International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

A Study on the Sound Recognition Method of Autonomous Vehicle using CNN

Taeho Kim[1], Minhyeok Yoo[2], Dae Kyeon Shin[3], Gooman Park[4], Seongkweon Kim*

Submitted: 06/06/2022 **Accepted:** 10/09/2022

Abstract- In this paper, a study on the algorithm that recognizes and judges sound source using convolutional neural network (CNN) is introduced. It is assumed that multiple of microphones are attached to receive sound information. The received sound information is then converted to visual information with the Mel-spectrogram which expands 1-dimensional sound information to 2-dimensional information. However, the shorter the extraction time by reducing n_mels, the lower the resolution of the image and the lower the performance as learning data. The value of n_mels = 64 is suggested to minimize the extraction time of Mel-spectrogram because this algorithm should be used in the autonomous vehicle. Through the computational experiment, 95% accuracy was obtained through CNN, machine learning.

Keywords*: Autonomous Vehicle; Audio Recognition; Deep Learning; Convolutional Neural Network; Mel-spectrogram*

1. Introduction

Vehicle drivers use both senses of sight and hearing to perceive and judge their surroundings. However, most of the technologies currently applied to autonomous vehicles rely on the driver's vision. However, judging the situation by visual information alone has limitations when compared to the overall judgment of humans through visual and auditory sensing. That is, it is impossible to make a three-dimensional judgment like human being using only visual information except auditory information. For example, imagine a situation when a siren of a police car or an ambulance can be heard from a distance, or a natural disaster such as a rockfall or an earthquake occurs. When visual information is not sufficient, autonomous vehicles that rely solely on visual information cannot make accurate judgments and responses. Therefore, auditory information must be used for autonomous vehicles to make three-dimensional judgment such as human level.

Therefore, for classifying and recognizing acoustic signals, we propose a system where five microphones are attached to a vehicle to receive sound information, and after conversion to visual information called mel-spectrogram which expands and receives the received one-dimensional sound signal in two dimensions, using machine learning model convolution neural network (CNN). We propose this system for analysis and recognition. As a result, an autonomous vehicle having an auditory function for determining the direction, distance and type of an acoustic signal can be realized. On top of that, this autonomous vehicle technology is capable of determining a three-dimensional surrounding situation like a human

2. Artificial neural network and CNN

This study used supervised learning that uses input and correct answer data to classify acoustic recognition and used artificial neural network model. The artificial neural

[1]Dept of Integrated IT Engineering, Seoul National University of Science and Technology, Republic of Korea
ORCID ID: 0000-0002-9627-3489
[2]Dept of Information Technology and Media Engineering, Seoul National University of Science and Technology, Republic of Korea
ORCID ID: 0000-0002-4296-5762
[3]Dept of Information Technology and Media Engineering, Seoul National University of Science and Technology, Republic of Korea
ORCID ID: 0000-0001-6535-7746
[4]Dept of Electronic IT Media Engineering, Seoul National University of Science and Technology, Republic of Korea
ORCID ID: 0000-0002-7055-5568
**Dept of Electronic IT Media Engineering, Seoul National University of Science and Technology, Republic of Korea*
ORCID ID: 0000-0001-5337-7731
**Corresponding author Email: kim12632@seoultech.ac.kr*

network learns through the process of modifying the algorithm by contrasting the output signal with the correct answer signal according to the input. The artificial neural network is composed of an input layer, a hidden layer, and an output layer, and each layer consisting of several nodes. The neural network learning is done by modifying the weights connected to each node. That is, in order to reduce the difference between the correct answer signal and the output signal according to the input, the weight of the node close to the correct answer signal should be increased. Thus, the process of adjusting the weight by comparing with the correct answer signal is called 'learning'. In addition, the neural network has a more complicated structure which allows for more sophisticated classification and increasing the number of hidden layers to two or more. This is called a deep neural network, and learning through the deep neural network is called deep learning. Deep neural networks are segmented by adding specific devices or functions to each node. Fig. 1 shows the structure of a CNN. The CNN is a structure in which a convolution layer and a pooling layer are added to the hidden layer. CNN, which plays a key role in acoustic recognition, is a deep neural network specialized in the classification of visual images. The process by which CNN trains image classification is as follows. First, when a training image is input to the convolutional layer, a filtering technique is applied to represent the image as a matrix including feature data, and each element of the matrix is converted to be suitable for data processing. Data extracted to some extent does not need to be used to determine all of them, so it undergoes a subsampling process that reduces feature data and then goes through a maximum pooling process that selects the largest value among the reduced data. As a result, this process reduces the size of the data in the input image and reduces the noise, which leads to better image discernment Therefore, CNN model is widely used for image recognition.

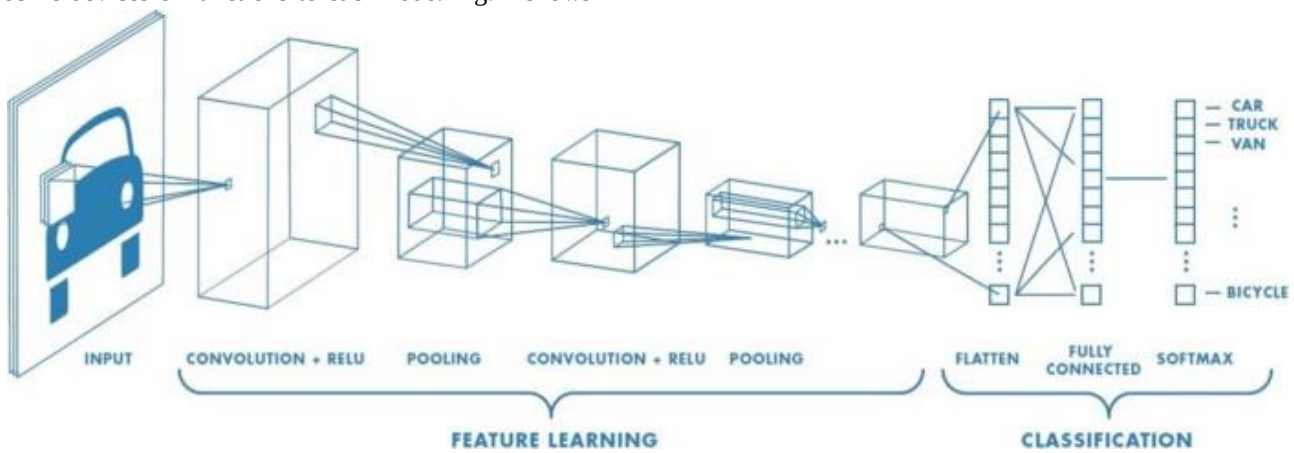


Figure 1- Structure of CNN[2]

3. Mel-spectrogram

The sound signal has amplitude information in the time axis and the frequency band. When a sound signal is received without special processing, it is inevitably expressed in one dimension. The mel-spectrogram expresses the one-dimensional sound signal information as one image. Fig. 2 shows a mel-spectrogram extracted from the sound of a vehicle crash. In the mel-spectrogram, the horizontal axis represents time and the vertical axis represents frequency. The amplitude values are expressed in color. The larger the amplitude value, the brighter the color. Through this, one-dimensional sound information can be changed into a two-dimensional image, and the overall information can be understood.

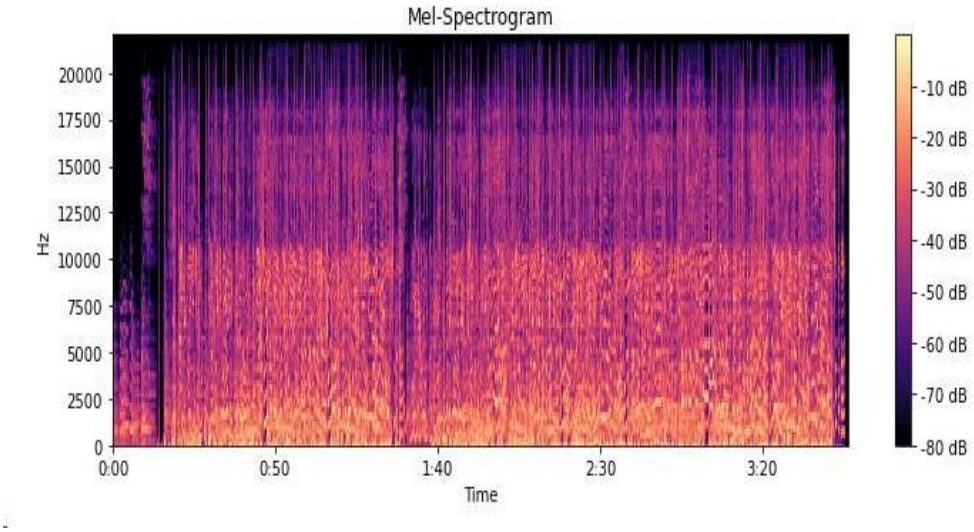


Figure 2- Mel-spectrogram

4. Sound recognition algorithm using CNN

Figure 3 shows the sound recognition algorithm that can be applied to autonomous vehicles. First, when the audio signal is input to the microphone sensors arranged at the corners of autonomous vehicle, the audio analysis algorithm extracts the mel-spectrogram of the audio signal. The algorithm performs sound source location recognition through Inter-aural Time Difference (ITD) and Inter-aural Level Difference (ILD). The mel-spectrogram is used as the input data of the completed CNN learning. CNN compares the learned mel-spectrogram with the input data through the convolution and determines the specific audio signal. The output determined from the specific signal leads to proper reaction. For example, when autonomous vehicle recognizes a siren sound from the front right side, it reacts to avoid the direction to the left side. The reason for using CNN is that it is more effective to extract a feature of information from two-dimensional image than from one-dimensional sound signal. In CNN, the process of extracting the feature of data is included in the neural network, so the processing is omitted in the algorithm. In addition, it is possible using a small number of learning data, because the two-dimensional mel-spectrogram diversifies the one-dimensional acoustic information. [3]

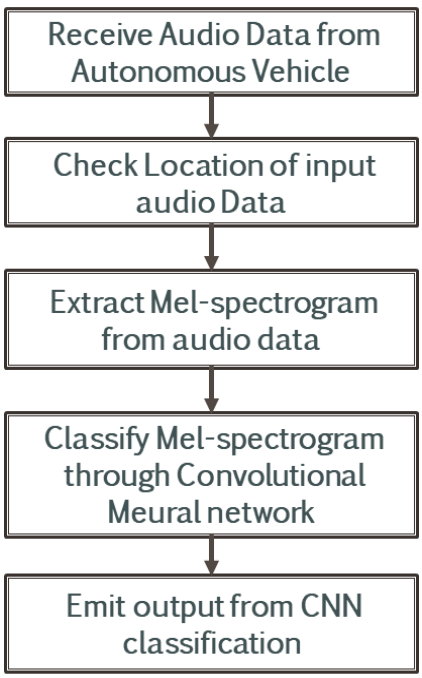
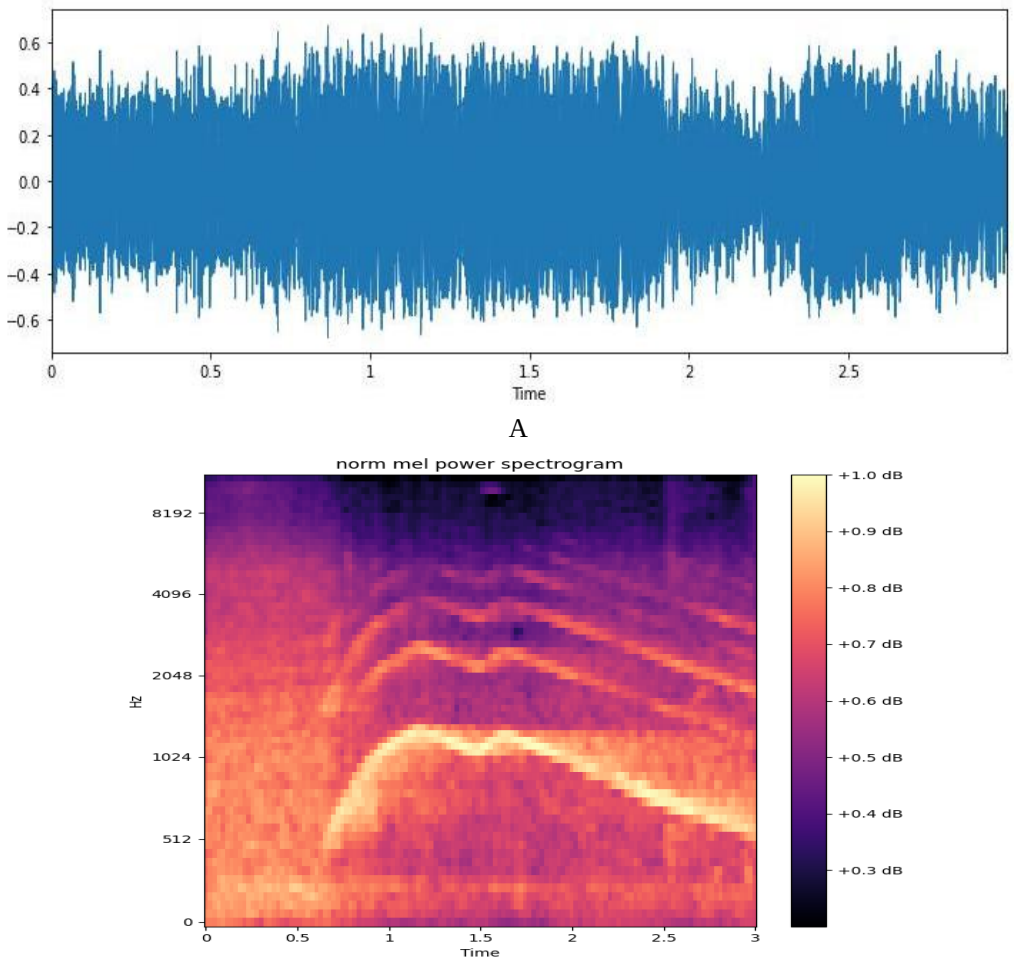


Figure 3- Sound Recognition Algorithm

5. Mel-spectrogram Extraction

Figure 4 shows the sound source data used in the experiment and the data converted to mel-spectrogram. Waveform has a one-dimensional information system that represents only amplitudes on the time axis. Mel-spectrogram has two-dimensional screen information by

representing amplitudes in color on the frequency and time axes. Therefore, when comparing information based on two criteria rather than comparing information on one basis, it becomes a basis for making more sophisticated judgments.[4]



Figuer 4- Visualization with sound data and Mel-spectrogram: A, Waveform of siren sound; B, Mel-spectrogram extracted from waveform A

When an autonomous vehicle is in an urgent situation, the sound recognition algorithm must make a quick judgment. Fig. 5 is a graph of mel-spectrogram extraction time and resolution which is from two-second audio samples. Seconds were rounded to the third decimal place. Values are measured five times and averaged to obtain consistent results. The output time of the algorithm can be expressed as the sum of the time taken to extract the mel-spectrogram and the time determined by the CNN. Although it is difficult to reduce the determination time of the CNN, it is possible to reduce the mel-spectrogram extraction time by reducing the sampling rate (n_mels) in the mel-spectrogram extraction process. However, the shorter the extraction time by reducing n_mels, the lower the resolution of the image and the lower the performance as learning data. In Fig. 5, since there is a minimum time that does not further decrease and its average time is about 1.48 seconds, the experiment was conducted with 64, which is the n_mels value that can maintain a proper resolution while minimizing mel-spectrogram extraction time.

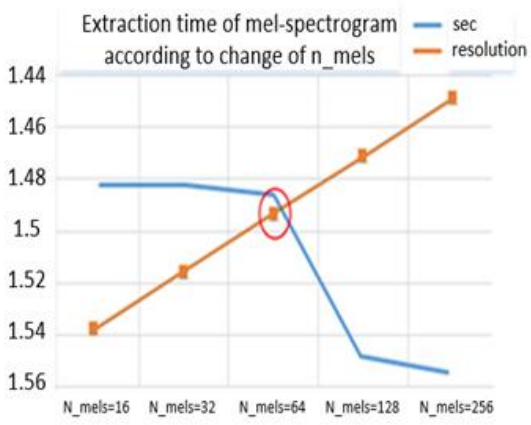


Figure 5- Extraction time of Mel-spectrogram according to change of n_mel

6. Sound recognition with CNN

Experiments were conducted to construct the actual CNN model and recognize sound source using vehicle driving sound and siren sound. A total of 385 learning data were used, with 188 siren sound data and 197 vehicle driving data. These two types of data are visually distinct and can be used a learning data for CNN. Fig. 6 shows the experimental result for CNN's binary classification of sound recognition. 'Epoch' represents the number of learning, 'acc' represents training accuracy for each epoch, and 'val_acc' represents verification accuracy for each epoch. Experiments were conducted to design binary classifications of siren and driving, and the results showed significant results with about 95% verification accuracy at 35 epoch. High accuracy was obtained by using 385 learning data, and efficient results were obtained by using relatively small amount of learning data.

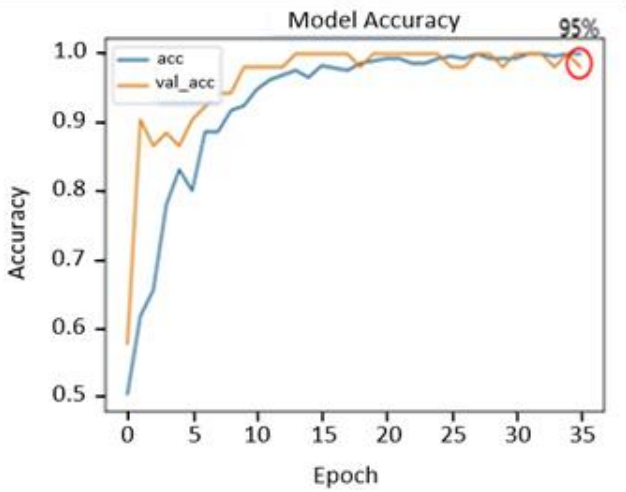


Figure 6- Experimental result for CNN's binary classification of sound recognition

7. Conclusion

In this paper, we have conducted a study on the algorithm that recognizes and judges sound source using CNN. To conduct this study, microphones are attached to receive sound information in the autonomous vehicle. The received sound information is then converted to visual information and the Mel-spectrogram was utilized, which expands one-dimensional sound information to two-dimensional information. The value of n_mels = 64 is suggested to minimize the extraction time of Mel-spectrogram with explanation of extraction process. Through the experiment, 95% accuracy was obtained through CNN learning. As a result, we conclude that sound source recognition through CNN can effectively recognize dangerous situations around autonomous vehicles. This paper implements the voice recognition system for visual installation of autonomous vehicles using CNN, and it is verified that recognizing sounds that can fly in road traffic is very accurate, and based on this, it is expected that the stability of autonomous vehicles can be improved further.

8. Acknowledgement

This study was supported by the Research Program funded by the SeoulTech (Seoul National University of Science and Technology).'

References

[1]. Y. U. Kim, "A Study on Multiple Sound Source Recognition Algorithm to Improve the Abient Detection Ability of Autonomous Vehicle", 2019

[2]. Sarma, C. A. ., S. . Inthiyaz, B. T. P. . Madhav, and P. S. . Lakshmi. "An Inter Digital- Poison Ivy Leaf Shaped Filtenna With Multiple Defects in Ground for S-Band Bandwidth Applications". International Journal on Recent and Innovation Trends in Computing and Communication, vol. 10, no. 8, Aug. 2022, pp. 55-66, doi:10.17762/ijritcc.v10i8.5668.

[3]. Towards Data Science. http://towardsdatascience.com

[4]. O'REILLY, "Hands on Machine Learning with scikit learn & tensorflow", Oreilly & AssociatesInc, 2017.

[5]. Kumar, S., Gornale, S. S., Siddalingappa, R., & Mane, A. (2022). Gender Classification Based on Online Signature Features using Machine Learning Techniques. International Journal of Intelligent Systems and Applications in Engineering, 10(2), 260–268. Retrieved from https://ijisae.org/index.php/IJISAE/article/view/2020

[6]. K. H. YU, "Deep-Learning based scream recognition and implementation of scream alarm system", 2018.

[7]. Silveira, M. R. ., Cansian, A. M. ., Kobayashi, H. K. ., & da Silva, L. M. (2022). Early Identification of Abused Domains in TLD through Passive DNS Applying Machine Learning Techniques. International Journal of Communication Networks and Information Security (IJCNIS), 14(1). https://doi.org/10.17762/ijcnis.v14i1.5256

[8]. Jalaj Thanaki, "Python Natural Language Processing: Advanced machine learning and deep learning techniques for natural language processing", Packt Publishing, 2017.