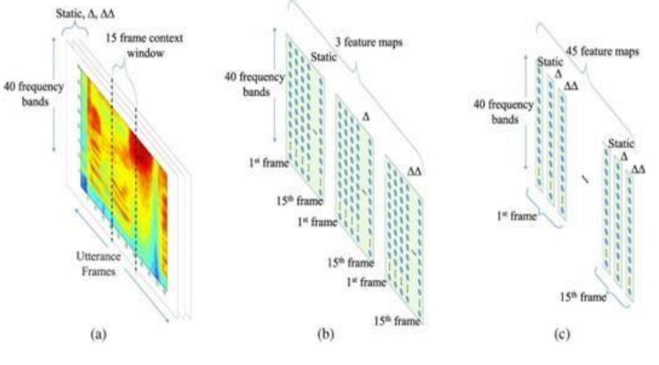

Figure 1: Voice input functions for CNNs in two different ways

2.2 Sound Recognizing using Machine Learning:

Sound analysis is a challenging task, associated to various modern applications, such as speech analytics, music information retrieval, speaker recognition, behavioral analytics and auditory scene analysis for security, health and environmental monitoring. ML methodologies focusses on hand- crafted audio features and traditional statistical classifiers such as SVMs. Machine Learning help us to make sound segmentation and classification easy. ML recognizes between different types of sounds, segment an audio signal to homogeneous parts or group sound files based on their content similarity and then classify them. Firstly, using Featured Extraction we need to find a way to go from low-level to high-level voluminous audio. Secondly, the whole audio is splitted to short-term windowing i.e. it is the audio signals split into short term signal. So that it can be thoroughly checked. Lastly, as and when all the sound is detected then the ML methodologies further tries to find loops in the sound so as to detect any suspicious sound of someone pretending to be someone else.

2.3. Echo detector:

2.3.1 Sound detector:

The use of a “Sound Detector” can be used for security purposes and surveillance that can be installed in streets, hospitals, shopping malls, etc. It will differentiate between a safe sound and a dangerous sound. In case of any unwanted sound, it will trigger the alarm. So as a result, it can be helpful while detecting unusual criminal activities. It will be useful for making the environment safe and secure. As, there are other sound sensors in market, which detects sound of clap to turn on the light. Likewise, this security sensor is going to detect the sound of people screaming, dog bark, mirror break and people crying and so on few anonymous sounds, etc. It will detect the intensity of the sound. The accuracy of this sensor can be changed for the ease of usage. This sensor will first notice the sound and its intensity and process a output signal to its microcontroller. After that, it will do the required processing.

2.3.2 Sensor:

A SENSOR is a device that detects the changes in electrical or physical quantities and produces an output as an acknowledgement of change in the quantity. Sensors are used in everyday life and objects like touch sensitive elevator buttons, IR sensor for television remote, passive infrared sensor, automatic door opening sensor in malls.

2.4 *Sound Detection System Using Hidden Markov Models:*

- The proposition of the sound-based intrusion detection system using the hidden Markov models is being done. In this the development of the sound-based intrusion detection system is done using
- Markov models. In this microphone are installed in the environment to catch the audio signals. then
- the audio signals are used to extract the relevant features and are then used to train the HMM. Each HMM has a specific event and is used according to the set of that audio samples. Then the same features of the sound signals are again extracted during the testing phase and then compared with HMM to determine the most relevant event. And if the event is classified as intrusion after detection, the alarm will raise. After the implementation it is found that the accuracy of the system was 92.5 per-cent which is more than the traditional motion detection system which was only 83.3 per-cent. In short, this system is very effective as compared to the other systems, especially in the scenarios where video or motion detection is not effective.domains, and it can also specify how to handle emails that fail the SPF check.

2.5 Audio based event-detection for Multimedia Surveillance:

With the increasing use of audio sensors in monitoring and surveillance applications, event detection using audio streams has become an important research problem. This article presents a layered approach to audio-based event detection for monitoring. The proposed method first classifies a given audio frame into sound and non-sound events and then further classifies them into normal and exciting events. We model the events using a Gaussian mixing model and optimize the parameters for four different audio characteristics ZCR, LPC, LPCC and LFCC. Experiments were performedto evaluate the effectiveness of human activity sensing features in various normal and excited states.The results show that the proposed top-down event detection method significantly outperforms single-step methods.

2.6 Echo-NET using IOT and Machine Learning Algorithm:

The key in IoT lies in real-time data capture and analytics using various types of sensors. It plays a crucial role in creating IoT devices. Sensors are innovating very fast. Sensors detect external information, replacing it with a signal that humans and machines can distinguish. IoT makesit possible to collect data in almost any situation. For instance, we use sensors in medical care, nursing care, industrial, logistics, transportation, agriculture, disaster prevention, tourism, regionalbusinesses, and many more. IOT sensors can detect various phenomenon like pressure, temperature Heat, including gyro sensors, sound sensor, odor sensor etc. so therefore our anonymous sound canbe easily detected by the IOT sensors can be uploaded directly on the internet so that people around can see.

2.7 Analysis and Design of the sensor:

Like other sensors, this is the sensor that can be easily used and can easily modified and its setup also didn't take too much of time. This Sensor can easily be installed and pre-installed in houses, societies, outside any company office or big MNCs. Its costing is also not very high so it makes affordable to every single person out there in cheap rates Its setup includes programmed languagelike java and python which are pre-installed and are calibrated with ease and the whole sensor connected with IoT So that the alarm as well as the people around the area come to know regarding the anonymous activity or criminal activity in the area so as to facilitate early possible and provide security to the suspected individual. This sensor has a microphone for detecting sound intensity. The electret microphone detects the sound wave and then sends it to the sensor circuit board whichconsists of a voltage comparator IC LM393 and potentiometer. Comparator IC LM393 process thissignal and convert it into a Digital Output. The potentiometer is used to adjust the sensitivity of the sensor.

2.8 Softwares and Hardwares Required:

III. DEFINITIONS OF TERMS

In this section, we aim to provide clear and formal definitions of essential terms and concepts vital to understanding the research. These definitions are intended to enhance clarity and promote a common understanding of the subject matter. Below, you will find concise explanations of the key terms used throughout this paper:

Sensor: A SENSOR is a device that detects the changes in electrical or physical quantities and produces an output as an acknowledgement of change in the quantity. Sensors are used in everyday life and objects like touch sensitive elevator buttons, IR sensor for television remote, passive infrared sensor, automatic door opening sensor in malls.

Speed Sensors: Speed Sensors are basically the sensors used for detecting speed of an object or a vehicle. It includes wheel speed sensor, LIDAR, speedometer (in vehicles), doppler radar, air speed indicatoretc.

Audio Event Detection: Audio Event Detection (AED) uses audio analysis to identify the daily audio show has attracted a lot of attention lately. A lot Environmental and wildlife monitoring system, smart home and video event detection. AED is also used in autonomous driving to avoid incidents related to visual object detection.

CRNN: CRNN (Convolutional Recurrent Neural Network) models are commonly used in audio detection systems, particularly for tasks such as speech recognition and music genre classification. These models are effective in handling time-series data and are well-suited for detecting patterns in audio signals. In audio detection systems, CRNN models typically take in the raw audio waveform as input and apply a series of convolutional layers to extract local features.

Pyaudio: Pyaudio is a Python library that provides a way to interact with your computer's audio hardware. Itcan record and play back audio data and perform various data processing tasks. Pyaudio is built ontop of the PortAudio library, a cross-platform C library for audio I/O. With Pyaudio you can: Record audio data from a microphone or other input device.

Librosa: Librosa is a specialized Python library designed for the analysis and manipulation of audio and music data, making it an invaluable tool for sound detection tasks. In particular, Librosa excels at extracting meaningful features from audio signals, which are essential for distinguishing betweendifferent sound classes, such as gunshots, window breaks, and baby cries

IV. METHODOLOGY

4.1 Developing echoNet:

i) Why Python:

Python scripts and APIs can be tailor-made into effective network monitoring and forensics tools. Their versatility makes them ideal in assorted applications including cybersecurity, data mining, Internet of Things, cloud simulation, grid implementation, etc.

ii) Advantages of Python:

The software development companies prefer Python language because of its versatile features and fewer programming codes. Nearly 14 percent of the programmers use it on operating systems like UNIX, Linux, Windows, and Mac OS. The programmers of big companies use Python as it has created a mark for itself in the software development with characteristic features like:

• Interactive • Interpreted • Modular • Dynamic • Object-oriented • Portable • High-level • Extensible in C++ and C

iii) Benefits of Python:

Extensive Support Libraries It provides large standard libraries that include areas like string operations, Internet, web service tools, operating system interfaces, and protocols. Most of the highly used programming tasks are already scripted into it, limiting the length of the codes to be written in Python.

Integration Feature Python integrates the Enterprise Application Integration that makes it easy to develop Web services by invoking COM or COBRA components. It has powerful control capabilities as it calls directly through C, C++ or Java via Jython. Python also processes XML and other markup languages as it can run on all modern operating systems through the same bytecode.

Improved Programmer's Productivity The language has extensive support libraries and clean object-oriented designs that increase two to tenfold of the programmer's productivity while using languages like Java, VB, Perl, C, C++ and C.

4.2 Materials and Methods:

In this section, we will use two supervised neural network algorithms. CNN and CRNN. These neural network algorithms are state of the art. Classification algorithms used for audio detection and classification in recent IoT-based CPS, (including shot) and other security-oriented applications.

1. CNN models in Audio Detection Systems:

Convolutional neural networks are considered the best learning algorithms for understanding image content. CNNs are inspired by the organization of animal visual cortex, and the deeper they go and the more difficult the is to train, the higher the performance

2. CRNN Models in Audio Detection Systems:

CRNN models are commonly used in audio detection systems, particularly for tasks such as speech recognition and music genre classification. These models are effective in handling time-series data and are well-suited for detecting patterns in audio signals. The output of the convolutional layers is then passed through a series of recurrent layers, such as LSTM (Long Short-Term Memory) or GRU (Gated Recurrent Unit), which can capture temporal dependencies in the audio signal. CRNN models can be trained using supervised learning techniques, where a labeled dataset of audio samples is used to train the model to recognize specific sounds or events.

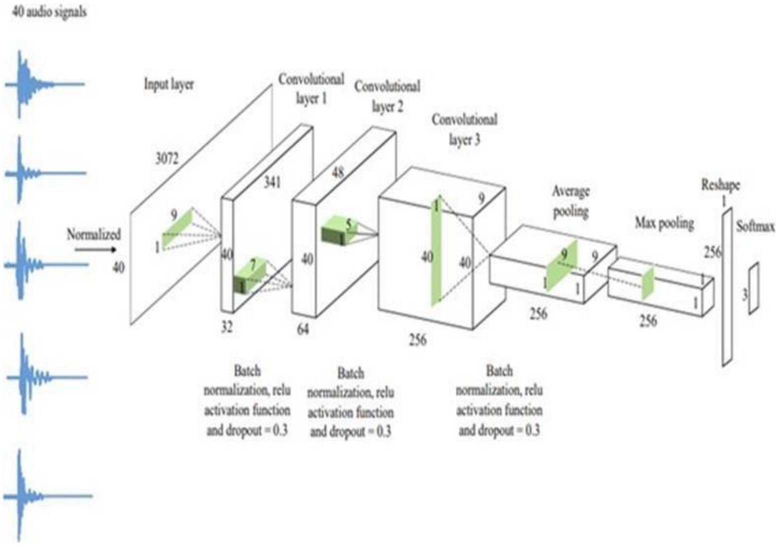

Figure 2: CNN Model Architecture

4.3 Sound Detection:

Sound detection requires modules that allow us to work with sounds. Python provides libraries to work with sound and its properties.

>Pyaudio is a Python library that provides a way to interact with your computer's audio hardware. It can record and play back audio data and perform various data processing tasks. Pyaudio is built ontop of the PortAudio library, a cross-platform C library for audio I/O. With Pyaudio you can Record audio data from a microphone or other input device.

>Librosa is a specialized Python library designed for the analysis and manipulation of audio and music data, making it an invaluable tool for sound detection tasks. In particular, Librosa excels at extracting meaningful features from audio signals, which are essential for distinguishing between different sound classes, such as gunshots, window breaks, and baby cries.

>PyTorch is a dynamic and popular deep learning framework that has gained significant tractionin recent years, and it is well-suited for sound detection tasks. When it comes to sound detection, particularly for identifying various sound classes like gunshots, window breaks, and baby cries, PyTorch offers a robust platform for developing accurate and efficient models.

>Keras key advantages is its user-friendly and modular design. It allows developers to define neural network layers, specify activation functions, and connect them easily to construct custom models for sound detection. Keras also provides a wide range of pre-built layers and models, whichcan be adapted and fine-tuned to suit specific sound classification requirements. Moreover, Keras seamlessly integrates with TensorFlow as its backend, leveraging TensorFlow's capabilities for audio data preprocessing and feature extraction.

V. DISCUSSION

Deep Learning:

Neural networks (ANNs) were inspired by the way neurons in the human brain perform various functions. In humans, different signals activate different neurons. Neurons is like map input signals into cognitive representations. In the ANN, each neuron computes a weighted sum of the input signals and passes the result through a linear or non-linear activation function. Just like the human brain, inputs pass through layers of interconnected neurons that are converted into outputs. The first layer of neurons that receives input signals is called the input layer, and the last layer that generates output signals is called the output layer. All layers between the input and output layers are called hidden layers. The origins of ANNs go back to 1943, when McCulloch and Pitts created a linear model to recognize two types of inputs. In 1958, Rosenblatt proposed a pattern recognition algorithm, the perceptron, with two training layers. A backpropagation algorithm formulated in was applied to modify the link weights of the model to minimize the difference between the computed output and the desired one. The next breakthrough happened when in 2006, Geoffrey Hinton showed that multiple stacked layers of neural networks can be trained using a method called greedy layer-wise pre-training [66]. The term “deep learning” or “deep neural networks” started to be commonly used to refer to the neural nets with two or more hidden layers. By adding more layers and more units within a layer, a deep network can represent functions of increasing complexity.

One constraint for training deep models with millions of parameters is the need for large datasets. In the supervised learning scheme, having a large number of labelled data makes it possible for the model to learn the mapping function from the input data samples to the output values. More recently, features in Internet have made it easier to create large labeled datasets. With approximately 3.2 million tagged images, ImageNet was first published in 2009 and the current sample count has grown from to approximately 14 million.

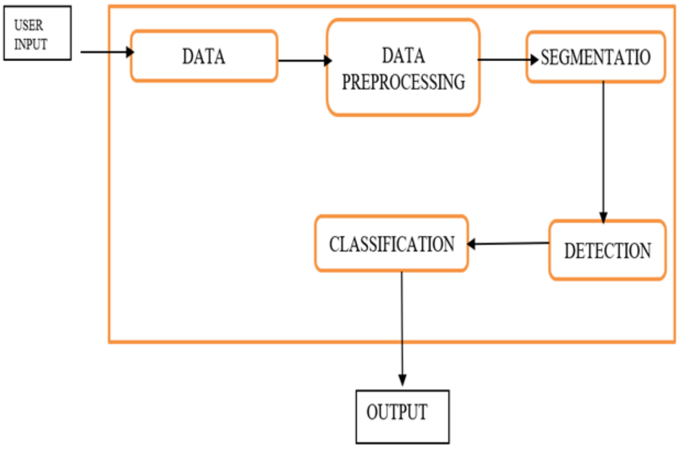

Figure 4 Architecture of echoNet

Feature Extraction:

We use the processed and normalized log-scale mel spectrogram as input features our model. A spectrogram is a 2D representation of sound in time-frequency represents the power of a specific frequency signal in brightness or color terms They retain more information than many manual features. Log-mail feature maps show local locations in the time and frequency domains. Mel-spectrograms have proven to be good characterizations of sound for many in-depth studies sound event detection method. Their difference from normal spectrograms is the use of large-scale filter banks to simulate non-linear auditory perception of sound. We used a window size of 46ms, which corresponds to half the amplitude of the sound output signal 128 filter bank for each frame of audio signal. Window size and number of filter banks are the same as. Then we count this filter is the logarithm of the bank value and its normalization.

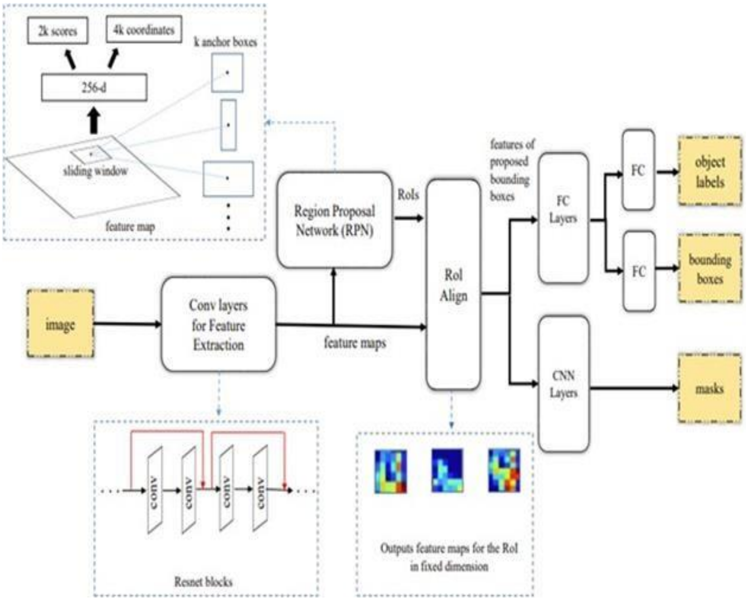

Figure5RCNNFramework

VI. FUTURE WORK

7.1 Edge Computing:

Optimize the system for edge computing, enabling sound analysis to occur on the device itself without the need for constant internet connectivity. This enhances privacy and reduces latency.

7.2 Localization and mapping:

Incorporate the ability to pinpoint the location of the detected sound source, allowing for more precise response actions. Combining acoustic data with data from other sensors (e.g., cameras) can provide a more comprehensive situational understanding

7.3 Cross-Platform Compatibility:

Ensure compatibility with a wide range of hardware platforms, from low-cost microcontrollers to high-end smartphones and dedicated security systems.

VII. CONCLUSION

The development and implementation of our Echonet system is to detect anonymous sounds, such as gunshots, window breaks, and baby cries, represent a significant technological advancement in the realm of audio surveillance and safety. Throughout this project, we have explored cutting-edge machine learning techniques, signal processing, and hardware integration to create a robust and efficient solution that addresses critical security and well-being concerns.

Our system's ability to differentiate and identify these distinct sounds showcases its potential in a variety of applications. For security purposes, it offers real-time threat detection in public spaces, potentially aiding law enforcement agencies in responding to incidents promptly. Additionally, its use in residential settings can provide peace of mind for homeowners, ensuring timely responses to emergencies such as break-ins or incidents involving infants.

Furthermore, our project underscores the importance of ethical considerations in the development and deployment of audio surveillance systems. We must balance the benefits of enhanced security and safety with privacy concerns, ensuring that the collected data is used responsibly and within legal boundaries. Implementing measures such as anonymizing data and obtaining informed consent when necessary will be crucial in achieving this balance. Implement a robust notification system that can instantly alert users or relevant authorities when a potentially threatening sound is detected. This could involve integration with smartphone apps or security systems. Invest in research to develop more advanced and efficient machine learning models for sound recognition. Deep learning architectures like convolutional neural networks (CNNs) and recurrent neural networks (RNNs) can potentially enhance accuracy and reduce false positives.

REFERENCES

1. Ossama Abdel-Hamid et al. "Convolutional neural networks for speech recognition". In: IEEE/ACM Transactions on audio, speech, and language processing 22.10 (2014), pp. 1533–1545.
2. Rajeev Aggarwal et al. "Noise reduction of speech signal using wavelet transform with modified universal threshold". In: International Journal of Computer Applications 20.5 (2011), pp.
3. Pradeep K Atrey, Namunu C Maddage, and Mohan S Kankanhalli. "Audio based event detection for multimedia surveillance". In: 2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings. Vol. 5. IEEE. 2006, pp. V–V.
4. Yan Zhang and Dan-jv LV. "Selected features for classifying environmental audio data with random forest". In: The Open Automation and Control Systems Journal 7.1 (2015).
5. Carlini, N.; Wagner, D. Towards evaluating the robustness of neural networks. In Proceedings of the IEEE Symposium on Security and Privacy, San Jose, CA, USA, 22–26 May 2017.
6. Freesound. Available online: freesound.org/help/faq/ (accessed on 1 January 2021).
7. How to Set up Alexa Guard on an Amazon Echo. Available online: cnbc.com/2019/05/14/how- to-set-up-alexa-guard-on-an amazon-echo.html (accessed on 1 January 2021).
8. N. Morgan, "Deep and wide: Multiple layers in automatic speech recognition," IEEE Trans. Audio, Speech, Lang. Process., vol. 20, no. 1, pp. 7–13, Jan. 2012.
9. What Is a Spectrogram. Available online: tomroelandts.com/articles/what-is-a-spectrogram (Accessed on 1 January 2021)
10. DG Aggelis et al. "Acoustic emission monitoring of degradation of cross ply laminates". In: The Journal of the coustical Society of America 127.6 (2010), EL246–EL251.