Design of Audio based Accident and Crime Detection using Simple Architecture of Neural Network

Afis Asryullah Pratama[1], Sritrusta Sukaridhoto[1], Mauridhi Hery Purnomo[2], Achmad Basuki[3], Vita Lystianingrum[4] and Rizqi Putri Nourma Budiarti[5]

[1]*Department of Electronic Engineering, Politeknik Elektronika Negeri Surabaya, Surabaya, Indonesia*
[2]*Department of Computer Engineering, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia*
[3]*Department of Creative Multimedia Technology, Politeknik Elektronika Negeri Surabaya, Surabaya, Indonesia*
[4]*Department of Electrical Engineering, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia*
[5]*Engineering Department, Universitas Nahdlatul Ulama Surabaya, Surabaya, Indonesia*

Keywords: Audio Recognition, Mel Spectrogram, CNN, RNN, Surveillance System.

Abstract: The accident and crime happened on the road still increasing nowadays. Those two events were considered as emergency event that need a quick response. In this research, a method to detect accident and crime were proposed. The proposed method uses audio data as input and extracting the Mel spectrogram as the feature, which later be fed to our simple neural network architectures. We classify our dataset into engine_idling, car_crash, and gun_shot classes to represent normal, accident, and crime condition on the road. Our simple CNN architecture obtains accuracy of 95.31% and 93.75% with 200ms and 1000ms segment duration respectively, and our simple RNN architecture obtains 86.67% and 58.67% by using 200ms and 1000ms segment duration respectively. We can conclude that the best simple architecture was performed by CNN architecture with 200ms segment duration.

1 INTRODUCTION

The transportation technology was being developed day by day, this has an impact not only to the system of the vehicles but also the number of vehicles and its passengers, in Indonesia there were 146,858,759 vehicles which include passenger cars, buses, freight cars, and motorcycles in 2018 (BPS, 2018b; Mahfuzhon and Setyawan, 2018). The increment of vehicles and its passengers also increase the number of accidents happened. In 2018, there are 109,215 accidents and 29,472 deaths were recorded in Indonesia (BPS, 2018a). Most of the death cases from car accident were happened due to the late treatment for the casualties (Kattukkaran, George, and Haridas, 2017).

Other than car accidents, crime also an emergency event that needs a quick response. In 2018 there are 8,423 mugs and 90.757 snitches happened in Indonesia. But according to statistics but only 23.44% in 2017 and 23.99% in 2018 was reported (Badan Pusat Statistik, 2019). The low reporting rate of crimes mostly caused by the lack of awareness and information about where to report it.

Therefore, we need a system to detect accidents and crimes with capability to deliver the emergency events happened on the road. In this research, we propose a method to detect car crash, idling engine and gunshot as representatives of accident and crime sounds. we use audio data of mentioned events for surveillance purpose. We use several audio segmentation parameters and neural network architectures to find the best result for the case we focused on.

2 SYSTEM DESIGN

There are many methods for audio recognition such as analyzing both time and frequency domain of sample audio gives an accuracy of 65%-82% (Sammarco and Detyniecki, 2018), or by extracts the audio feature using MFCC and inferenced with DNN which gives an accuracy of 98.4% (Arslan and Canbolat, 2018). The MFCC and DNN method could

Pratama, A., Sukaridhoto, S., Purnomo, M., Basuki, A., Lystianingrum, V. and Budiarti, R.
Design of Audio based Accident and Crime Detection using Simple Architecture of Neural Network.
DOI: 10.5220/0010940300003260
In *Proceedings of the 4th International Conference on Applied Science and Technology on Engineering Science (iCAST-ES 2021)*, pages 76-81
ISBN: 978-989-758-615-6; ISSN: 2975-8246

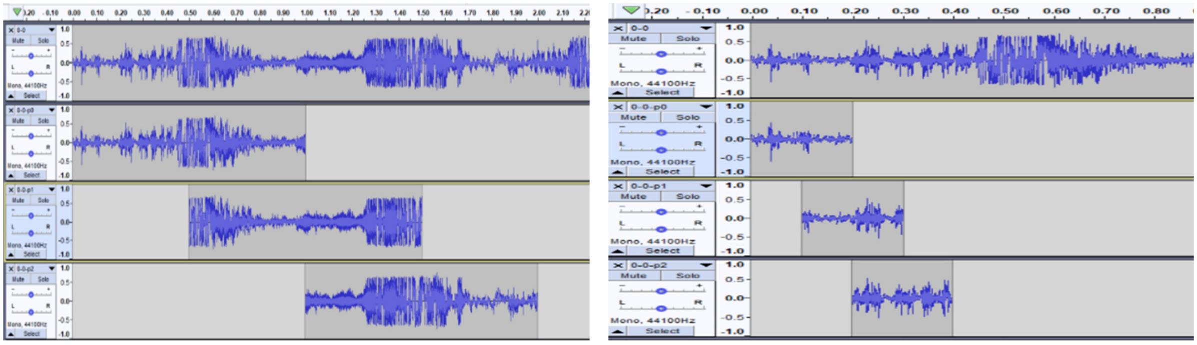


Figure 1: Audio segmentation with (1000ms length and overlap 50%, left), (200ms length and overlap 50%, right).

give a better result but it could be unimplementable in a small memory device, therefore we need a simpler architecture which requires less memory.

Our proposed method extracts the features of audio signal using Mel spectrogram and uses neural network with thresholding to get the conclusion. Complete steps of our proposed method were shown in Figure 2.

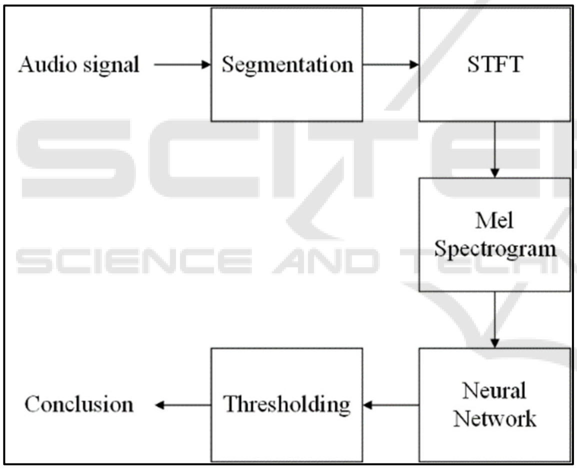


Figure 2: Methods workflow.

Our methods consist of several steps, which are segmentation, framing, windowing, short time Fourier transform (STFT), Mel spectrogram extraction, neural network and thresholding with following details.

2.1 Segmentation

We use single channel audio signal with sampling rate 44100Hz as the input. Then we slice the audio into smaller segments with 2 different parameters, first with 1000ms length and the other with 200ms length both with 50% overlap ratio Figure 1.

We use different parameters to analyze the effects of segment length to our proposed method's performance.

2.2 STFT

STFT were used to compute Fourier transform of the segment faster. STFT step consists of 3 subprocesses as shown in Figure 3.

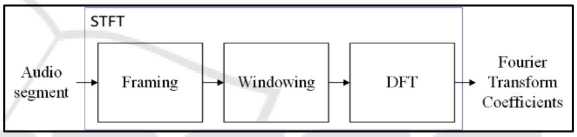


Figure 3: STFT steps.

2.2.1 Framing

For each segment will be framed into smaller frames with frame width 1764 samples, and hop length 441 samples. Frame width for Fast Fourier Transform (FFT) must be the power of 2, therefore 1764 fulfil the requirement needed.

2.2.2 Windowing

The discontinuity for each frame's edge could cause a spectral leakage. Thus, we implement windowing method using Hann window function.

$$w(k) = 0.5 \cdot \left(1 - cos\left(\frac{2\pi k}{K-1}\right)\right), k = 1 \ldots K \qquad (1)$$

w : window coefficients
k : coefficient index
K : window width

We use Hann window with window width 1764 samples so we could fit the whole frame with the window.

2.2.3 FFT

FFT is a Discrete Fourier Transform (DFT) with a faster computation process. Therefore, we use it to obtain the Fourier transform coefficients within a shorter time. DFT equation is shown in (2).

$$\hat{x}\left(\frac{k}{N}\right) = \sum_{n=0}^{N-1} x(n) \cdot e^{-2i\pi n\frac{k}{N}} \tag{2}$$

$\hat{x}$: Discrete Fourier series
k : [0, N-1]
N : Fourier width
x : Input samples
n : sample index

We use FFT with N=1764, so we could calculate Fourier transform coefficients for each windowed frame.

2.3 Mel Spectrogram

Human hearing perception of frequencies is not linear but logarithmic, meaning that human hearing has a higher resolution at high frequencies. The perception graph is described in the Mel scale which can be mapped with (3).

$$Mel(f) = 2595 \cdot log\left(1 + \frac{f}{500}\right) \tag{3}$$

Mel : Mel coefficients
f : total filter banks

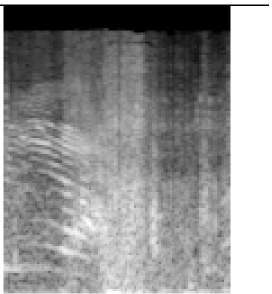
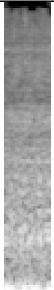
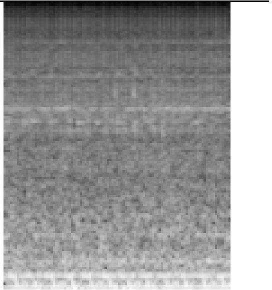
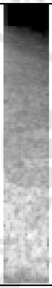
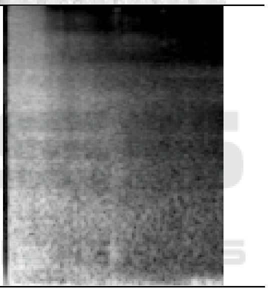
With equation above, we calculate 128 Mel filter banks then convolute it with Fourier transform coefficients that has been converted into dB unit.

2.4 Neural Network

In this research, we use Neural network (NN) in two different architectures, which are Convolutional Neural Network (CNN), and Recurrent Neural Network (RNN). We use two different architectures to analyze which performs better in the focused case.

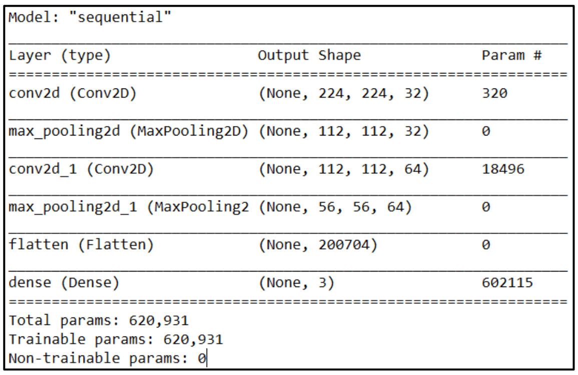
We use a total of 133 voice data with the label car_crash from YouTube channels CarCrashesTime and CarCrashesTime2, and a total of 201 voice data with the engine_idling label and 361 voice data with the gun_shot label from the UrbanSound8K dataset (Car Crashes Time - YouTube n.d.; Car Crashes Time - YouTube n.d.; Salamon, Jacoby, and Bello 2014). Those 3 labels represent accident, normal, and crime condition on the road.

Table 1: Pre-processed datasets.

Class	Mel Spectrogram	
	200ms	*1000ms*
Car_crash		
Engine_idling		
Gun_shot		

2.4.1 CNN

In this study, a CNN architecture was built which consists of two 2-dimensional convolution layers with the activation function of Rectified Linear Unit (Relu) and 2-dimensional max pooling, then continued with a fully connected layer with the SoftMax activation function as follows.



```
Model: "sequential"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
conv2d (Conv2D)              (None, 224, 224, 32)      320
_________________________________________________________________
max_pooling2d (MaxPooling2D) (None, 112, 112, 32)      0
_________________________________________________________________
conv2d_1 (Conv2D)            (None, 112, 112, 64)      18496
_________________________________________________________________
max_pooling2d_1 (MaxPooling2 (None, 56, 56, 64)        0
_________________________________________________________________
flatten (Flatten)            (None, 200704)            0
_________________________________________________________________
dense (Dense)                (None, 3)                 602115
=================================================================
Total params: 620,931
Trainable params: 620,931
Non-trainable params: 0
```

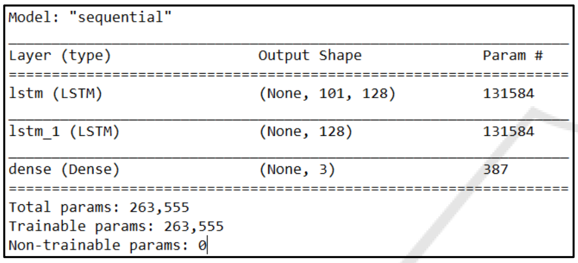
Figure 4: Simple CNN architecture.

The architecture has an input of a 224x224 sized matrix, and provides confidence values from 3 classes as an output.

The training process were done by using the training dataset which divided into training and validation datasets with the ratio of 8:2. The training process is done with the parameters learning rate = 0.0001, epochs = 40 and the Adam optimizer.

2.4.2 RNN

In this study, a CNN architecture was built which consists of 2 Long Short Term Memory (LSTM) layer, then continued with a fully connected layer with the SoftMax activation function as shown in Figure 5.



```
Model: "sequential"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
lstm (LSTM)                  (None, 101, 128)          131584
_________________________________________________________________
lstm_1 (LSTM)                (None, 128)               131584
_________________________________________________________________
dense (Dense)                (None, 3)                 387
=================================================================
Total params: 263,555
Trainable params: 263,555
Non-trainable params: 0
```

Figure 5: Simple RNN architecture.

The architecture has an input of a 101x128 sized matrix, and provides output in the form of confidence values from 3 dataset classes.

We train our model using training dataset which divided into training and validation data with ratio of 8:2. And train it with the same parameter as CNN.

2.5 Thresholding

Thresholding was used to ignore inference result with confidence value lower than assigned threshold value. The threshold value was configured by trial-and-error method.

3 EXPERIMENT AND RESULT

This section explains the results of the experiments of our proposed method.

3.1 CNN Test

Our CNN model was tested using 150 data from 3 different classes with result shown at Figure 6.

The CNN model with 200ms and 1000ms segment duration obtains accuracy of 95.31% and 93.75% respectively.

3.2 RNN Test

We test our RNN model with 150 data from 3 different classes with following results.

Our RNN model with 200ms and 1000ms segment duration obtains accuracy of 86.67% and 59.97% respectively.

3.3 Comparison

We compare the accuracy of our proposed method with the research of (Sammarco and Detyniecki 2018) and (Arslan and Canbolat 2018). The detailed comparison shown in Table 2.

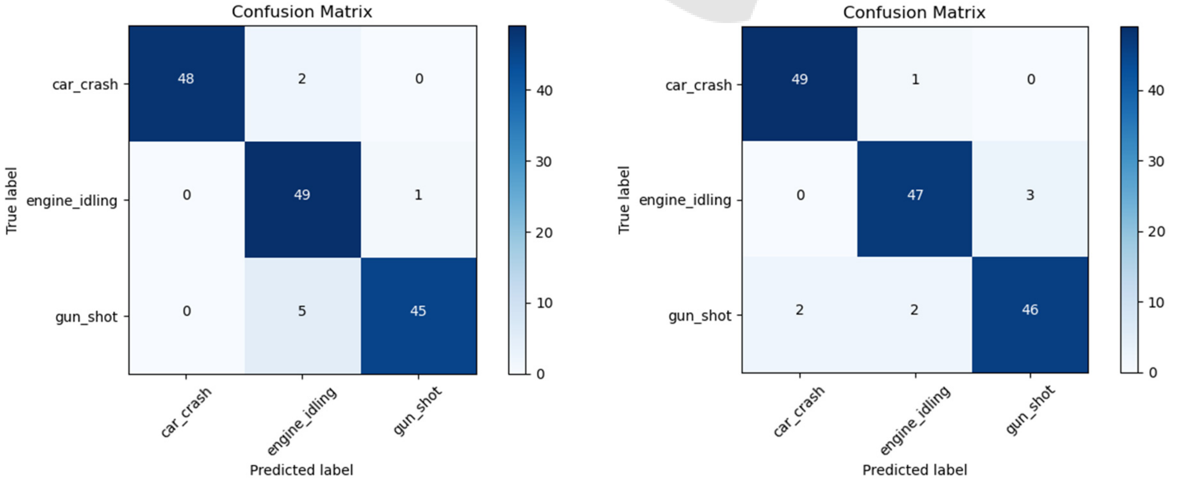


Figure 6: CNN test with (200ms segment duration, left), (1000ms segment duration, right).

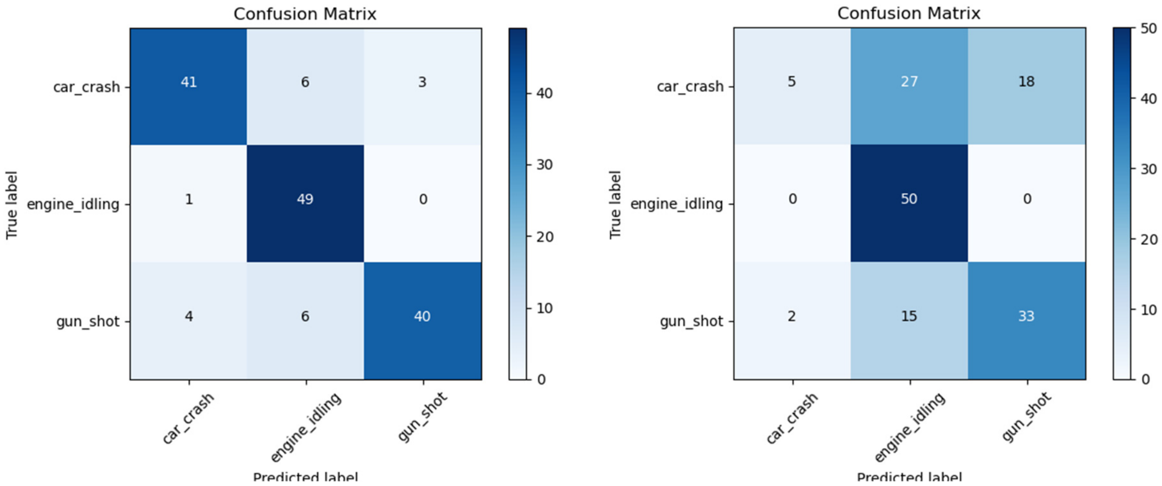


Figure 7: RNN test with (200ms segment duration, left), (1000ms segment duration, right).

Table 2 : Method comparison detail.

Method	Accuracy (%)
Our simple CNN	93.75 – 95.31
Our simple RNN	59.97 – 86.67
(Sammarco and Detyniecki 2018)	65 – 82
(Arslan and Canbolat 2018)	98.4

4 CONCLUSION

Our proposed method is able to classify car crash, engine idling and gun shot from the audio data with various accuracy. The shorter segment duration could give a higher accuracy for accident and crime detection, applied for both CNN and RNN architecture. CNN has a better accuracy than RNN architectures with lowest accuracy of 93.75%. The best model was performed by CNN model with 200ms segment duration.

In the future, we hope our simple neural network architecture could be improved with more dataset and implemented in an embedded system with limited memory resources as an early warning surveillance system.

ACKNOWLEDGEMENTS

This research was registered in World Class Professor program with number 082/E4.1/AK.04.PT/2021. Special gratitude to World Class Professor program who funded this research and thanks everyone who participated in this Audio Based on Road Accident and Crime Detection using Neural Network Research Team. We like to give our thanks for the help, support, and knowledge to given in this research.

REFERENCES

Arslan, Yuksel, and Huseyin Canbolat. 2018. “Performance of Deep Neural Networks in Audio Surveillance.” In *2018 6th International Conference on Control Engineering and Information Technology, CEIT 2018*,.

Badan Pusat Statistik. 2019. “Statistik Kriminal 2019.” *Badan Pusat Statistik*: 1–218.

BPS, Badan Pusat Statistik. 2018a. “Jumlah Kecelakaan, Koban Mati, Luka Berat, Luka Ringan, Dan Kerugian Materi Yang Diderita Tahun 1992-2018.” https://www.bps.go.id/linkTableDinamis/view/id/1134 (January 8, 2021).

———. 2018b. “Perkembangan Jumlah Kendaraan Bermotor Menurut Jenis, 1949-2018.” https://www.bps.go.id/linkTableDinamis/view/id/1133 (January 8, 2021).

“Car Crashes Time - YouTube.” https://www.youtube.com/c/CarCrashesTime2 (July 9, 2021a).

———. https://www.youtube.com/user/CarCrashesTime (July 9, 2021b).

Kattukkaran, Nicky, Arun George, and T. P.Mithun Haridas. 2017. “Intelligent Accident Detection and Alert System for Emergency Medical Assistance.” In *2017 International Conference on Computer Communication and Informatics, ICCCI 2017*,.

Mahfuzhon, Adnan, and Gembong Edhi Setyawan. 2018. “Rancang Bangun Alat Pendeteksi Kecelakaan Mobil Menggunakan Sensor Akselerometer Dan Sensor 801s Vibration.” *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer (J-PTIIK) Universitas Brawijaya.*

Salamon, Justin, Christopher Jacoby, and Juan Pablo Bello. 2014. “A Dataset and Taxonomy for Urban Sound

Research." In *MM 2014 - Proceedings of the 2014 ACM Conference on Multimedia*,.

Sammarco, Matteo, and Marcin Detyniecki. 2018. "Crashzam: Sound-Based Car Crash Detection." In *VEHITS 2018 - Proceedings of the 4th International Conference on Vehicle Technology and Intelligent Transport Systems*,.