October 2020

Spatial Stereo Sound Source Localization Optimization and CNN Based Source Feature Recognition

Cong Xu

University of South Florida

Spatial Stereo Sound Source Localization Optimization and CNN Based Source Feature Recognition

by

Cong Xu

A thesis submitted in partial fulfillment
of the requirements for the degree of
Master of Science in Electrical Engineering
Department of Electrical Engineering
College of Engineering
University of South Florida

Major Professor: Ravi Sankar, Ph.D.
Alexandro Castellanos, Ph.D.
Kwang-Cheng Chen, Ph.D.

Date of Approval:
October 8, 2020

Keywords: 3D Localization, Speech Signal Processing, CNN Deep Learning, Stereo Sound

Dedication

To my mom,

Chunling Wang

and

My lovely wife,

Meiling Kang

Acknowledgments

I would like to acknowledge my mentor and advisor Dr. Ravi Sankar for his wise guidance and generous support in my project. At the same time, I would like to acknowledge Dr. Alexandro Castellanos and Dr. Kwang-Cheng Chen for their valuable time in participating in my defense committee. In addition, I would also like to thank Sai Bharadwaj Appakaya, who helped me a lot in my project and provided a lot of pertinent suggestions. Finally, I want to thank my family and my wife for their help during my life and research project. They have always been my motivation to keep improving myself.

I also would like to acknowledge the sources that provide me with audio data material, including SoundBible, Freesound, GameSound, ZapSplat, AudioMicro and SoundGator. The free audio material of these companies enables the deep learning model of the audio recognition system to be realized.

Table of Contents

List of Tables iii

List of Figures..iv

Abstract..v

Chapter1: Introduction..1
1.1 Background...1
1.2 Challenge 5
1.3 Motivation and Research Objectives 7
1.4 Thesis Expectations...7
1.5 Thesis Organization 9

Chapter 2: Literature Review...10
2.1 Relevant Research Background 10
2.1Vocalization and Mechanical Wave Theory..11
2.2Sound Source Feature Recognition...12

Chapter 3: Sound Source Localization By TDOA Methods...13
3.1 Basic Concept and Mathematical Model of TDOA...13
3.2 Spatial Stereo Capture Sensor Array..15
3.3 Acoustic Signal Processing and Stripping in Stereo 19

Chapter 4: Source Characteristics Identify By CNN Deep Learning....................................21
4.1 Basic Concept and Mathematical Model of CNN Deep Learning...................21
4.2 Sound Source Recognition with CNN Deep Learning Methods.....................24
4.3 The Comparison of CNN with other recognition systems 25

Chapter 5: System Design and Implementation...27
5.1The Architecture Design of Localization and Recognition System..................27
5.2 Pre-processing before Stereo Signal Localization 32
5.3 Implementation of Sound Source Localization Subsystem 33
5.4 Implementation of Sound Source Recognition Subsystem..............................35
5.5 Results comparison with other researchers' SVM research 39

Chapter 6: Conclusion and Future Work 42
6.1 Conclusion 42
6.2 Future Recommendation 43
6.3 Summary 44

References 45

List of Tables

Table 1: Training category for CNN source recognition 24

Table 2: Classification result comparison between CNN and SVM.. 41

List of Figures

Figure 1: Data Processing System for Sound Source Localization and Recognition System 2

Figure 2: TDOA sources localization in X-Y plane space 13

Figure 3: 3D TDOA Localization by traditional microphone array 16

Figure 4: Optimized 3D TDOA sensor matrix 17

Figure 5: CNN deep learning data processing diagram 21

Figure 6: Spectrogram image preprocessing procedure 22

Figure 7: System architecture for sound source identification and localization 28

Figure 8: Implementation of 3D microphone sensor array 29

Figure 9: Hardware connection mode and A/D conversion of sensor array 30

Figure 10: Use trigger mechanism to identify sound source segments 31

Figure 11: Noise cancellation by Low-pass FIR filter 32

Figure 12: TDOA calculation demonstration when the bit rate is 115200 bps 34

Figure 13: Deep learning network architecture based on supervised learning 35

Figure 14: Pixelated spectrogram for training model 36

Figure 15: CNN convolution system design 37

Figure 16: CNN sound source recognition model training results 38

Figure 17: Model testing result by dog audio 39

Figure 18: SVM classification system data structure 40

Figure 19: SVM classification result by using general sounds 41

Abstract

In the process of propagating as a carrier of information in space, in addition to transmitting the information itself, the acoustic signal also contains the position information of the sound source itself and its related physical characteristics. Acoustic signal uses the medium (such as air, water, steel, etc.) in the space to transmit mechanical vibration and longitudinal waves from the sound source to the outside world. The traditional single audio collection device cannot collect position information and sound source characteristic information. Therefore, the signals processed in the audio signal processing process are all mixed source acoustic signals after spatial reverberation. The significance of studying the sound source recognition and positioning in the three-dimensional space is to help the computer to reshape the specific information of the sound source by using artificial intelligent acoustic processing, effectively separate a single sound source from the environment or synthesize a fine stereo source for virtual reality scene. For instance, in an acoustic environment containing multiple background noises, it is not possible to filter all the background noises from the received audio signal. At this time, the system can be used to perform deep learning of convolutional neurons on the information of the target sound source, and finally strip off the undesired sound signal. The realization of this research will be able to help better machine learning algorithms in the audio field and other fields of speech signal processing.

In this thesis, the location and recognition of the sound source will be achieved through two major parts. In the first part, the stereo signals in the three-dimensional space will be collected by the sensor array, and the correlation of the signals of each radio unit will be compared. After the comparison, the delayed signal will be measured by the time difference of arrival algorithm to obtain the position information of the sound source. In the second part, the original multi-source signal is integrated into a relatively independent unit signal by determining the location information of the audio, and the audio is subjected to cross-comparison in the time domain and the frequency domain after noise reduction processing. Then, the convolutional neural network is used to identify the target audio features. Finally, the results calculated by the two parts are combined to realize the analysis of the sound source of the acoustic signal.

The sound source recognition system that integrates sound source recognition and sound source localization can effectively identify active signals in background environmental noise in an indoor environment and obtain a good sound source recognition accuracy rate. The application of this system will be able to effectively help the computer system to perceive the surrounding environment and realize effective three-dimensional coordinate monitoring of specific sound sources. This research has a wide range of applications in acoustic recognition and sound source location and monitoring.

Chapter 1: Introduction

1.1 Background

Acoustic waves, as a mechanical wave transmission mode of physical shock, exist widely in all corners of the earth. People perceive various sound waves around the environment through their ears and auditory ossicles and use the brain to analyze various information contained in the sound waves to help humans perceive the surrounding environment. When an object emits sound, this mechanical wave often carries the physical characteristics of the sound source, so that people can quickly distinguish the specific characteristics and distance of the sound source after hearing the corresponding audio information. Acoustic positioning and sound source recognition have broad applications in passive positioning, especially when obstacles block the sound source's line of sight or people cannot visually observe the target. Sound source localization is a passive localization method that the sound source is only be measured and processed when a specific source emits sound. With the rapid improvement of chip performance and digital signal processing capabilities, real-time analysis of sound source information is technically possible. The sound source information is physically divided into two parts, namely the vocal feature information of the sound source and the spatial position information of the sound source. The vocal feature information refers to the resonance frequency, the energy intensity of the sound source, sound loudness, and the characteristics of the carrier medium when the sound-producing object undergoes mechanical oscillation. The spatial position information refers to the independent position information of the sound source in the three-dimensional space.

After the signal processing system collects and analyzes these two types of information, the deep learning network can use the sound waves received by the sensors to understand the characteristics of the sound source and locate the sound source.

Sound source information detection has a very wide range of applications in the Internet industry. In the field of the Internet of Things, a medium scale unmanned smart factory can simultaneously control the collaborative operation of more than one hundred numerical control equipment. When the equipment has a sound failure or a potential safety hazard, it often first emits abnormal sounds, such as electric sparks, friction sounds, or impact sounds. Traditional position sensors and vision sensors cannot provide effective early warning of hidden dangers, and major safety accidents often occur. Therefore, when the sound source information detection system intervenes, the fault information can be captured in real-time when the equipment emits abnormal noise, and the early warning information can be reported.

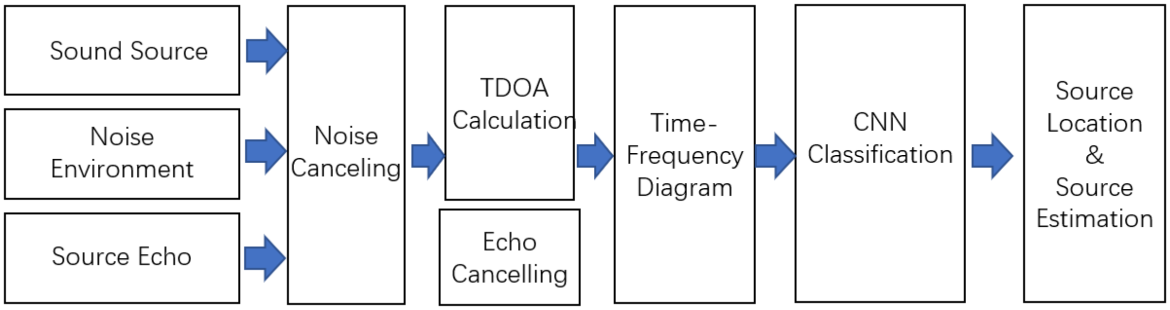


Figure 1: Data Processing System for Sound Source Localization and Recognition System

Figure 1 shows the processing flow of the acoustic system's capture and recognition of sound source signals including environmental noise and echo. As shown in Figure 1, the sound source information detection system is mainly composed of a sound source three-dimensional position detection unit based on the time difference of arrival (TDOA) principle and a sound source feature recognition unit based on convolutional neural network(CNN)algorithm. In terms of sound source localization, researchers usually use the generalized cross correlation (GCC) method to solve the different delays of the homologous signal in the sensor array. Examples of solving TDOA using GCC can be found in [1].In order to truly restore the accuracy of the sound source detection system, the test audio used in this experiment is the real sound received by the sensor array indoors, and the system response simulation of similar frequency is carried out with reference to the received audio in the simulation system.

As a branch of the audio recognition system, the sound source detection system has broad applications in the recognition of audio signals. This technology can extract independent sound source information from environmental reverberation and various environmental noises and helps the speech recognition system to obtain more useful sound source data before data processing. The sound source recognition system can help computer software reconstruct the real-time information and three-dimensional position coordinates of the occurring objects, and it has an especially important application in sensor fusion in the field of Internet of Things. At the same time, for the speech recognition system, the sound source recognition system can help the speech recognition software perform the preprocessing of the sound source in the audio recognition process and can effectively strip the useless echo and environmental noise. For the field of immersive virtual reality technology, the sound source recognition system can help the virtual reality system to better restore the real-life information and sound source characteristics, thereby helping the system to achieve a better immersive experience.

The sound source information recognition system is mainly composed of two main parts. The first part is the GCC-TDOA calculation unit to calculate the specific position of the sound source in the three-dimensional space. The second part is the CNN sound source recognition unit used to analyze the spectral characteristics of the sound source and classify the sound source. In the first part, the audio signal first uses digital signal processing technology to perform signal preprocessing such as noise reduction and reverberation suppression on the background noise through a dynamic digital filter, and then uses the GCC algorithm to solve the TDOA of the audio signal containing multiple audio tracks , and output the three-dimensional position coordinates of the sound source. In the second part, the multi-source signal is first combined in phase with the TDOA value calculated in the previous step and merged into a whole independent audio track to make up for the lack of sound source information due to insufficient sound reception or system noise reduction . Then, on the basis of synthesizing the sound track, the convolutional neural network is used to identify and classify the spectrogram of the sound source audio information, and the supervised learning method in the deep learning algorithm is used to distinguish different sound sources, and the system is trained to finally recognize the sound source information. The sound source information includes the collision sound, friction sound, and vibration sound of different materials, the sound of animals or human voice.

The goal of this thesis project is to establish an audio processing system that can locate and recognize sound-producing objects in an indoor environment. This system is expected to be able to achieve the following functions and have the following application prospects:

High robustness is mandatory as the indoor environment usually contains environmental noise, echo and multi-source reverberation. For the system to collect sound source information as accurately as possible in the presence of various noises, the system must have high system robustness in a complex environment to ensure the reliability of system input data. This function allows the sound source recognition system to be used normally in factories with noisy backgrounds. Operators can use the sound source recognition system for industry product line monitor Internet of Things system and report early warning information in real-time when the equipment emits abnormal noise.

Real-time system offers a low latency as an important factor to ensure the accuracy of sound source positioning. When the sound source object emits sound, the sound wave will produce a time difference of several milliseconds between being transmitted to different sensors. Therefore, the system needs to process the audio signal to ensure the accuracy of the TDOA solution. At the same time, the real-time system can effectively monitor moving objects. This feature is mainly used in the field of autonomous driving and security surveillance. When the visual detection system enters the blind area of vision, the real-time sound source positioning system can help the security system to notice the sound source located outside the line of sight, and use the sound source recognition to match whether the sounding object poses a threat to the vehicle or property.

There are many factors that interfere with the accuracy of the recognition system in a real environment, the most important of which is the indoor building echo and system noise reduction, which make the signal lose important audio characteristics. In order to improve the accuracy of the recognition system, it is required that the system can automatically correct incorrect or missing audio source information. This feature can enable the sound source detection system to improve the system's identification ability and accuracy as much as possible even when the sound source detection system interferes with the complex building environment or noise.

1.2 Challenge

At present, there are many related types of research on sound source detection, but most of them are limited to a single sound source ranging or orientation determination. Few research teams can completely extract all the sound source information covered by the sound, such as combining the sound source location to understand the material, type, species, movement state of the sounding object. As the application of machine learning algorithms, there is a big difference in the technical realization of sound source information recognition and visual information recognition. The more significant difference is that the sound source information is solved indirectly instead of being able to be identified directly on the image like visual information recognition. Therefore, the sound source identification is more difficult.

The sound source needs to be extracted from a large amount of background noise and system reverberation, which requires remarkably high audio signal processing efficiency. In addition, there are few people involved in the field of algorithms that combine real-time sound source localization and adaptive sound source audio correction, so there are few references that can be used during the experiment. These are undoubtedly the problems that need to be overcome for this experiment.

In order to obtain the three-dimensional position information of the sound source, multiple sets of audio sensors need to be used to form a sensor array. To improve the accuracy of position information, it is required to extend the adjacent distance between sensors as much as possible within the allowable range to obtain a larger TDOA value. Such a sensor arrangement will bring about the problem of data fusion between the sensors. Different sensors will lose the detailed information of the sound source due to the weak sound intensity during the sound reception process due to different positions. Therefore, in the process of sound source identification, it is necessary to fuse the information collected by different sensors and sort out a soundtrack signal covering all sound source information. This process is undoubtedly very challenging.

1.3 Motivation and Research Objectives

The motivation of this project to study sound source information recognition and localization is to use deep learning algorithms and to apply the convolutional neuron algorithm originally used for visual information recognition to the acoustic field. To achieve this goal, collecting and analyzing the frequency spectrum information of the sound source is the main work of this research.

This research has a rich application range and broad prospects in the field of acoustics, especially the use of artificial intelligence algorithm training system to automatically recognize the information characteristics of the sounding object and the specific position in the three-dimensional space. Acoustic detection is an extension and supplement of visual detection, which allows people to perceive sound source information beyond obstacles through objects, which plays a particularly important role in the field of sound wave visualization. At the same time, this technology can also provide visual alternatives based on sound source localization for many hearing-impaired people, helping these groups detect and warn of dangerous sound sources. It is foreseeable that the future of signal processing will be more innovations and applications of sensor fusion and intelligence, so this research is a direction worthy of in-depth exploration and innovation by researchers.

1.4 Thesis Expectations

Regarding my thesis of sound source information recognition, the expected goal is to achieve detection, location, and recognition of sound-producing objects, and at the same time provide an effective solution for sound source visualization. To realize this idea, knowledge in the field of sound source localization and recognition, and convolutional neural networks will be used. At the same time, the research on this subject can also provide experience and background knowledge for:

- Digital signal processing: The main signal processing methods used in the research of this subject most come from A/D conversion, digital filtering, coding and decoding techniques, and sampling in digital signal processing.
- Noise reduction technology: The processing of noise signals is the key to the success of the experiment. Reducing the impact of environmental noise on sound source determination can improve the ability to design noise reduction systems and have a deeper understanding of audio signal processing.
- Artificial intelligence algorithms: Experiments to try various classifiers, cross-validation, and integration methods can enrich the actual combat experience of deep learning algorithms and strengthen the optimization ability of audio artificial intelligence recognition algorithms.
- Parallel signal processing of multi-sensor arrays: Use multiple groups of sensors to form a sound receiving array and perform parallel signal processing on audio information. Parallel processing of multiple sets of signals not only speeds up the signal processing capability but also implements the processing of the signal fusion and synchronization problems that are faced in sensor networking. At the same time, experiments have also enhanced the relevant experience in sensor array design.

This research will use machine learning algorithms to extract unknown sound sources in the environment and identify the location information and sound source characteristics of the sound source, and the sound characteristics of the sound source have been detected. The realization of this technology will have a very wide range of applications in the field of sound source detection and virtual reality technology in the future. The new system that integrates three-dimensional sound source localization and sound source recognition can identify and track sound sources at the same time. Compared with pure sound source recognition, the sound source recognition system can better process the moving sound source in real time with the assistance of the sound source positioning system, thereby ensuring the accuracy of dynamic sound source recognition. The biggest advantage of this research is that it can track and identify the characteristic information of the sound source across visual obstacles, and continuously increase the accuracy of the system's sound source recognition through its deep learning algorithm.

1.5 Thesis Organization

Next chapter provides the related research background of this research and results of colleagues in related fields, and briefly introduces the theoretical basis of acoustic localization and sound source recognition. Chapter 3 mainly introduces the sound source localization algorithm and the characteristics of the time difference of arrival (TDOA) in two-dimensional space and specific application of TDOA in the sensor array in the three-dimensional space. Chapter 4 introduces how the sound wave recognition system classifies and recognizes audio information. The technology used in this experiment is the deep learning algorithm of convolutional neural networks, and the construction and design of convolutional neural networks are provided in this chapter. Chapter 5 introduces in detail how this experiment uses the previously described knowledge and concepts to realize the three-dimensional sound source positioning and sound source recognition and provides the recognition results. Finally, Chapter 6 summarizes and discuss related applications of this technology.

Chapter 2: Literature Review

2.1 Relevant Research Background

Sound source recognition technology has been a hot research subject in the field of signal processing. Yue et al and their research team demonstrated in [1], the algorithm and simulation results of using convolutional neural networks to locate sound sources in three dimensions. Their team used the GCC-PHAT algorithm to solve the TDOA value of the sound source signal and simulated the use of a six-microphone array to detect the sound source. In [2], Martin from MIT proposed a sound source recognition theory using the principle of sound excitation and resonance. He developed an algorithm by studying "simple statistical pattern-recognition techniques", enabling the computer to distinguish 25 different musical instruments. On the other hand, Boes et al [3] referred to the speech signal processing mechanism of humans and other organisms and developed a sound source recognition model of a recurrent neuron neural network with a three-layer neuron structure. The above research is to direct process-specific sound source signals, but in reality, it is necessary to strip and extract the received signal from a single sound source. One of the effective ways to strip the sound source is to measure the inclusion ratio of a single sound in the audio based on the Expectation-Maximization (EM) algorithm [4]. The advantage of this method is to continuously optimize the deep learning model in an iterative manner, and finally obtain more accurate information to distinguish different sound sources. In terms of counting the number of sound sources, Yamamoto et al. [5] proposed the use of support vector machines to count the number of different sound sources in the reverberation sound field, and this method is suitable for multi-microphone input. In the process of investigating the separation of sound sources, it is also found that speech recognition can also be applied to the recognition of moving sound sources [6].

2.2 Vocalization and Mechanical Wave Theory

Periodic oscillations of objects will produce mechanical waves that spread to the surroundings in the medium. These mechanical waves are the main source of sound. Due to the different ways of emitting mechanical waves, the sound can be divided into human voice, animal chirping, mechanical vibration, and blasting sound. Higashimoto and Sawada [7] exemplified the formation of the human voice and reconstructed the speech system mechanically. In their research, different sounds are produced because of the difference in pitch and the oscillation frequency of the sound source. Mel Frequency Cepstral Coefficient (MFCC) and Perceptual Linear Prediction (PLP) coefficients are often used to evaluate speech recognition systems, and these parameters can also be used for animal vocal analysis between different species [8].When dealing with the information of these different sound sources, Nadeu et al. [9] proposed a model for Acoustic Scene Analysis (ASA) and successfully applied it to scenes where sound sources overlap, and multiple sets of sound sources are simultaneously identified.

Multi-channel stereo uses signal delay and signal attenuation to reproduce sound sources in space. The two parameters of inter-channel time difference (ICTD) and inter-channel level difference (ICLD) are important data often used in stereo reproduction technology that in digital processing systems these two parameters are used to adjust the multi-channel sound source level and phase delay [10].

2.3 Sound Source Feature Recognition

The traditional sound recognition system uses the related technology based on the Mel frequency cepstral coefficient [14]. This method can effectively distinguish the acoustic and prosodic features in the nonlinear audio signal, but it requires artificial recognition and processing in the process of processing. Processing acoustic features, so the Mel frequency cepstral coefficient method is less objective than machine learning algorithms. The use of convolutional neural network to perform adaptive feature extraction of audio signals is a more effective means of sound source recognition. Zhang et al. [15] conducted an in-depth discussion on the emotion recognition of the speaker in audio files. They used multi-layer convolutional neurons to estimate the emotion of the speaker in 1200 audio samples in the CASS speech database and used the same data. A horizontal comparison of support vector machine (SVM) was carried out. The results show that the sample recognition accuracy based on CNN can be as high as 95.5%, which is higher than the sample recognition accuracy rate based on SVM technology (92.5%).

The use of CNN deep learning has also been applied in the recognition of bird sounds. In 2018, Incze et al. [16] classified and recorded bird calls and trained a deep learning network for bird acoustic recognition. They used spectrograms to record the time-domain and frequency-domain features of bird calls, and then input the spectrograms corresponding to these calls into the convolutional neuron network in picture format for pre-training. The trained deep learning model can achieve a more reliable bird recognition rate.

Chapter 3: Sound Source Localization By TDOA Methods

3.1 Basic Concept and Mathematical Model of TDOA

Acoustic positioning and electromagnetic wireless positioning have many similarities. In the field of wireless positioning, time difference of arrival (TDOA) is a common positioning method and has a wide range of applications. The main principle of using the TDOA to calculate the position of the emission source is that the emission source radiates outward in a straight line in the medium. Therefore, the position when reaching different detectors will produce corresponding amplitude and time differences due to path attenuation and arrival speed that the resulting time difference is the TDOA. Figure 2 shows that in the 2-D plane space, when the acoustic signal generated by the unknown source X is observed by the observation points A1, A2, and A3 at three different positions on this plane, three sets of arrival time differences will occur.

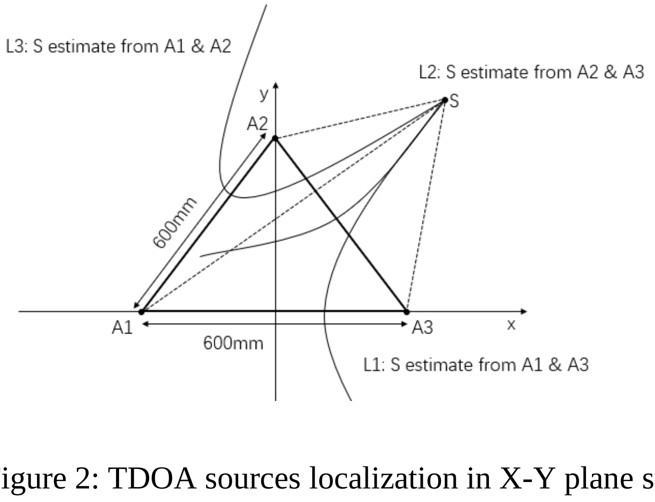

L3: S estimate from A1 & A2
L2: S estimate from A2 & A3
y
S
A2
600mm
A1
600mm
A3
x
L1: S estimate from A1 & A3

Figure 2: TDOA sources localization in X-Y plane space

At this time, the unknown source S can be determined by the hyperbolic equation the distance between each point and solve the plane coordinates of the unknown source.

Since the focus of this experiment is on sound source detection in an indoor environment, the sound source propagates oscillations through the air. The sound propagation [11] in the air is affected by the temperature and air density and changes accordingly. The specific relationship between sound velocity and medium and temperature is shown equation (1).

$$v_0 = \sqrt{\gamma RT} \qquad (1)$$

Among them, v_0 refers to the propagation speed of sound at a specific temperature, and γ specifies the ratio between the specific heat of pressure and the specific heat of constant volume. In the calculation process, the value of air is 1.4. R represents the gas constant. K represents the temperature in Kelvin (K) in the environment. In a dry air environment, the propagation speed of sound is v_0 =331.6+0.6 K-273.15 (m/s).

According to the above function, when the room temperature is 25℃, the sound propagation speed in dry air is 346.6 meters per second, that is, 346.6 millimeters per millisecond. The three sensors in the planar sensor array designed in the experiment constitute an equilateral triangle, with a side length of 600 mm. Then the coordinates of the three points are A1(-300,0), A2(0, $300\sqrt{3}$), and A3(300,0).Using TDOA to measure the time difference between two adjacent points, these three points can form a total of C (3, 2) = 3 sets of mutually independent TDOA values. The calculation method of TDOA value is shown in equation (2).

$$TDOA = v_0 \Delta t(A_{n+1} - A_n) \qquad (2)$$

The TDOA value solved in this equation refers to the distance difference between the sound source and the sensor estimated based on the time difference measured between two adjacent sensors A_n and A_{n+1}. In a random environment, due to the different sequences of sound propagation to different detection points, TDOA value will produce negative values in some cases. The positive and negative values of these values can help us predict the quadrant range of the sound source and narrow the search range. From the definition of the hyperbolic equation, it can be seen that the positions with the same TDOA value between any two detection points can form a set of hyperbolic trajectories, and the matching single curve rule can be screened out according to the sign of TDOA. The specific relationship is as shown in the equation (3).

$$\sqrt{(x-x_1)^2+((y-y_1)^2)}-\sqrt{(x-x_2)^2+((y-y_2)^2)}=TDOA(A_{n+1}-A_n) \qquad (3)$$

In this equation, x and y represent the possible trajectories of the sound source coordinates, x1 and y1 refer to the coordinate position of the detector A_{n+1}, and x2 and y2 refer to the coordinates of the position of the detector A_n. When the TDOA value is positive, it means that the sound source is closer to the detector A_{n+1}, and when the TDOA value is negative, it means that the sound source is closer to the detector An. Thus, irrelevant trajectory curves can be eliminated. Finally, overlap all the trajectories and find the common solution of all related trajectories to determine the position of the sound source on the plane. This method can also be extended to solve the TDOA of the detector in three-dimensional space and generate C(n, 2) groups of different TDOA values to determine the specific location of the sound source in the three-dimensional space and the specific details will be introduced in the next section.

3.2 Spatial Stereo Capture Sensor Array

Based on the TDOA positioning method in the plane space, the above method can be extended to the three-dimensional space by establishing a three-dimensional sensor array, and the position of the unknown sound source in the three-dimensional space can be measured.as described in section 3.1, when the number of sensors increases, the amount of calculation of the system will also increase exponentially. Therefore, in order to reduce the response delay of the system and finally realize real-time sound source localization, this experiment design uses the least number of sensors to realize the rapid response of the system, and optimizes the system architecture to reduce the amount of system calculations. Figure 3 shows the specific method of 3D sound source localization and the arrangement of the sensor array.

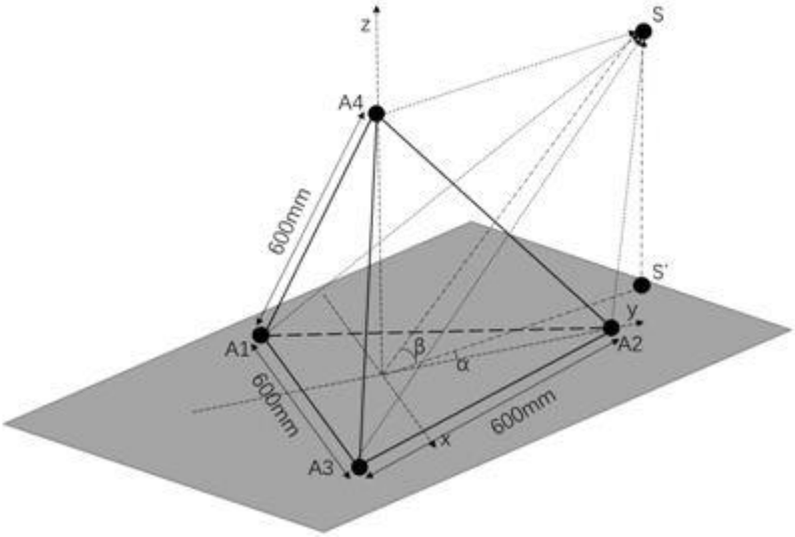


Figure 3: 3D TDOA Localization by traditional microphone array

In this experiment, a space regular tetrahedron structure is used to build a sensor array to minimize the number of sensors used. The three-dimensional extension of the two-dimensional TDOA solution using the traditional architecture is shown in equation (4).

$$\sqrt{(x-x_1)^2+(y-y_1)^2+(z-z_1)^2}-\sqrt{(x-x_2)^2+(y-y_2)^2+(z-z_2)^2} = TDOA(A_{n+1}-A_n) \quad (4)$$

Solving the position of the unknown sound source requires solving the TDOA values of all signals in pairs, and finally combining to solve a common solution. For the minimum sensor array in this experiment, a total of 4 microphones are needed to detect sound sources in three-dimensional space, so the resulting paired TDOA value is C(4,2)=6 groups. The 6 sets of TDOA pairs are A1-A2, A1-A3, A1-A4, A2-A3, A2-A4, and A3-A4. Bring these 6 sets of data into equation (4) to find the three-dimensional coordinates of the location sound source.

Although the above method minimizes the number of detectors, the method of simultaneously seeking common solutions for six groups of TDOA requires a lot of mathematical calculations, which will increase the system delay. In order to ensure that the system can process the signal in real-time, reducing the operation delay is the most important part of designing the system. Therefore, during the actual operation of this experiment, we optimized the original traditional TDOA algorithm to reduce system delay. Based on the original regular tetrahedron structure, the corresponding positions of the sensors have been re-arranged as shown in Figure 4:

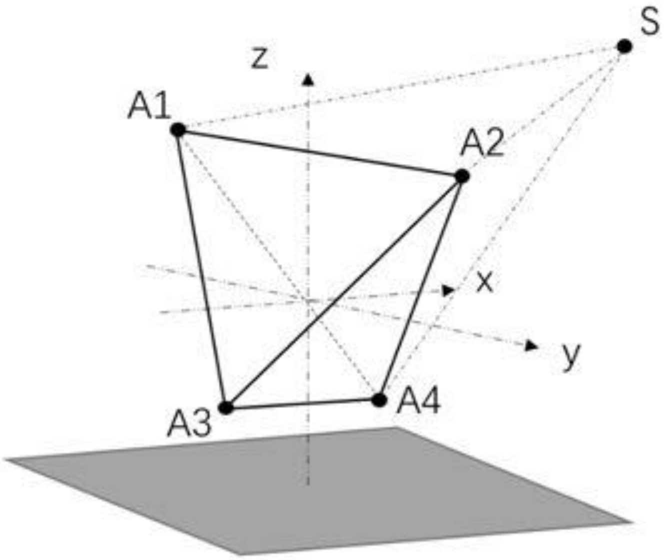


Figure 4: Optimized 3D TDOA sensor matrix

The difference between the optimized TDOA microphone sensor and the previous sensor array is that the regular tetrahedrons are arranged in a vertical manner in different planes, where the projections of the detection points A1 and A2 coincide with the y-axis in the Cartesian three-dimensional coordinates, A3 and The projection of A4 finally merges with the x-axis, and the midpoints of the two sets of sensors are arranged symmetrically about the three-dimensional origin. After determining the structure of the sensor array, there are two situations that will appear in the process of measuring the sound source signal and need to be treated differently. The first case is when the TDOA value of the sound source signal arriving at the observation point does not contain zero, that is, the sound source does not reach two or more observation points at the same time. The second case is that the data contains a TDOA value of zero which at least one group (two sensors) simultaneously receives the signal from the sound source.

For the first case, we can find out the sequence and time difference of the sound source signal arriving at the sensor detection point. According to the position of the sensor to lock the subspace range of the sound source, the hyperbolic equation set consisting of the three sets of TDOA values that received the signal first can be found, and a common solution can be found. This common solution is the coordinates of the sound source in the three-dimensional space. For the second case, the existence of a zero TDOA value means that the sound source is at the boundary of the subspace (the TDOA value between two sensors is zero) or a direction facing the sensor's endpoint (the TDOA value between the three sensors is zero). When the sound source is at the boundary of the subspace, only the two observation points need to be solved on this plane because the sound source and the origin-TDOA non-zero observation point are on the same plane; when the sound source is facing a certain sensor (i.e. when the TDOA value between the facing sensor and the other sensors is equal), the TDOA value can be added to equation (4) to get the position of the observation point. It can be seen from the above that by preprocessing and categorizing data, the difficulty of calculation can be effectively reduced, and thus the delay caused by the system processing data can be reduced. In this way, the sampling frequency of the system can be guaranteed to the greatest extent to improve the response speed of real-time sound source positioning.

3.3 Acoustic Signal Processing and Stripping in Stereo

Another important technology that can realize sound source localization and feature analysis is to extract the sound source information from the environment and strip it out of the background noise. The biggest difference between sound source information and background noise is the relevance of TDOA. Background noise usually includes ambient sound field reverberation and Gaussian noise. These sounds do not have obvious sound source characteristics because they are not emitted from a single sound source. Another situation is that the reflection of the indoor wall to the sound source will produce a system echo, which has characteristic information like the sound source and contains a certain amount of time delay. However, because the echo is absorbed by the sound wave through a longer path and obstacles, the intensity of the echo will be significantly lower than that when the sound source is directly transmitted to the sensor. Based on the above-mentioned differences in acoustic characteristics, we will discuss the specific strategies and algorithm details for extracting sound source information in this section.

The way to determine whether an audio signal contains an active signal is to calculate the correlation coefficient between the various signals in the sensor. The calculation method of the correlation coefficient [13] is shown in equation (5).

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}} \tag{5}$$

The above equation is to solve the correlation of two sets of data containing similar information. By calculating the correlation coefficient, it is better to find out whether the signals collected between the sensors are correlated. In the process of processing audio signals, it is necessary to perform A/D conversion and sampling on the continuous voltage signal from the acoustic-electric transducer, so the signals processed by the signal processing system are all sampled discrete digital signals. For discrete digital signals, equation (5) can be modified into equation (6) to facilitate the signal processing system to process discrete data.

$$Corr[x, y] = \frac{\sum_{n=0}^{N-1} x[n]y[n]}{\sqrt{\sum_{n=0}^{N-1} x^2[n]}\sqrt{\sum_{n=0}^{N-1} y^2[n]}} \tag{6}$$

It can be seen from section 3.1 that the maximum sensor distance difference of this system is 600 mm, and the maximum TDOA tolerance allowed by the system can be calculated to be 1.73 milliseconds. Therefore, the maximum TDOA allowed for all active signals is the upper limit of the above value. If the obtained cross-correlation time difference is greater than 1.73 milliseconds, then it indicates that the signal is an echo sound wave reflected from a wall or obstacle, and this set of data is eliminated. Based on this rule, we can compare the four sets of signals with cross-correlation coefficients and filter out the active signals within the range. Setting the upper limit of TDOA can greatly reduce the possibility of echo signals being collected incorrectly, thereby reducing the error rate of the system.

Chapter 4: Source Characteristics Identify By CNN Deep Learning

4.1 Basic Concept and Mathematical Model of CNN Deep Learning

The convolutional neural network is an efficient artificial intelligence algorithm in the field of deep learning. The CNN deep learning algorithm mainly performs convolution operations on image conversion information and filters, and continuously refines the graphic feature information contained in the picture. Finally, the neural network is used to cross-compare the feature information to finally estimate feature inside the image contained by the system. The realization of the CNN algorithm mainly passes through three important parts, which are image preprocessing, image convolution and pooling, and neuron network connection. The specific operation process of CNN deep learning is shown in Figure 5.

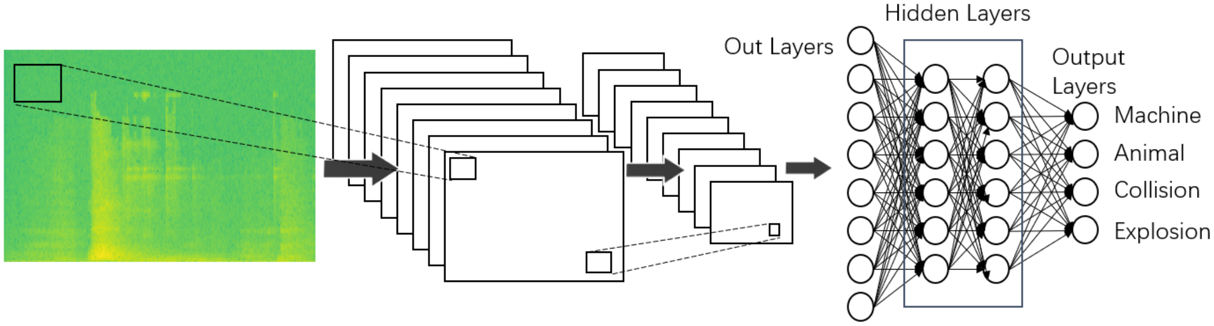


Figure 5: CNN deep learning data processing diagram

While evaluating image preprocessing, the process of feature extraction of audio files, it is not only relevant to understand the amplitude change of the audio at a specific moment, but also the change of the sound wave frequency. Therefore, in the process of analyzing the audio characteristics, the characteristic information cannot be obtained directly from the audio signal itself. In order to solve this problem, this thesis mainly obtains the spectrogram of the audio segment (data frame or block) by performing a short-time Fourier transform on the audio file in the process of studying audio features and performs feature matching and classification on the spectrogram. When the audio clip is converted into a spectrogram through a sequence of short-time Fourier transform, because the audio file contains noise information and there are many non-periodic signals in the sound source, the generated image has low contrast, which increases the difficulty of feature recognition. In order to improve the recognition efficiency of the system and increase the accuracy of sound source information recognition, it is necessary to preprocess the image before recognizing the image.

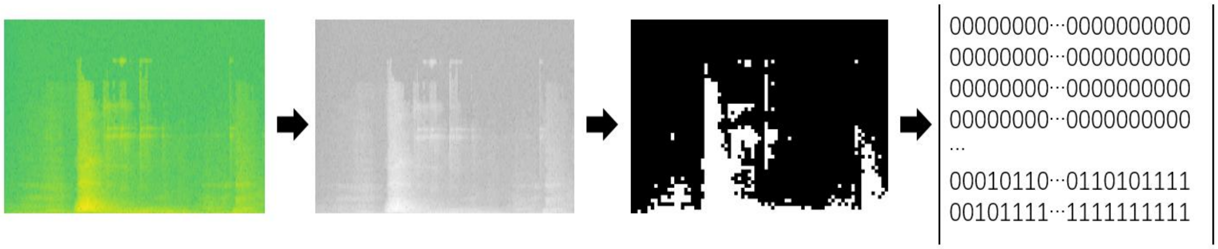


Figure 6: Spectrogram image preprocessing procedure

As shown in Figure 6, the spectrogram generated by the audio clip is first converted from a color image to a grayscale image. Because the original image uses RGB labeled pixels to display the audio frequency and energy intensity, the image needs to be converted into a grayscale image to convert the 3-layer matrix into a single layer matrix. After obtaining the grayscale spectrogram, the next step is to sharpen the spectrogram and extract features. In order to make the graphics more contrast, through pixilation, each pixel only retains two variables, 0 (black) and 1 (white). In this step, to maximize the contrast of image features and retain feature information, we set a threshold of 200 in the interval from 0 to 255 to eliminate unimportant noise information in the image. Finally, the data stored in the file is converted into a matrix containing only 0 and 1 in units of 25 pixels for the data input of the subsequent convolutional neuron deep learning algorithm.

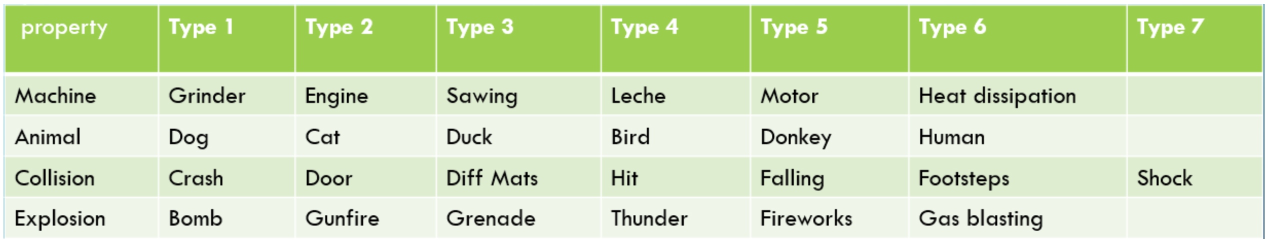
Image convolution and pooling comprises the convolutional neural network loops through three basic parts [13] to pick up graphics features. The first part is the data input. When a picture is put into the convolution system, it is called the input layer. In this layer, the picture contains all the original information or graphic features refined by the previous layer. The second part is the weight layer used for graphics convolution operations. The input layer extracts corresponding features and reduces the graphics matrix by performing convolution operations on the input layer and the weight layer and the reduced image layer is called the convolution layer. The third part is the pooling layer. The result of the convolution operation is still a multi-dimensional array. In order to further reduce the range of features, the image is convolved with a specific sub-level as a unit (such as a small 3x3 matrix) to select the maximum value or average value is used as the characteristic value of the sub-level. This process is called pooling. For pictures with large number of pixels, multiple convolution cycles are usually used to narrow the range of graphic features as much as possible.

Neural network subsystem includes algorithm that is a weighted decision network developed with reference to the information processing mechanism of human neurons [13]. In this network, the data is multiplied by the input parameters and the feature weights on each independent neuron, and then after the sum is processed by the activation function at the node, the output result is generated. Usually, a neural network system consists of an input layer, an output layer, and several hidden layers used to process data after weight calculation. The output result will be compared with the labeled data through supervised learning and the error value will be corrected to ensure that the system can obtain a higher accuracy value after multiple iterations of the calculation.

4.2 Sound Source Recognition with CNN Deep Learning Methods

In the model training process of analyzing the sound source, the difference in acoustic characteristics will directly affect the accuracy of the neural network. Therefore, in the process of selecting training resources, four groups of acoustic databases with obvious characteristic differences were defined, namely machine running sound, animal sound, collision sound, and explosion sound as shown in Table 1. Each category contains several common sound source types with different acoustic characteristics, the specific implementation details will be described in Chapter 5.

Table 1: Training category for CNN source recognition



property	Type 1	Type 2	Type 3	Type 4	Type 5	Type 6	Type 7
Machine	Grinder	Engine	Sawing	Leche	Motor	Heat dissipation	
Animal	Dog	Cat	Duck	Bird	Donkey	Human	
Collision	Crash	Door	Diff Mats	Hit	Falling	Footsteps	Shock
Explosion	Bomb	Gunfire	Grenade	Thunder	Fireworks	Gas blasting	

The main reason why these four types of sounds are selected for convolutional neural training is that they have a high sound intensity and high sound recognition. The running sound of the machine has a continuous and stable running frequency, so it is relatively smooth in the spectrogram; animal calls have obvious characteristic peaks, and can detect obvious pitch; the loudness of the collision sound is high, and the two collision objects can be clearly distinguished Its own material characteristics, such as the collision sound of wooden doors and iron doors when they are closed are obviously different; the explosion sound has the characteristics of a large amount of energy and strong echo vibration, and at the same time the attenuation speed is very fast, usually the first sound wave collected The signal strength is the highest, and it is clearly different from the crash sound. Using these characteristics, it can help the convolutional neural network to perform fast and effective supervised learning, and finally iterate a set of sound source recognition system with higher accuracy.

4.3 The Comparison of CNN with other recognition systems

The methods used to identify sound source information are roughly divided into linear estimation algorithms and deep learning algorithms. The representative of the linear estimation algorithm is linear predictive coding. This type of sound source analysis method is based on the estimation of the formant of the audio segment, and the residual audio is used to analyze the audio intensity and frequency after the influence of the formant is eliminated. The deep learning algorithm mainly uses the establishment of a neural network model to perform short-time Fourier analysis on the time domain and frequency domain of the sound and performs convolution analysis with other spectrograms in the database, and finally trains a suitable recognition model. In the actual operation of the system, linear predictive coding is very sensitive to errors, and small errors will cause the system's predictive filter to become very unstable. However, the deep learning algorithm based on the convolutional neural network has high stability of the system with the help of a large amount of training data. Even if there is a certain error in the system, the stability of the system recognition can be ensured through repeated training.

In addition to the convolutional neural network deep learning algorithm, the support vector machine (SVM) algorithm is also a commonly used method for using deep learning algorithms to identify sound source information. The research results are compared horizontally to test which method is more advantageous in achieving sound source recognition.

Chapter 5: System Design and Implementation

5.1The Architecture Design of Localization and Recognition System

The purpose of this system is to locate the unknown sound source in the unknown coordinate position in the three-dimensional space and distinguish the sound source type through the sound wave characteristics. In order to realize this idea, this experiment is realized through three parts, namely sound signal preprocessing, sound source localization, and sound source feature recognition. Audio signal preprocessing refers to the processing of noise reduction and echo removal on the multivariate signals collected from the sensor array, to ensure that each channel contains as many sound source information signals as possible. Sound source localization is a sound wave traceability system based on the TDOA and related localization algorithms described in Chapter 3, which can effectively estimate the coordinates of the sound source in space. The sound source recognition part is based on the CNN deep learning network described in Chapter 4 to recognize audio segments and realize high-precision estimation of sound source features.

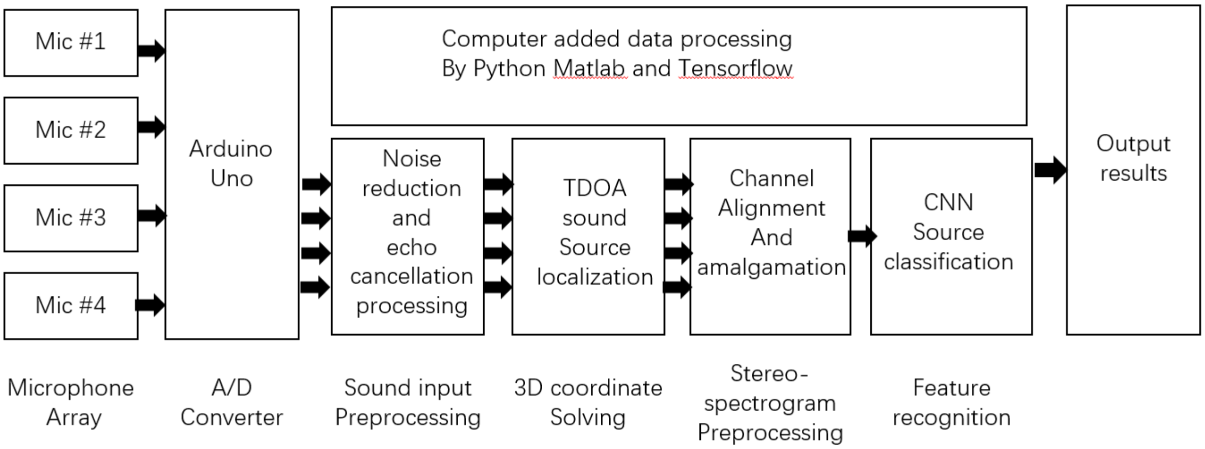

Mic #1
Mic #2
Mic #3
Mic #4
Arduino Uno
Computer added data processing
By Python Matlab and Tensorflow
Noise reduction and echo cancellation processing
TDOA sound Source localization
Channel Alignment And amalgamation
CNN Source classification
Output results
Microphone Array
A/D Converter
Sound input Preprocessing
3D coordinate Solving
Stereo-spectrogram Preprocessing
Feature recognition

Figure 7: System architecture for sound source identification and localization

The system composition and data flow of sound source identification and location are shown in Figure 7. From the figure, we can see that when a certain point in the space emits sound waves, the multi-channel stereo sound is first obtained through different microphone sensors, and the four sets of analog signals in the sensors are converted into corresponding digital signals using Arduino Uno and transmit to the computer by serial communication. After receiving these four sets of channel information, the system first identifies the origin of the sound source by the trigger equation. Then uses the delay between each set of channels to perform pairwise cross-correlation operations to obtain the TDOA value between the channels and then substituted into the three-dimensional hyperbolic equations (shown as equation(7)) to calculate the position coordinate of the sound source and obtain the three-dimensional coordinate value.

- $\sqrt{(x-x_1)^2+(y-y_1)^2+(z-z_1)^2}-\sqrt{(x-x_2)^2+(y-y_2)^2+(z-z_2)^2} = TDOA(A_{1st}-A_{2nd})$

- $\sqrt{(x-x_1)^2+(y-y_1)^2+(z-z_1)^2}-\sqrt{(x-x_3)^2+(y-y_3)^2+(z-z_3)^2} = TDOA(A_{1st}-A_{3rd})$

- $\sqrt{(x-x_2)^2+(y-y_2)^2+(z-z_2)^2}-\sqrt{(x-x_3)^2+(y-y_3)^2+(z-z_3)^2}=$ $TDOA(A_{2nd}-A_{3rd})$ (7)

After the system obtains the TDOA of the sound source and the sound source coordinates, the four sets of channels are processed through phase compensation, and then the amplitude is superimposed to obtain a set of single channel signals with homologous signals. The spectrogram generated after the soundtrack fusion generates the classification result of the sound source information by matching the features of the CNN network to the trained deep learning network, and finally outputs it with the position information. In this way, the position information and characteristic information of the sound source can be obtained at the same time.

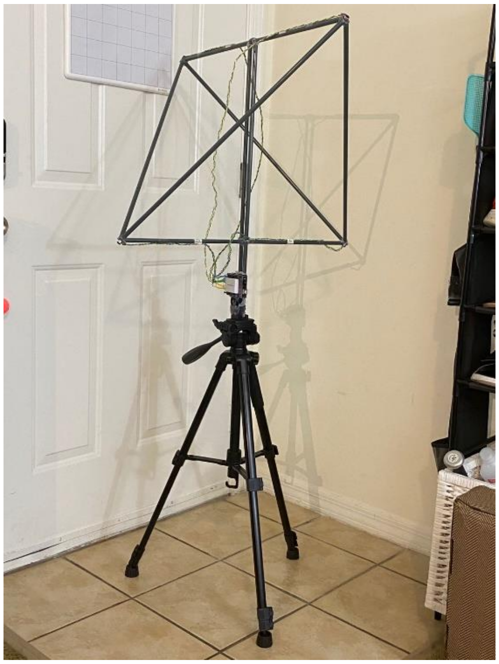


Figure 8: Implementation of 3D microphone sensor array

The core of obtaining sound source position information lies in the difference in the time of reaching different measurement points when the sound source emits sound waves, hence, a need for an array of microphone sensors. In order to capture this difference and trace back to the source, the related design of the sensor arrangement was described in Chapter 3. The specific implementation is as shown in Figure 8 using iron brackets to build a regular tetrahedron with every side length of 600mm and is erected at an appropriate height set as 3D zero point. In terms of sensor selection, in order to adapt to the Arduino single-chip processing system and have higher acoustic sensitivity, the MAX4466 Electret Microphone Amplifier Module produced by HiLetgo [17] was selected for this experiment. MAX4466 is a variable gain audio amplifier module that can amplify the input audio signal by 25 to 125 times. At the same time, the built-in clipping mechanism can automatically correct the voltage level, and the effective detection bandwidth is 20-20 kHz.

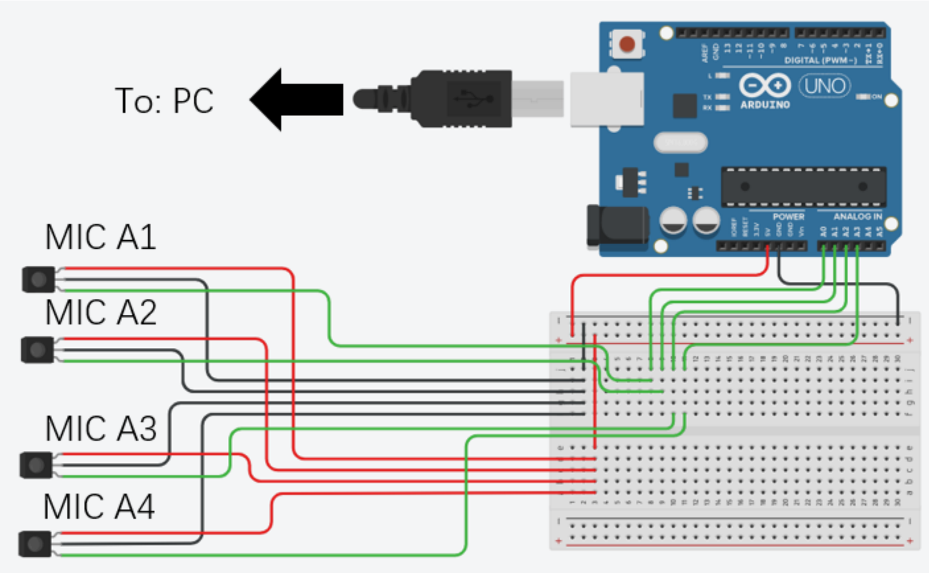
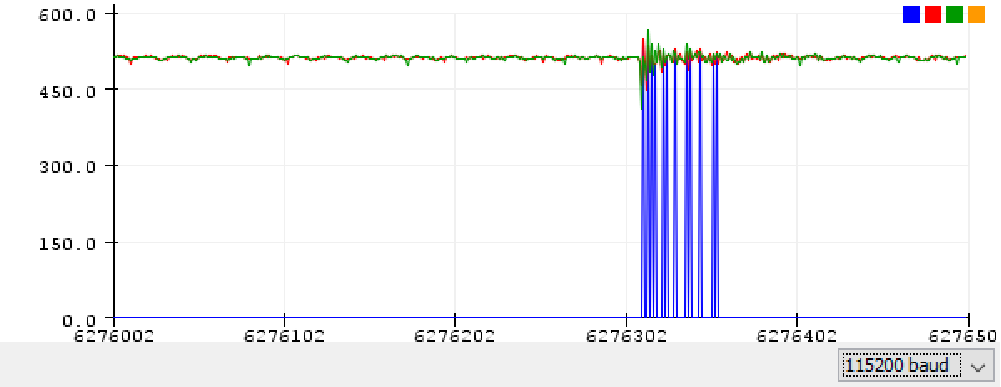


Figure 9: Hardware connection mode and A/D conversion of sensor array

The Analog to Digital Converter is essential. By looking up the Arduino datasheet, we know that Arduino supports 5v/1024 analog-to-digital conversion, so the static horizontal voltage is 2.5v/512. We can set the horizontal voltage zero point of the audio signal to 512 based on this information so that we are processing audio data when you can get more positive and negative amplitude space. The connection method of the microphone array and the Arduino A/D channel is shown in Figure 9.



115200 baud

Figure 10: Use trigger mechanism to identify sound source segments

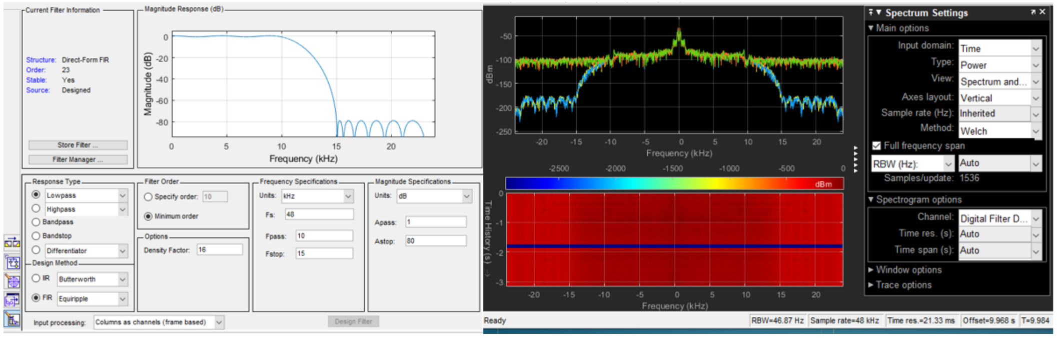
Since the time when the sound source utters is unknown, in most cases the signal collected by the sensor is environmental noise which necessitates the application of sound input processing. To improve the effectiveness of serial communication, the sound source trigger mechanism is used in the process of collecting signals to help the sensor intercept useful audio fragment. The specific trigger method (shown in Figure 10) is as follows: first sample the environmental noise in the initial state of power-on. In order to obtain the sound source characteristic information as complete as possible, the system bit rate is set to 2 Mbps, the noise sampling time is set to 1 millisecond, and a total of 200 samples. Calculate the average value and peak value of the sampled data to obtain the amplitude range of the noise. After the comparison test, when the peak value of the input terminal is greater than 1.5 times the absolute value of the maximum noise amplitude, it can be regarded as the new input signal as the sound source signal rather than the noise signal, which is used as the initial state for intercepting the sound source characteristics, and this function also can judge the end position of the sound source signal.

When the amplitude of the input signal reaches the trigger condition, it starts serial communication to the PC and sends the intercepted audio samples to the PC buffer in turn until the signal amplitude does not meet the trigger condition. It is worth noting that, in order to match the accuracy of the subsequent audio recognition, the upper limit of the audio single sampling time is two seconds to ensure that the test segment and the training reference length in the convolutional neural network are equal.

5.2 Pre-processing before Stereo Signal Localization

When we successfully intercepted the audio fragments of a sound source, because the audio is still mixed with environmental background noise and sound source echo reflected by obstacles, it is not possible to directly calculate TDOA. To eliminate the negative effects caused by these noise audios, the first thing we must do is to use digital filters to filter out environmental background noise.

As shown in Figure 11, through the Fourier transform of the audio segments, we can clearly see that there are a lot of high-frequency noises in addition to the sound source signal in the audio file. In order to reduce this part of the high-frequency noise, A FIR low-pass filter with order of 23 and cutoff frequency of 15 kHz is designed to filter out high-frequency noise outside



Structure: Direct-Form FIR
Order: 23
Stable: Yes
Source: Designed

● Lowpass
○ Highpass
○ Bandpass
○ Bandstop
○ Differentiator

○ IIR Butterworth
● FIR Equiripple

○ Specify order: 10
● Minimum order

Density Factor: 16

Units: kHz
Fs: 48
Fpass: 10
Fstop: 15

Units: dB
Apass: 1
Astop: 80

Input processing: Columns as channels (frame based)

Input domain: Time
Type: Power
View: Spectrum and...
Axes layout: Vertical
Sample rate (Hz): Inherited
Method: Welch
☑ Full frequency span
RBW (Hz): Auto
Samples/update: 1536

Channel: Digital Filter D...
Time res. (s): Auto
Time span (s): Auto

RBW=46.87 Hz | Sample rate=48 kHz | Time res.=21.33 ms | Offset=9.968 s | T=9.984

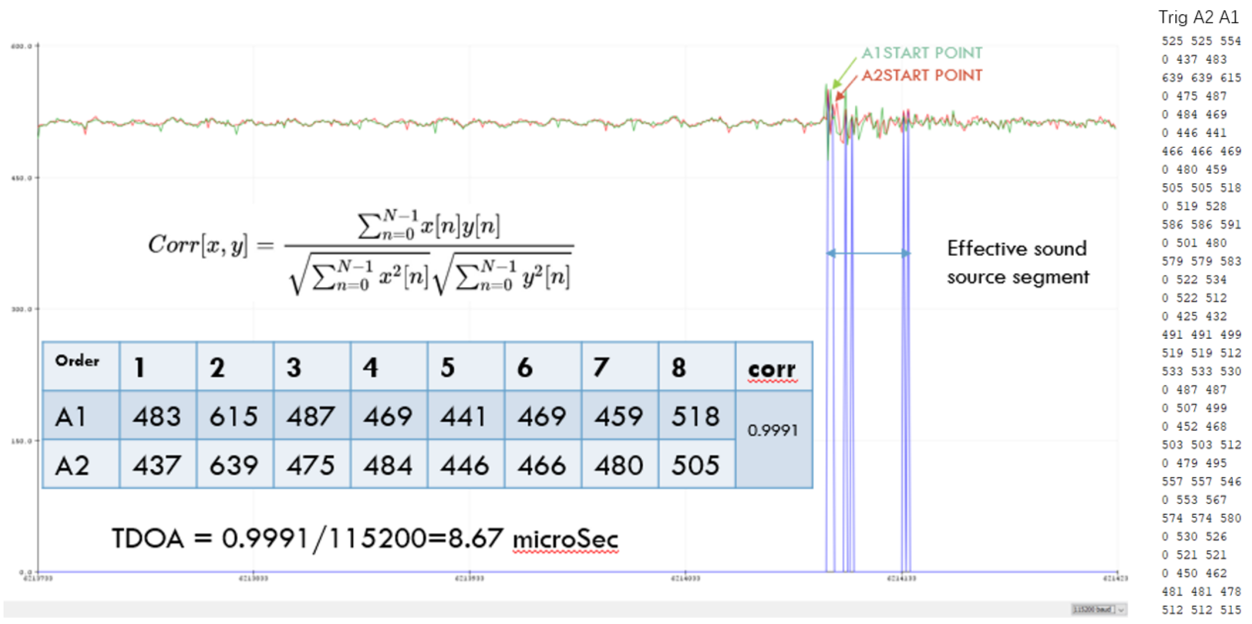
Figure 11: Noise cancellation by Low-pass FIR filter

of 15 kHz. In comparison, the frequency domain response of audio segments has been significantly improved, and no obvious sound source distortion is found after playing the filtered audio.

After the background noise is reduced, the next step is to process the echo signal in the audio. Because the echo signal has a high similarity with the sound source signal, it cannot be eliminated by directly passing through the filter, and if the estimation delay algorithm is used to measure each group of signals, the system calculation will be increased and the recognition efficiency of the system will be reduced. Therefore, the method used in the process of removing echo signals is to calculate the signal correlation. We know that the shortest distance between the two sensors is 600 mm, and the echo signal needs to travel a longer path, so by using the method introduced in Chapter 3, we can find the sound source signal with a delay of more than 1.73 milliseconds, and the signal that exceeds the limited time that this signal is the echo signal that is not transmitted directly but is reflected back.

5.3 Implementation of Sound Source Localization Subsystem

After the acoustic signal is preprocessed, the TDOA data extraction process can be carried out. Through the previous preprocessing, the trigger mechanism has identified the first sound source to reach the sensor, and this first sensor that receives the signal is taken as the zero point of the time delay of the arrival time difference. Because four sensors are used in this experiment, C (4,2) = 6 sets of TDOA pairs will be generated. The calculation sequence of the TDOA pair is combined according to the order in which the sound source signal arrives at the sensor, which are A1st-A2nd, A1st-A3rd, A1st-A4nd, A2nd-A3rd, A2nd-A4nd, and A3rd-A4nd.



$$Corr[x,y] = \frac{\sum_{n=0}^{N-1} x[n]y[n]}{\sqrt{\sum_{n=0}^{N-1} x^2[n]}\sqrt{\sum_{n=0}^{N-1} y^2[n]}}$$

Order	1	2	3	4	5	6	7	8	corr
A1	483	615	487	469	441	469	459	518	0.9991
A2	437	639	475	484	446	466	480	505	

TDOA = 0.9991/115200=8.67 microSec

```
Trig A2 A1
525 525 554
0 437 483
639 639 615
0 475 487
0 484 469
0 446 441
466 466 469
0 480 459
505 505 518
0 519 528
586 586 591
0 501 480
579 579 583
0 522 534
0 522 512
0 425 432
491 491 499
519 519 512
533 533 530
0 487 487
0 507 499
0 452 468
503 503 512
0 479 495
557 557 546
0 553 567
574 574 580
0 530 526
0 521 521
0 450 462
481 481 478
512 512 515
```

Figure 12: TDOA calculation demonstration when the bit rate is 115200 bps

As shown in Figure 12, the estimated value of TDOA can be obtained by performing cross-correlation calculations on the data between the two channels. After obtaining the estimated value, substituting it into Equation 7 can get the corresponding hyperbolic equations. Since this equation uses three-dimensional coordinates for calculation, it is difficult to directly solve the equations. Therefore, we can iterate the general solution between each TDOA pair at a time while temporarily fixing the z-axis coordinate, and eventually find the common solution of all hyperbolas. The coordinates of this solution are the three-dimensional coordinates of the sound source estimated by the system.

5.4 Implementation of Sound Source Recognition Subsystem

As shown in Figure 13, the main method used in this experiment to analyze sound source recognition and classification is a deep learning network architecture based on supervised learning. The features embodied by the sound source information include time-domain features (amplitude) and frequency-domain features. The spectrogram can fully express the time domain and frequency domain information of audio at the same time, so it is very suitable for sound source information identification. We have discussed in Chapter 4 how to preprocess the audio frequency spectrogram to facilitate the operation of the convolutional neural network. This part is mainly used to explain how to implement the convolutional neural network. The construction of deep learning network models usually requires a lot of data for training, but because the sound source information designed by this research can be directly found in the database is very limited, it is impossible to directly train the sound source characteristics through the existing database. Hence, we need to build the training database ourselves. In order to ensure that the deep learning model (Based on the convolutional neuron deep learning network mentioned in Chapter 4) can have enough training data, we plan to build a crawler to obtain audio clips from royalty-free audio sources including SoundBible [18], Freesound [19], GameSound [20], ZapSplat [21],

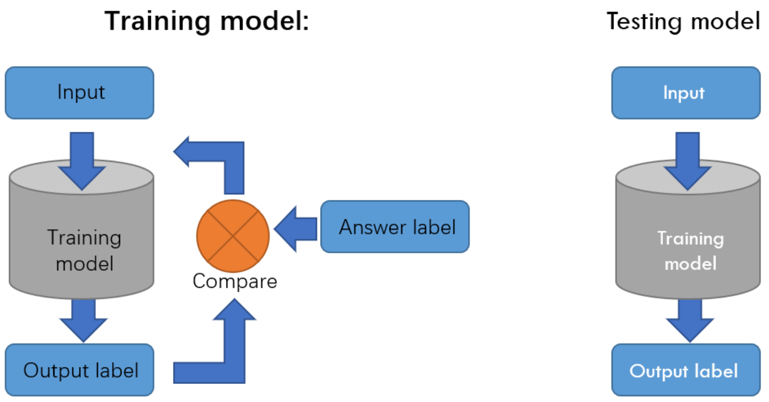


Figure 13: Deep learning network architecture based on supervised learning

AudioMicro [22] and SoundGator [23]. From the information contained in Table 1, we have selected 50 sets of audio tags with obvious characteristic differences among the four major sound source types. These tags are used as keywords to filter out the corresponding audio files in the audio website and import them one by one to generate the corresponding spectra in the system.

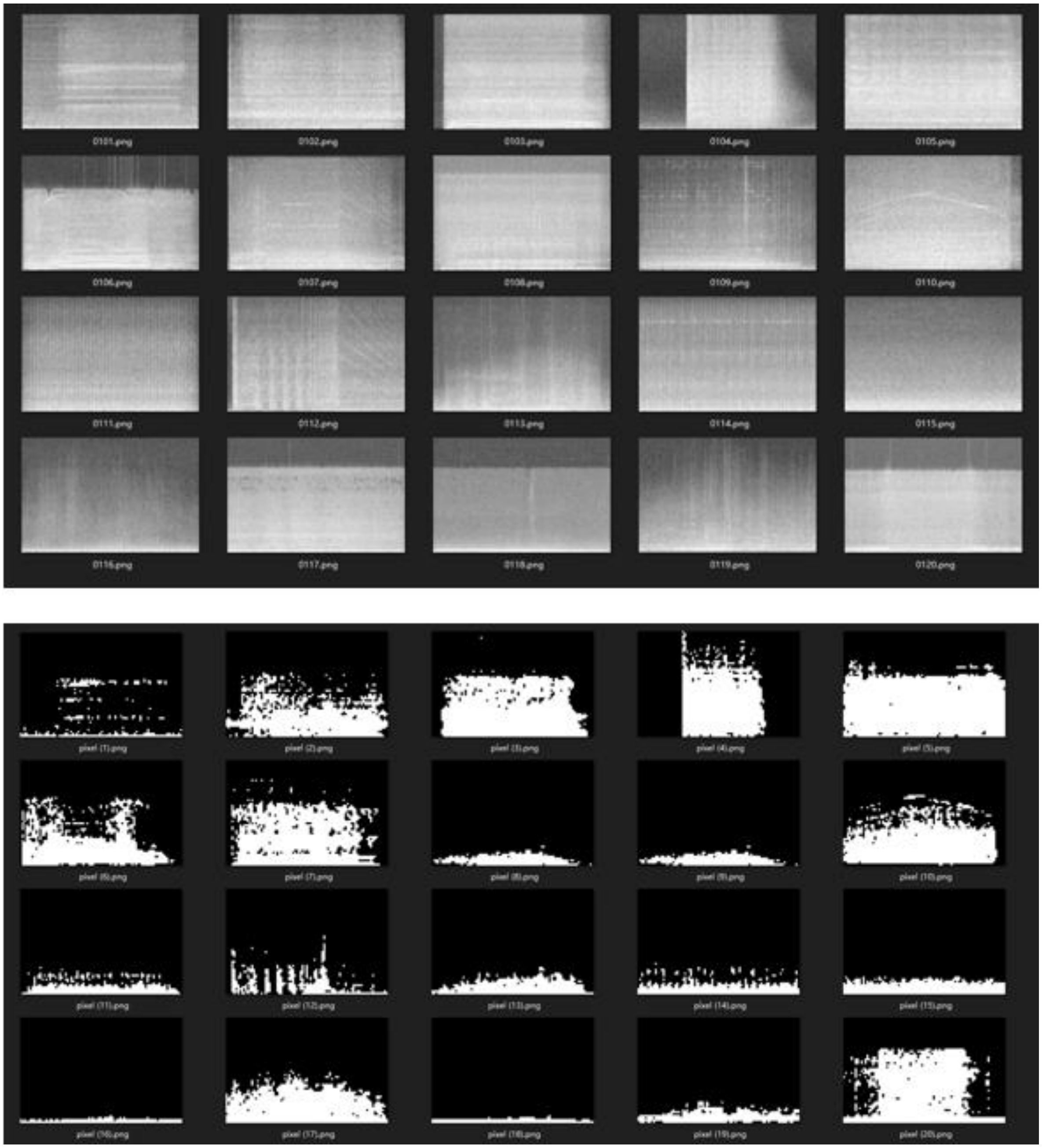


Figure 14: Pixelated spectrogram for training model

In order to realize the recognition of sound source information through the convolutional neural network, the first task is to build a deep learning model. To ensure the effectiveness of learning, this research mainly uses supervised learning to build deep learning models (Figure 13). Supervised learning is mainly composed of two parts, one part is a training model built for the system to learn by itself, and the other part is a test model used to test the effectiveness of the model. These two models are combined to form a complete deep learning network.

Figure 14 shows the spectrogram after pixilation. The pixelized image only needs to consider the convolution calculation of two values of 0 (black) and 1 (white) when processing, which greatly simplifies the amount of calculation of the convolutional neuron network. Such a processing method can speed up the comparison of images and shorten the processing time while retaining the main acoustic time-frequency image characteristics. The current input picture displayed in the system is 350X225. First, all images are normalized to generate a 200x200 graphic matrix. After generating a new graph, we design the convolutional layer, and the activation type is “Relu”.

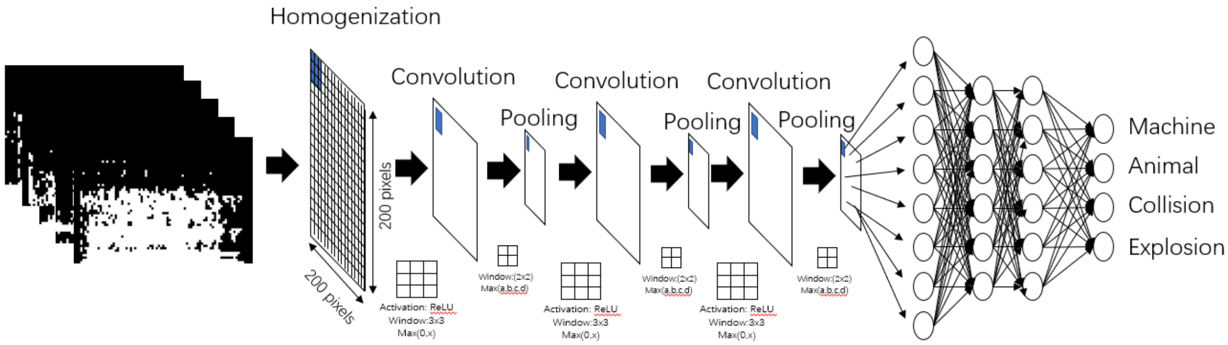
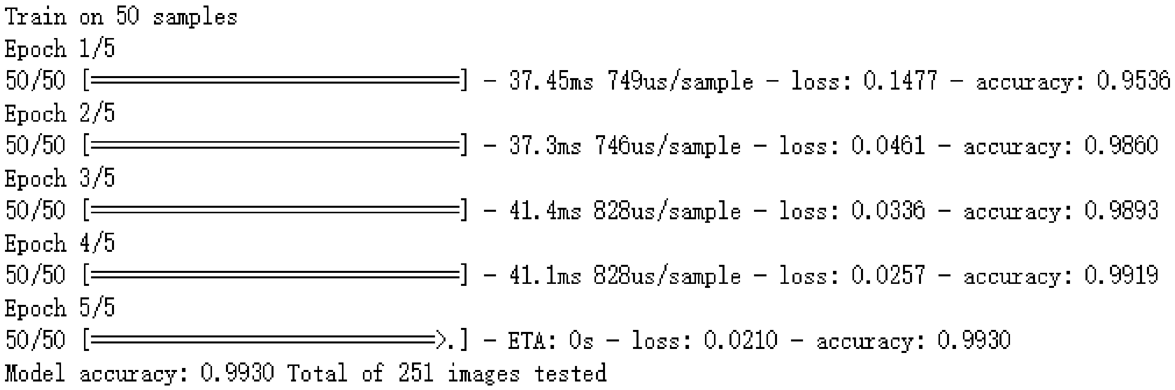


Figure 15: CNN convolution system design

This system uses a three-layer convolution structure in the design of the convolutional neuron network as shown in Figure 15. A 3x3 convolution window is used in the first layer of the convolutional network to generate 16 different convolutional layers. The convolutional layer

After 2x2 window pooling, the maximum value is obtained, and the convolution calculation of the next level is performed. In the second convolution, 32 convolutional layers are constructed using the same window, and expanded to 64 layers in the third convolution. After completing three cycles of convolution and pooling, the data is transmitted to the neural connection network. The neurons in the first hidden layer are 500, the neurons in the second hidden layer are 500, and the final output layer is four classification outputs. After completing the architecture configuration of the convolutional neural network, the training file is loaded into the model through the python program, and the system automatically compares the generated results with the label. Finally, the training of the convolutional neural network is realized.

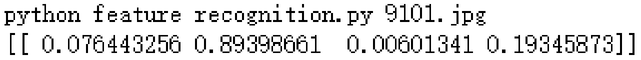


```
Train on 50 samples
Epoch 1/5
50/50 [==============================] - 37.45ms 749us/sample - loss: 0.1477 - accuracy: 0.9536
Epoch 2/5
50/50 [==============================] - 37.3ms 746us/sample - loss: 0.0461 - accuracy: 0.9860
Epoch 3/5
50/50 [==============================] - 41.4ms 828us/sample - loss: 0.0336 - accuracy: 0.9893
Epoch 4/5
50/50 [==============================] - 41.1ms 828us/sample - loss: 0.0257 - accuracy: 0.9919
Epoch 5/5
50/50 [============================>.] - ETA: 0s - loss: 0.0210 - accuracy: 0.9930
Model accuracy: 0.9930 Total of 251 images tested
```

Figure 16: CNN sound source recognition model training results

Complete the model training of the convolutional neural network, and the results are shown in Figure 16. In this experiment, each group of the same type of sound source has 7 to 8 groups of different sound source training to ensure the diversity of training materials.

After five rounds of model training, the accuracy of the final convolutional neural network can reach 99.3%.From the experimental results, we can see that the automatic classification of sound sources through the convolutional neuron network system can obtain higher accuracy after multiple rounds of training. The processing capacity of these data can also become more accurate as the amount of training increases.



```
python feature recognition.py 9101.jpg
[[ 0.076443256 0.89398661  0.00601341 0.19345873]]
```

Figure 17: Model testing result by dog audio

After the model verification was completed, an audio file of a dog barking was selected from the database and input into the system to test the model's resolution ability as shown in Figure 17. The result of the data shows that the audio coincides with the animal in the system at 89%, the mechanical sound at 8%, the collision sound is close to 0%, and the explosion sound at 19%. It can be determined that the dog's barking belongs to the model of animal sounds. The results show that combining the trained model can effectively identify and classify sound source features. The sound source recognition system can effectively use the trained model for sound source recognition after three-dimensional positioning, so it has reliable applications in system responsivity and real-time sound source processing.

5.5 Results comparison with other researchers' SVM research

The support vector machine (SVM) is also usually used for sound recognition and classification. During the experiment at this time, it was found that Rong [24] had done a similar sound source recognition project in the speech recognition system. Rong uses SVM as the main method of audio recognition in his research and conducts deep learning training by collecting sound source information in general scenes. Therefore, when comparing the experimental results of this experiment with Fong's SVM project, we can discover the differences in the results of two different deep learning methods in voice recognition.

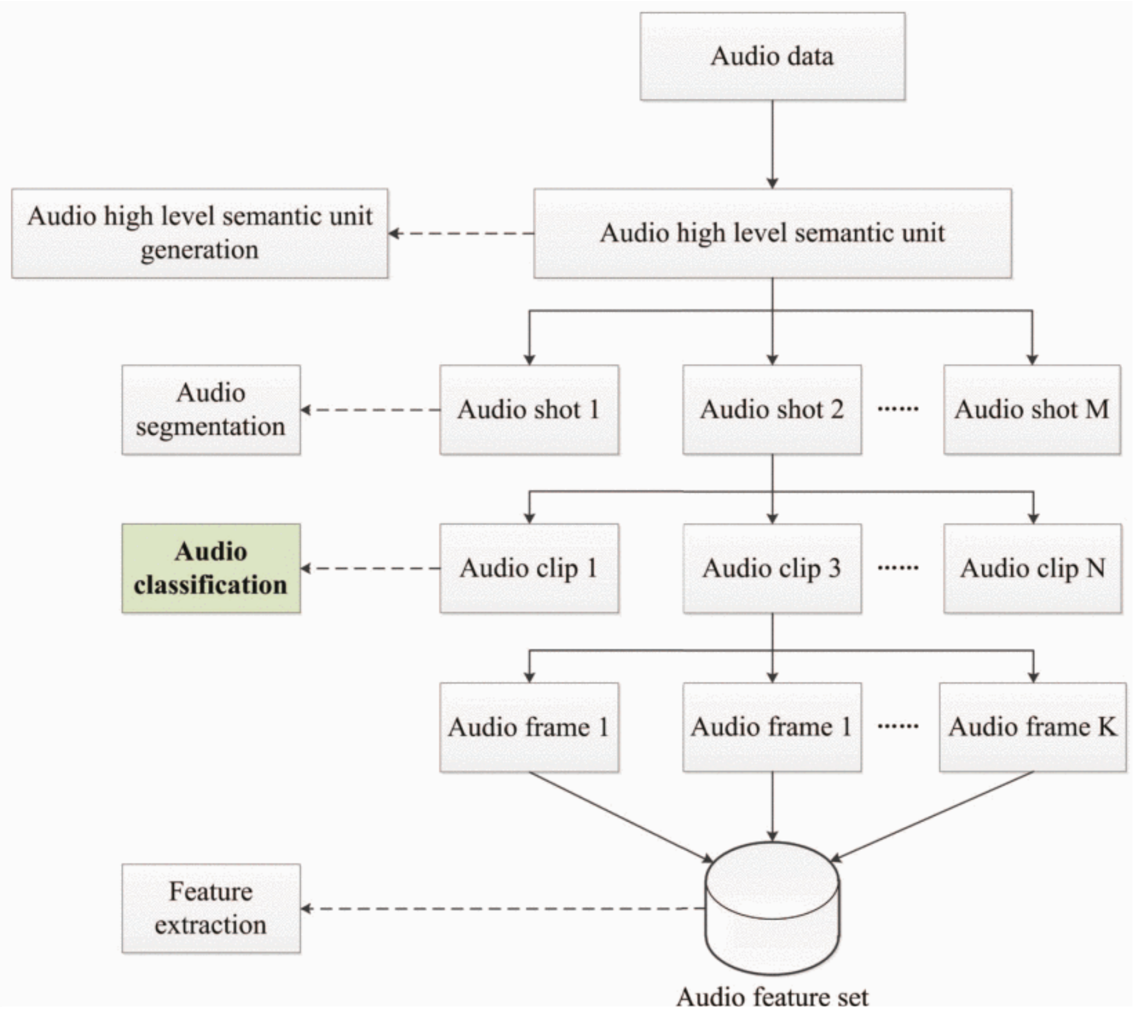


Figure 18: SVM classification system data structure

Figure 18 shows the data flow architecture of the system that his team uses SVM to recognize sounds. When he collects sound clips, he first intercepts multiple segments of the sound, and then compares the features of independent audio segments through the SVM system, and finally generate estimates of sound information. This system uses a four-layer audio filtering system to gradually extract feature segments in audio through audio frame-by-frame filtering, effective sound interception, audio structure grouping, and advanced audio extraction. This method has a similar effect to the feature audio trigger interception mechanism that we use when the sound source signal is preprocessed. Therefore, the comparison of these two sets of experimental results has a certain reference value for testing effectivity between SVM and CNN in the deep learning algorithms.

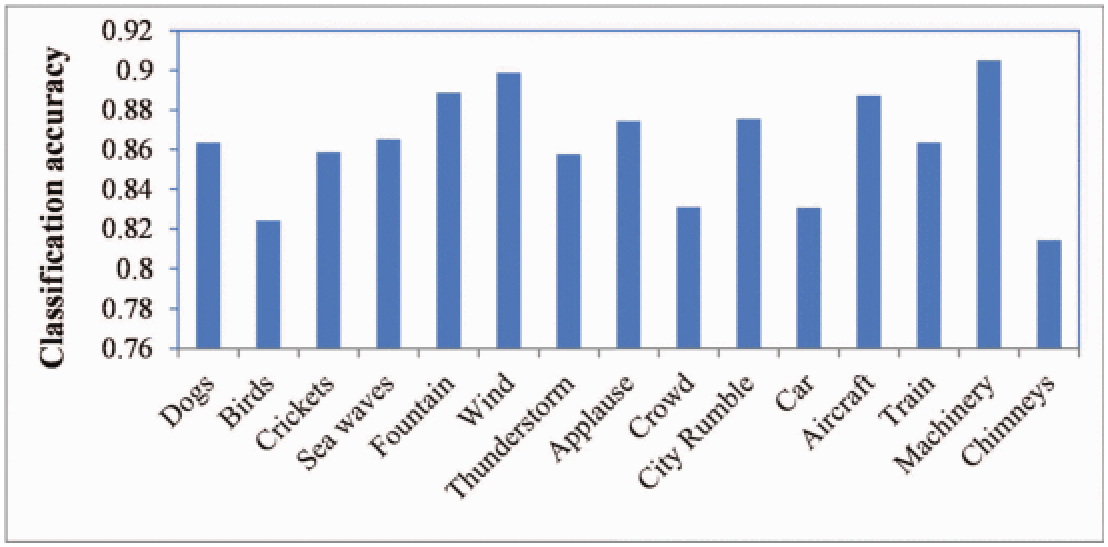


Figure 19: SVM classification result by using general sounds

The result of sound classification using SVM is shown in Figure 19. In the same figure, we can see that this experiment uses 15 different sound types for comparison. We use the accuracy of "dog barking" and our own experiments. The comparison results are shown in Table 2.

Table 2: Classification result comparison between CNN and SVM

Type	CNN Deep Learning	SVM Deep Learning
Model Accuracy	0.993	0.876
Sample “Dog” Prediction	0.893	0.860

Comparing the prediction results between the two models, it can be found that the CNN deep learning algorithm can obtain higher classification accuracy than SVM after a large amount of experimental data training, and has higher accuracy in predicting specific sound sources.

Chapter 6: Conclusion and Future Work

6.1 Conclusion

The originator of the sound has its own unique characteristics, and these characteristics are often transmitted to the remote receiver along with the sound wave in the process of making the sound. Establishing a sound source perception system that integrates sound source location information and sound source characteristic information can help researchers understand the characteristic information of the sound source itself and understand the sound source location. The three-dimensional position location based on the TDOA method can quickly and effectively find the three-dimensional coordinates of the sound source, and it is a highly reliable passive location algorithm. At the same time, the CNN deep learning algorithm used this time can well summarize the similarities between audios and classify them by extracting audio features in the spectrogram. Combining these two technologies can quickly and effectively understand the specific properties of the sound source, and help people better understand the specific location and state of the sound source when the sound is emitted.

The experimental results show that the construction of the sound source processing system in this research can effectively realize the simultaneous recognition of sound source characteristics and the sound source localization in three-dimensional space. The accuracy of sound source recognition depends on the number of training sessions and the amount of data involved in training. Therefore, this system will have higher accuracy through long-term training in the future. The system's sound source location has been optimized for traditional TDOA, so that it can quickly locate sound sources in three-dimensional space and lay a solid foundation for the ability to track and identify sound source information in real time in the future.

6.2 Future Recommendation

The localization and recognition of the sound source can help the artificial intelligence system better understand the surrounding environment, especially the need to understand the sound source information through obstacles. For example, when the autonomous driving system is on a congested road section, the sight of many vehicles is often blocked by other surrounding vehicles or obstacles. Currently, traditional optical sensors, optical radar, or ultrasonic sensors cannot directly penetrate the obstacles to understand other moving objects. Therefore, the auxiliary detection method based on sound source localization and recognition technology can detect and track the sound source information outside the visual range, and can predict the movement of the sound source according to the trajectory comparison, and improve the safety of unmanned driving. At the same time, sound source location and recognition can help the hearing-impaired people better avoid potential dangers and better perceive the world around them through the sound source visualization system. This technology has many very practical technologies waiting to be further developed.

This experiment was completed in an indoor environment due to the epidemic, so complete experimental data was not obtained in terms of data collection, dynamic sound source capture, and noise resistance in outdoor environments. In future studies, more improvements will be made to the above shortcomings. The optimized system will be able to realize real-time capture of dynamic sound sources in the future, or real-time processing and identification of multiple environmental sound sources in complex conditions such as mobile sensor array platforms, and further optimize the overall anti-noise ability of the system.

6.3 Summary

In conclusion, the sound source recognition and location system is an audio processing technology with rich application prospects and practicability, which enables the computer to independently identify the type, characteristics and location of the sound source. This system is mainly composed of a stereo microphone array, a digital signal processing system, a TDOA traceability positioning system and a convolutional neuron network deep learning system. The audio file processed by the system can parse out the type of object, the vocal feature, and the location of the sound source. This system and its derived sound source behavior prediction system can be equipped with artificial intelligence driving systems or other environmental detection systems in the future, and it has very broad prospects in many automation fields in the future.

References

[1]X. Yue, G. Qu, B. Liu and A. Liu, "Detection Sound Source Direction in 3D Space Using Convolutional Neural Networks," in 2018 *First International Conference on Artificial Intelligence for Industries* (AI4I), 2018, pp. 81-84: IEEE

[2] K. Martin, (1999, January 01). Sound-source recognition : A theory and computational model. Retrieved August 01, 2020, from https://dspace.mit.edu/handle/1721.1/9468

[3] M. Boes, D. Oldoni, B. De Coensel and D. Botteldooren, "A biologically inspired recurrent neural network for sound source recognition incorporating auditory attention," The 2013 International Joint Conference on Neural Networks (IJCNN), Dallas, TX, 2013, pp. 1-8

[4] T. Heittola, A. Mesaros, T. Virtanen and M. Gabbouj, "Supervised model training for overlapping sound events based on unsupervised source separation," 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, Vancouver, BC, 2013, pp. 8677-8681

[5] K. Yamamoto, F. Asano, W. F. G. van Rooijen, E. Y. L. Ling, T. Yamada and N. Kitawaki, "Estimation of the number of sound sources using support vector machines and its application to sound source separation," 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03)., Hong Kong, 2003, pp. V-485

[6] K. Nakadai, H. Nakajima, G. Ince and Y. Hasegawa, "Sound source separation and automatic speech recognition for moving sources," 2010 IEEE/RSJ International Conference on Intelligent Robots and Systems, Taipei, 2010, pp. 976-981

[7] T. Higashimoto and H. Sawada, "A mechanical voice system and its adaptive learning for the mimicry of human vocalization," Proceedings 2003 IEEE International Symposium on Computational Intelligence in Robotics and Automation. Computational Intelligence in Robotics and Automation for the New Millennium (Cat. No.03EX694), Kobe, Japan, 2003, pp. 1040-1045 vol.2

[8] P. J. Clemins, M. B. Trawicki, K. Adi, Jidong Tao and M. T. Johnson, "Generalized Perceptual Features for Vocalization Analysis Across Multiple Species," 2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings, Toulouse, 2006, pp. I-I

[9] C. Nadeu, R. Chakraborty and M. Wolf, "Model-based processing for acoustic scene analysis," 2014 22nd European Signal Processing Conference (EUSIPCO), Lisbon, 2014, pp. 2370-23

[10] E. De Sena, Z. Cvetković, H. Hacıhabiboğlu, M. Moonen and T. van Waterschoot, "Localization Uncertainty in Time-Amplitude Stereophonic Reproduction," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 28, pp. 1000-1015, 2020

[11] The Speed of Sound. (n.d.). Retrieved August 13, 2020, from https://www.mathpages.com/home/kmath109/kmath109.htm

[12] E. Kim and D. Choi, "A 3D Ad Hoc Localization System Using Aerial Sensor Nodes," in IEEE Sensors Journal, vol. 15, no. 7, pp. 3716-3723, July 2015

[13] I. Zafar, Hands-on convolutional neural networks with TensorFlow: Solve computer vision problems with modeling in TensorFlow and Python. Birmingham, UK: PACKT Publishing Limited, 2018

[14] G. Diğken and T. İbrikçi, "Recognition of non-speech sounds using Mel-frequency cepstrum coefficients and dynamic time warping method," 2015 23nd Signal Processing and Communications Applications Conference (SIU), Malatya, 2015, pp. 144-147

[15] B. Zhang, C. Quan and F. Ren, "Study on CNN in the recognition of emotion in audio and images," 2016 IEEE/ACIS 15th International Conference on Computer and Information Science (ICIS), Okayama, 2016, pp. 1-5

[16] Á. Incze, H. Jancsó, Z. Szilágyi, A. Farkas and C. Sulyok, "Bird Sound Recognition Using a Convolutional Neural Network," 2018 IEEE 16th International Symposium on Intelligent Systems and Informatics (SISY), Subotica, 2018, pp. 295-300

[17] A.HiLetgo, Electret Microphone Amplifier MAX4466 Module Adjustable Gain Blue Breakout Board for Arduino. Retrieved October 19, 2020, from http://www.hiletgo.com/ProductDetail/1952713.html ,2020

[18] M. Koenig, SoundBible.com. Retrieved October 21, 2020, from http://soundbible.com/

[19] Freesound. Retrieved October 21, 2020, from https://freesound.org/

[20] GameSounds.xyz - Royalty free or public domain game music and sounds. (n.d.). Retrieved October 21, 2020, from https://gamesounds.xyz/

[21] Download Free Sound Effects &amp; Royalty Free Music. (2020, October 20). Retrieved October 21, 2020, from https://www.zapsplat.com/

[22] Royalty Free Music and Sound Effects Library w/ 300k+ high quality tracks from GRAMMY winning artists and Hollywood's top sound studios. (n.d.). Retrieved October 21, 2020, from https://www.audiomicro.com/

[23] Free Sound Effects. (n.d.). Retrieved October 21, 2020, from http://www.soundgator.com/

[24] F. Rong, "Audio Classification Method Based on Machine Learning," 2016 International Conference on Intelligent Transportation, Big Data & Smart City (ICITBS), Changsha, 2016, pp. 81-84,