Research Article

Zhuhe Wang, Nan Li*, Tao Wu, Haoxuan Zhang, and Tao Feng

Simulation of Human Ear Recognition Sound Direction Based on Convolutional Neural Network

https://doi.org/10.1515/jisys-2019-0250
Received Nov 22, 2019; accepted Apr 11, 2020

Abstract: In recent years, more and more people are applying Convolutional Neural Networks to the study of sound signals. The main reason is the translational invariance of convolution in time and space. Thereby the diversity of the sound signal can be overcome. However, in terms of sound direction recognition, there are also problems such as a microphone matrix being too large, and feature selection. This paper proposes a sound direction recognition using a simulated human head with microphones at both ears. Theoretically, the two microphones cannot distinguish the front and rear directions. However, we use the original data of the two channels as the input of the convolutional neural network, and the resolution effect can reach more than 0.9. For comparison, we also chose the delay feature (GCC) for sound direction recognition. Finally, we also conducted experiments that used probability distributions to identify more directions.

Keywords: Convolutional Neural Network, simulated human head, dual-channel raw data, GCC, probability distributions

1 Introduction

In the field of artificial intelligence, computer hearing [8] remains in an early stage of research than computer vision [10]. With the development of technology, computer hearing has become an important research topic in the field of artificial intelligence. It is one of the essential symbols of intelligent robots, and it is also an important means to realize human-computer interaction and interaction with the environment [5].

There are three mainstream methods for sound source localization: time delay estimation method [9], the beam forming method [7] and a machine learning method [16]. The delay estimation is built on the generalized correlation function based on periodic cross-spectral density and the generalized cross-correlation algorithm based on the cross-power spectrum. This algorithm was suggested by Knapp *et al.* In order to get a more accurate time delay estimation, people focus on the design of the microphone array, such as linear arrays [6], circular arrays [15], distributed arrays [3], and non-coplanar arrays of any shape [1]. The sound source localization method based on time delay estimation is firstly to estimate the time difference of sound reaching different microphones. Then, the position of the sound source can be found by a geometric relationship. The advantage is that the amount of calculation is unimportant, but it will produce cumulative errors and cannot be used for multi-source positioning. The second method is the beam forming method. It is primarily divided into the subspace method and beam scanning method [4, 11, 17, 20]. The subspace method mainly performs feature decomposition on the covariance matrix of the output data of an array and obtains a noise subspace orthogonal to the signal subspace corresponding to the signal component. Sound local-

***Corresponding Author: Nan Li:** School of Artificial Intelligence, Beijing Technology and Business University, Beijing, China; Email: linan@th.btbu.edu.cn
Zhuhe Wang, Tao Wu, Haoxuan Zhang, Tao Feng: School of Artificial Intelligence, Beijing Technology and Business University, Beijing, China

ization is performed using orthogonally between the signal and noise subspaces. Classical methods include multi-signal classification [19] and estimation of signal parameters by rotational invariance technique [21]. The beam scanning method can position the array signal in a specific direction. The classical method is the steering response power phase transformation. The benefit of this method is that its spatial resolution is not affected by the sampling frequency, and arbitrary precision can be achieved under certain conditions. However, if directional noise occurs in practical applications, and its energy is not too dissimilar from the sound source, it is judged that the sound source may use the noise source as the sound source according to the largest features in the correlation matrix. Besides, this method needs to search the entire space to determine the reliable source, and the accuracy of the estimation is related to the degree of subdivision of the space, and the calculation is complicated.

Machine learning is a popular method in recent years, and they mainly include the two parts of feature extraction and machine learning methods. Extracted features mainly include time delay estimation feature, covariance matrix, and short-time power spectral density function spectrum. Machine learning methods mainly include support vector machines, multi-layer perceptron, Gaussian mixture models, Convolutional Neural Network and so on. In [14] the covariance matrix of the received signal of the microphone array is mainly used as the feature input. The feed-forward neural network model was chosen for sound source localization. In [13], the delay feature (GCC) is extracted by the generalized cross-correlation method, and the support vector machine is employed for positioning. In [18], the delay feature is further extracted, The network selects convolutional neural networks and multilayer perceptrons for reliable source coordinate localization. However, these methods use a complex microphone matrix and do not achieve end-to-end sound direction recognition.

2 Overview of the principle of acoustic scattering

It is well known that in terms of sound source localization, the number of microphones required is three or more. This article uses a dual microphone with input characteristics including time delay (GCC) and raw data. To some extent, when the sound source distance is the same, the raw data and time delay of the sound source into the sensor in the front and back directions are basically the same.

The basic principle of this paper is the sound scattering feature. As shown in Figure 1. When the sound wave emitted by the sounder reaches the simulated human head, when an obstacle such as the auricle and the irregular surface of the human face is encountered, part of the sound wave is deviated from the original

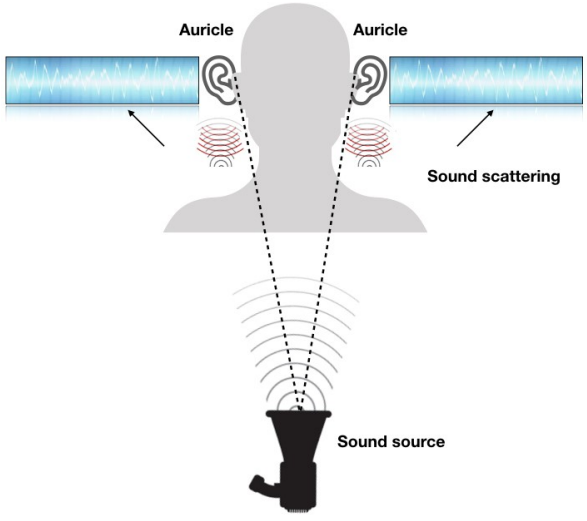

Figure 1: Schematic diagram of acoustic scattering

predetermined propagation path and spreads from the periphery of the obstacle. This makes the difference in the audio data received by the microphone. With this feature we try to use dual microphones for sound source orientation.

3 The proposed approach

In this paper, our model includes two parts of feature extraction and convolutional neural network. In terms of feature extraction, we use the dual-channel raw data and time delay estimates as feature input for CNN, respectively. On CNN, we propose two kinds of network structures Net1 and Net2. The specific operation process is shown in Figure 2.

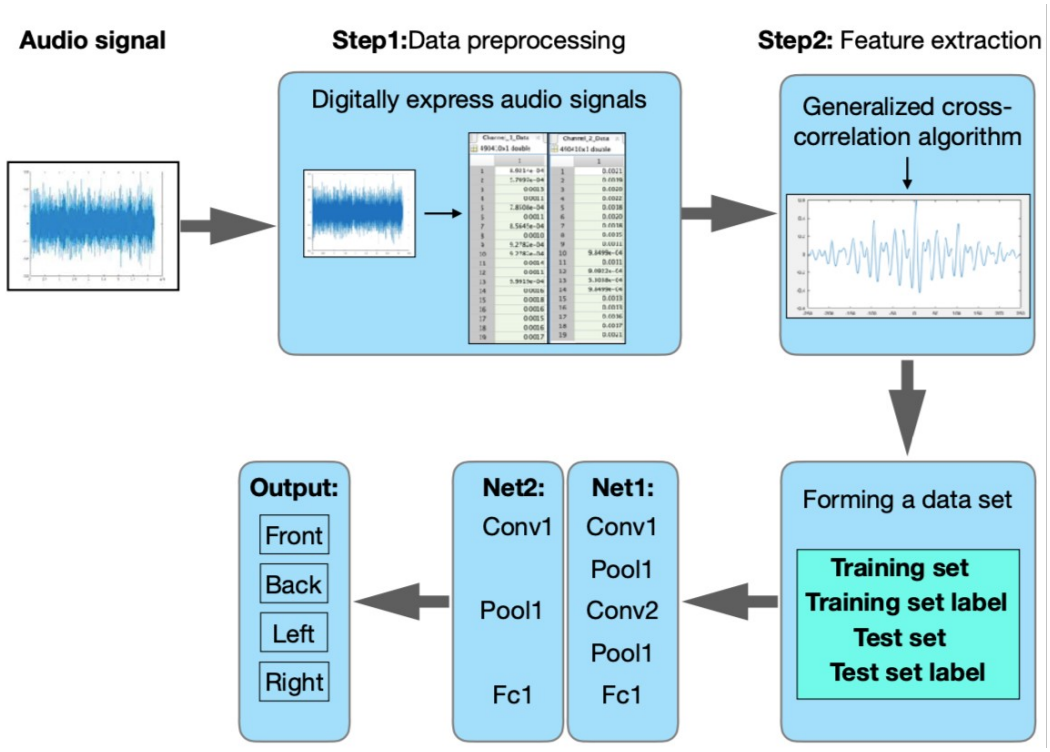

Figure 2: Sound direction recognition model flow chart

3.1 Data acquisition and preprocessing

In this section, we will describe step one following Figure 3. The collection of audio signals is selected in an anechoic room environment and an ordinary room environment. The sound is emitted in four different directions of the simulated human head, and these auditory data are collected by Pulse.

For the original dataset, we intercept directly from the raw data in each direction, and the intercept length is L. All data form a dataset in binary form. The dataset mainly includes four parts, namely, a training set, a training set label, a test set, and a test set label.

For the GCC dataset, we need to go through the process of feature extraction. The specific operation is: extracting the delay feature by using the generalized cross-correlation algorithm based on the cross-power

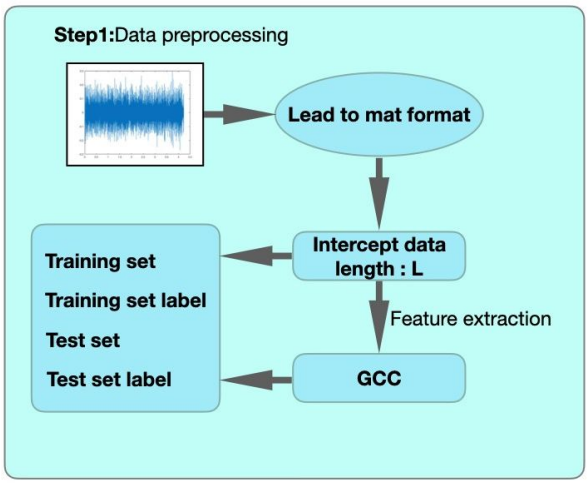

Figure 3: Data acquisition and pre-processing flowchart

spectrum from the original data with the intercept length L. The feature extraction method will be introduced in the next part.

3.2 Feature extraction

In this part, we will introduce the extraction method of the delay feature. The commonly used time delay estimation methods have Generalized Cross-Correlation, Least Mean Square [2] and Cross-power Phase [12]. In this paper, we use the Cross-power Phase to get time delay estimation. This method was first proposed by Knapp *et al.*, as shown in Figure 4:

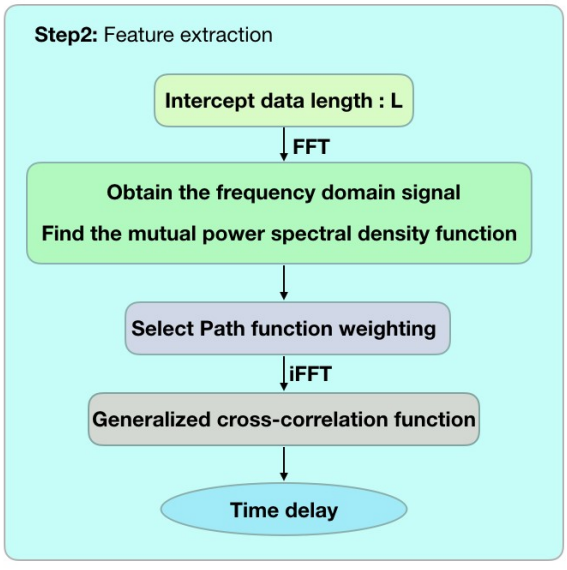

Figure 4: Delay feature extraction flow chart

First, the original data of length L is transformed into a frequency domain signal by fast Fourier transform. The mutual power spectral density function $G_{x_y}(\tau)$ is found.

Next, we perform frequency domain weighting on it. However, some of our audio data is collected in a common room and there may be interference from noise and reverberation, which do not make it possible to observe peaks. In order to make peaks easier to observe, we first need to perform a filtering process on the signals, that is, weighing in the frequency domain, it can strengthen the spectral component of the source signal in the received signal, achieve the function of improving the signal to noise ratio, and then can obtain the higher TDOA delay estimation accuracy. We use Path as the weighting function to get:

$$R_{xy}^{g} = \int_{-\infty}^{+\infty} AG_{x1x2}(\tau)e^{-j2\pi\tau}d\tau \tag{1}$$

Where $A = \frac{1}{|G_{x1x2}(\tau)|}$ represents a generalized cross-correlation weighting function.

Finally, the generalized cross-correlation function based on the cross-power spectrum can be obtained through the inverse Fourier transform. The peak is the delay between the arrival of the sound and the two microphones.

After obtaining the generalized cross-correlation function, we will extract the effective values to form the feature matrix. First measure the distance d between the two microphones, According to the formula:

$$\Delta\tau_{max} = \frac{d}{v}Fs \tag{2}$$

Where v = 340m/s represents the velocity of sound in the air, F_s represents the sampling frequency. Therefore, the number of useful values is 2 τ_{max}. The feature matrix X_{x_y} is:

$$X_{xy} = \left(R_{xy}(-\Delta\tau_{\max}), R_{xy}(-\Delta\tau_{\max}+1)\ldots R_{xy}(\Delta\tau_{\max}-1), R_{xy}(\Delta\tau_{\max})\right)^{T} \tag{3}$$

In this paper, we intercept the length of the original data $256 \div 65536 = 3.9ms$. The distance d between the two microphones was measured to be 13cm. Then the number of valid values is:

$$\Delta\tau_{\max} = \frac{0.13}{340} \times 65536 = 25 \tag{4}$$

We select 50 valid values to form the data set X_{x_y}, in the form of:

$$X_{xy} = \left(R_{xy}(-25), R_{xy}(-24)\ldots R_{xy}(24), R_{xy}(25)\right)^{T} \tag{5}$$

3.3 CNN architecture

This paper designs two architecture, namely Net1 and Net2. Net1 contains two hidden layers. The extracted features pass through the input layer, then they will go through the first hidden layer and are input to the second hidden layer. After the matching operation, they are input into the fully connected layer. Finally, the classification result will output via the Softmax layer, Net2 only contains a hidden layer, and its parameters are shown in Figure 5.

In speech recognition, convolutional neural networks still perform very well. Speech recognition technology is used to model the relationship between spoken speech signals and textual content. However, speech signals are diverse, and we need to be able to distinguish between different signals for sound source localization. For example, different speakers, different timbres and tones, different sound sources, and environmental factors, the convolutional neural network uses the local connection and weight sharing to make it have good translation invariance. We can use this function to overcome the diversity of speech signals. Another aspect is the choice of the number of network layers. This article designed Net1 and Net2. Their hidden layers are two layers and one layer. The reason for not using a deeper network structure lies in the pooling layer. In terms of speech recognition, the largest pooled result is better than the mean pooling, which can reduce the number of output nodes, thereby reducing the amount of calculation and increasing the speed of operation. This is highly critical in practical applications. Besides, choosing maximum pooling can increase the robustness of

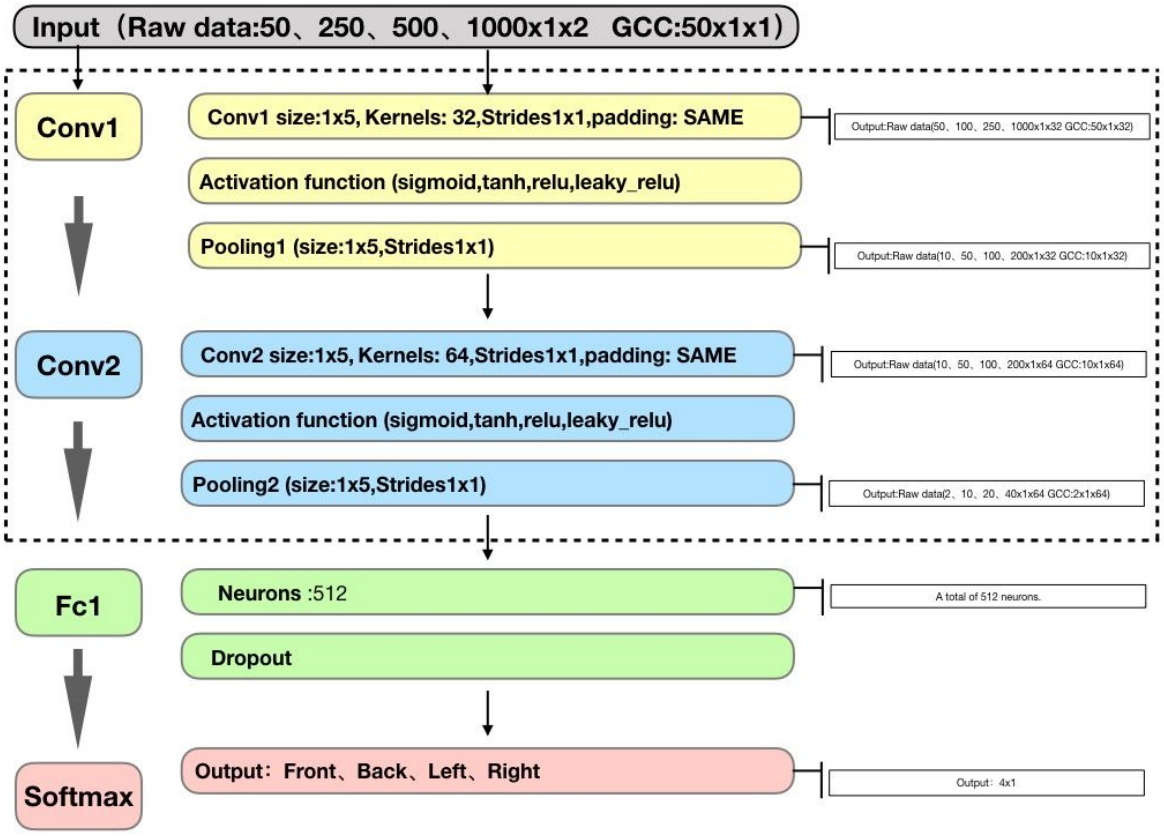

Figure 5: Schematic diagram of the convolutional neural network structure. Net1 contains two convolutional layers, while Net2 contains only one convolutional layer

speech features. After the maximum pooling method is selected, if the number of network layers is too much, it will lead to the loss of features, resulting in a lower recognition rate of experimental results.

In order to maintain the map and features of activated neurons by functions, preserve features and remove some of the data redundancy. We use activation functions to add non-linear factors. In this paper, four commonly used activation functions are selected: sigmoid, tanh, ReLU, and Leaky-ReLU. The cause is that the ReLU activation function is ineffective for negative numbers. In the time delay estimation matrix, we can clearly see that the negative number occupies a large proportion, so we choose different activation functions to compare the recognition accuracy.

4 Experimental evaluation

4.1 Experimental preparation and data acquisition

Experimental environments are the near-field environment of an anechoic chamber and the near-field environment of an ordinary room. As shown in Figure 6. Lab equipment includes Simulated Head, Pulse, Computer, Tape Measure, Router, Marker, and Audio Materials.

Take an anechoic room environment as an example. As shown in Figure 7. First connect the device, place the simulation head in the middle of the anechoic chamber, connect the sensor on the head of the simulation to the Pulse and connect it to the computer through the router. Audio is recorded at 45 degrees, 90 degrees (Left), 180 degrees (Back), 270 degrees (Right), and 360 degrees (Front), respectively, based on the front of the human head. The distances are 40cm, 80cm, 120cm, and 200cm, respectively, and they are marked; finally,

the audio material is played to record sound signals. The types and duration of recorded sounds are shown in Table 1:

Table 1: The type and duration of the audio.

Distance	Experimental environment	
	Anechoic room	**Ordinary room**
40cm	Speaking, Pulse gun, MP3—each 3min	Speaking, Pulse gun, MP3—each 3min
80cm	Speaking, Pulse gun, MP3—each 3min	Speaking, Pulse gun, MP3—each 3min
120cm	Speaking, Pulse gun, MP3—each 3min	Speaking, Pulse gun, MP3—each 3min
200cm	Speaking, Pulse gun, MP3—each 3min	Speaking, Pulse gun, MP3—each 3min

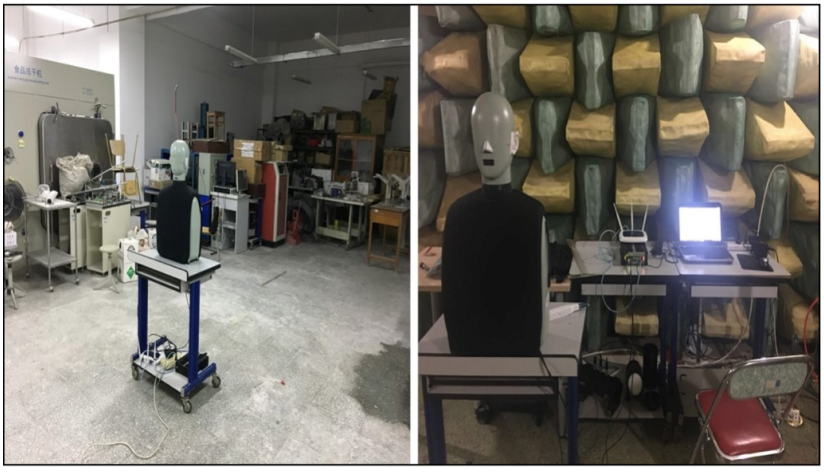

Figure 6: The left picture shows the ordinary room environment, and the right picture shows the anechoic room environment

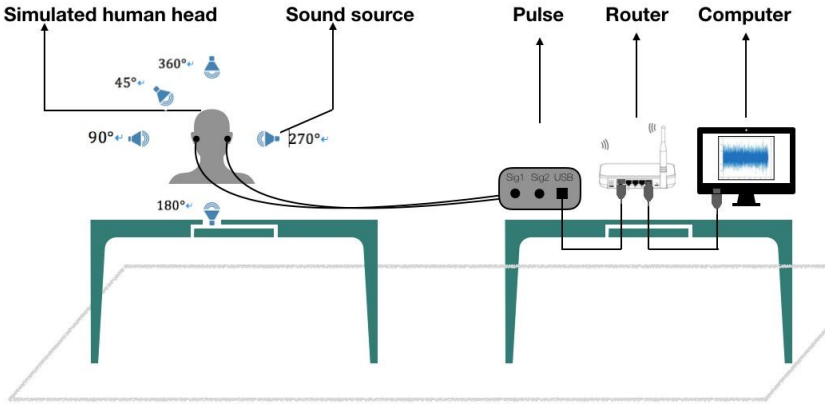

Figure 7: Schematic diagram of experimental equipment connection

4.2 Data set

The collected audio data are intercepted 2/3 as training data, and the remaining 1/3 is used as test data. The data set of each room includes two parts. One is a dual-channel data set based on the original data, and the other is a GCC dataset obtained by time delay estimation.

Among them, Set the value of L to 50,250,500,1000. The dataset contains four files, training set, the training set label, test set, and test set label, which are composed of a header file and real data. Header files of the training set and test set are the experiment number (3331), total data number, data line number, data column number, followed by the first number of channel 1 data, the first number of channel 2... The 50th (250th, 500th, 1000th) number of channel 1 and the 50th (250th, 500th, 1000th) number of channel 2. Training set labels and test set labels are also composed of header files and real data in the format of the experiment number (2049), the total number of data, and the numerical representation of each direction. The total amount of data is given in Table 2.

Table 2: Total amount of data of different size data.

Data size	Data number of Training set	Data number of Testing set
50×1	900000	400000
250×1	180000	80000
500×1	90000	40000
1000×1	45000	20000

The production of GCC dataset starts by taking a piece of audio from the raw data in each direction. Set L to 256. The value of the ordinate corresponding to −25 to 25 on the abscissa in the graph is extracted. These data are then integrated into a feature matrix of size 50 × 1. The dataset also includes the training set, the training set label, the test set, and the test set label. The header file is the same as the raw data, and the real data is the feature matrix in each direction. The total number of training set data is 180000, and the total number of test set data is 80,000.

4.3 Experimental results

4.3.1 Experimental results based on raw data

Using the raw data dual-channel data set, the data is a feature matrix of size 250 × 1. The network uses Net1 and Net2. The experimental results are shown in Tables 3 and 4:

Table 3: Experimental results based on Net1.

Net1								
	Anechoic room				**Ordinary room**			
	Sigmoid	**tanh**	**ReLU**	**Leaky_ReLU**	**Sigmoid**	**tanh**	**ReLU**	**Leaky_ReLU**
40cm	67.9%	92.7%	89.6%	89.9%	71.1%	79.6%	80.5%	80.3%
80cm	69.8%	90.3%	91.0%	88.9%	60.7%	70.7%	71.6%	70.5%
120cm	66.6%	91.6%	90.1%	91.2%	57.1%	72.8%	72.8%	72.3%
200cm	68.9%	90.2%	88.4%	88.3%	46.8%	60.0%	59.5%	60.7%

Table 4: Experimental results based on Net2.

Net2	Anechoic room				Ordinary room			
	Sigmoid	**tanh**	**ReLU**	**Leaky_ReLU**	**Sigmoid**	**tanh**	**ReLU**	**Leaky_ReLU**
40cm	66.2%	80.2%	81.6%	81.5%	68.2%	77.1%	78.8%	78.3%
80cm	64.8%	79.9%	81.1%	80.7%	57.6%	65.8%	67.6%	66.5%
120cm	52.0%	80.9%	80.1%	79.8%	44.5%	68.3%	67.9%	68.4%
200cm	43.2%	79.8%	79.9%	79.6%	25.3%	55.8%	57.0%	56.9%

4.3.2 Experimental results based on delay features

Using a delayed data set, the data is a feature matrix of size 50×1, The net uses Net1 and Net2. The experimental results are shown in Tables 5 and 6:

Table 5: Experimental results based on Net1

	Anechoic room				Ordinary room			
	Sigmoid	**tanh**	**ReLU**	**Leaky_ReLU**	**Sigmoid**	**tanh**	**ReLU**	**Leaky_ReLU**
40cm	64.6%	59.3%	59.6%	55.3%	62.7%	57.8%	51.4%	49.8%
80cm	64.8%	63.8%	63.6%	60.1%	51.2%	49.6%	46.8%	47.2%
120cm	58.8%	60.5%	56.0%	57.6%	55.7%	49.5%	48.6%	31.2%
200cm	58.7%	57.6%	56.6%	57.0%	42.7%	40.0%	39.5%	33.7%

Table 6: Experimental results based on Net2

	Anechoic room				Ordinary room			
	Sigmoid	**tanh**	**ReLU**	**Leaky_ReLU**	**Sigmoid**	**tanh**	**ReLU**	**Leaky_ReLU**
40cm	63.7%	59.6%	60.8%	52.6%	62.0%	58.2%	54.5%	51.9%
80cm	64.0%	61.0%	65.1%	60.4%	51.6%	51.2%	57.4%	40.9%
120cm	58.7%	57.4%	60.8%	57.2%	54.9%	52.2%	56.4%	50.3%
200cm	59.3%	56.4%	59.6%	56.5%	57.5%	52.4%	52.5%	35.3%

4.3.3 Experimental results based on different data sizes

From the above experimental results, we can see that the recognition results of the dual-channel raw data are better than the delay features. On this basis, we also explored the impact of each data size on the recognition results. The experimental data size is set to 50, 500, 1000 respectively. In addition, we can see that the results of Net1 are better than that of Net2, so only Net1 is chosen for the experiment. The experimental results are shown in Tables 7 and 8:

Table 7: The accuracy of different data sizes in the anechoic chamber.

Sound distance	Activation function / Data size											
	Sigmoid			tanh			ReLU			Leaky_ReLU		
	50	500	1000	50	500	1000	50	500	1000	50	500	1000
40cm	66.4%	73.8%	76.7%	75.0%	92.6%	83.9%	75.9%	94.7%	90.7%	74.2%	92.3%	77.6%
80cm	63.2%	73.3%	66.2%	72.6%	94.0%	88.5%	75.3%	92.8%	87.8%	71.9%	90.1%	72.3%
120cm	57.0%	70.7%	24.9%	71.7%	94.6%	90.8%	74.7%	92.1%	91.2%	72.4%	93.8%	89.1%
200cm	31.3%	73.0%	25.4%	73.7%	92.7%	93.2%	74.8%	93.0%	91.6%	73.9%	89.9%	93.4%

Table 8: The accuracy of the identification of different data sizes in common rooms.

Sound distance	Activation function / Data size											
	Sigmoid			tanh			ReLU			Leaky_ReLU		
	50	500	1000	50	500	1000	50	500	1000	50	500	1000
40cm	60.7%	72.6%	68.8%	68.2%	83.1%	72.7%	69.8%	84.8%	78.9%	69.6%	83.8%	76.7%
80cm	50.0%	64.3%	58.9%	55.8%	74.4%	66.3%	57.0%	76.0%	73.0%	55.9%	76.0%	72.0%
120cm	42.5%	65.8%	48.7%	57.0%	74.7%	55.0%	58.6%	76.5%	54.0%	58.8%	76.8%	54.7%
200cm	35.8%	24.9%	24.9%	44.1%	65.1%	62.9%	44.9%	64.3%	61.9%	44.0%	65.0%	62.2%

4.4 Analysis of results

4.4.1 Analysis of recognition results of different features

In an anechoic room environment, the comparison of the delay features and the identification results of the raw data is shown in Figure 8.

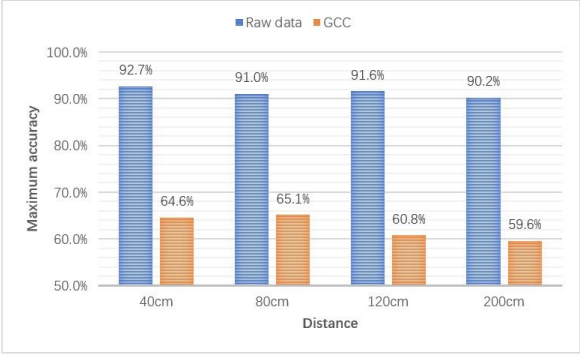

Figure 8: Based on the comparison of different feature recognition results

We can clearly see that the recognition results of the two-channel raw data are much better than the GCC features. It is well known that in conventional sound source localization methods, at least three microphones are required for positioning using delay estimation. Therefore, the time delay estimation is used as the input of the neural network, and the model cannot distinguish the front and back (left and right). Because the features in both directions are almost identical. However, when we used the two-channel raw data as the input of the neural network, we got a good recognition result. We speculate that the possible reason is that the sounds in different directions may differ more significantly in the raw data. Therefore, the recognition accuracy based on the original data is higher than the delay feature. In the discussion that follows, we focus on the raw data.

4.4.2 Influence of different indoor environments on experimental resultsn

Our experimental environment is divided into anechoic chambers and ordinary room. The reason why the ordinary room is chosen is that they are closer to the real environment. They are characterized by the possibility of noise and reverberation. The experimental results are shown in Figure 9:

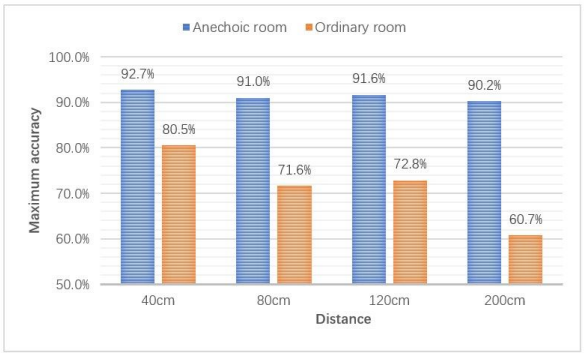

Figure 9: Recognition results in different indoor environments based on raw data

As can be seen from Figure 8, the recognition rate of the ordinary room is lower than that of the anechoic chamber environment. Moreover, the recognition rate is getting lower as the distance increases. The possible reason is that the ordinary room has more noise and reverberation interference than the anechoic chamber. Therefore, this model is intended to be applied to actual engineering, and audio is also pre-processed to eliminate noise and reverberation interference.

4.4.3 The effect of different data sizes on the recognition accuracy

In the previous section, we discussed that the recognition rate of the ordinary room is lower than that of anechoic room. Therefore, we used different data sizes for experiments. The reason is that under the premise of ensuring accuracy, using smaller data as much as possible can greatly reduce the recognition time and improve efficiency. The result is shown in Figure 10.

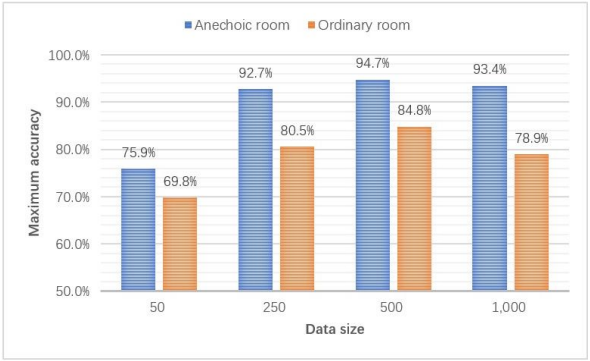

Figure 10: The result is compared after changing the data size

In the figure, after changing the data size, the maximum recognition result is obtained at different distances. Although we have chosen larger data, it seems that the recognition accuracy has not improved much. However, we have found that when we choose larger data, the accuracy of the training set is high, which can

reach almost 1. The test accuracy can reach more than 0.9 in the anechoic room environment, but not in the ordinary room, Perhaps overfitting is one of the reasons. We will continue to explore in the next work.

4.4.4 The influence of network structure on the recognition accuracy

Taking an anechoic room environment as an example, we compare the effect of different network structures on the recognition accuracy. The classification results on two different networks are shown in Figure 11.

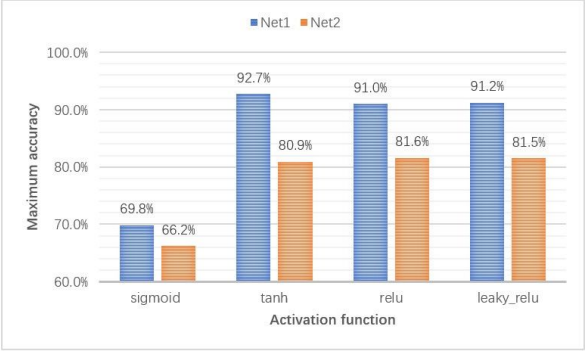

Figure 11: Different network identification results of raw data

For the original feature, Net1 is about 0.1 more accurate than Net2. This phenomenon seems to indicate that the deep model is more practical. However, using the deep model is not realistic due to data size and maximum pooling restrictions. Moreover, our test accuracy can reach 0.92. For delay data, the accuracy of Net1 and Net2 is not really satisfactory.

Another improvement for the network is activating the function. For raw data, tanh is chosen as the activation function to obtain the highest recognition accuracy. Moreover, ReLU and Leaky-ReLU also have very good results. However, the accuracy of the sigmoid is much lower. The possible reason for this is that the input is far from the origin of coordinates and the gradient of the function becomes very small, almost zero. Eventually, it leads to gradient saturation. For the delay feature, although the recognition accuracy of the sigmoid is not very different from that of other activation functions, the sigmoid requires an exponential operation and has a long running time.

5 High resolution directional model based on probability distribution

In this article, we only conducted positioning experiments for the four directions of the simulated human head. In order to achieve high-resolution orientation, we hope to expand the direction of the sound source to eight and more directions, but we do not want to increase Multi-directional training data, therefore, we propose a high-resolution positioning model based on probability distribution.

Firstly, the network model is trained by using four directions of audio data. When the training is completed, we input a test data of 90 degrees, and perform numerical processing on the final output unit through the Softmax layer to obtain a sound source probability of P_1 in the direction of 90 degrees. The probability of the sound source in the 180-degree direction is P_2, the probability of the sound source in the 270-degree direction is P_3, the probability of the sound source in the 360-degree direction is P_4, and when the recognition is correct, there is: $P_1 > P_2, P_3, P_4$.

Based on this well-trained network model, we want to predict more directions. We guess that there is a certain relationship between the directions of the source points. If such an algorithm is found, the positive direction is related to other directions. And it gives a new method of discriminating between right and wrong in the oblique direction. It is a possibility to use the n-class model to predict the 2n-class problem. Therefore, we design an algorithm to convert the probability of the model output into an unknown direction. And we will call this algorithm a high-resolution algorithm, the specific steps are as follows:

Step one: First, we define the number classifications n of the model which will be trained, divide 360° equally into n-class and collect sound data in n directions for feature extraction. Then we make a data set and train the neural network, and finally save this well-trained network model.

Step two: We divide 360° equally into 2n-class, and collect the sound data between n and n+1 direction's center, a total of n data is collected. The feature of the new data is extracted by the extraction feature of step one. The data set is only included in the test set. The new test set is input into the trained network model, and the model outputs n probability values. We record them as $P_1\ P_2 \ldots P_n$.

Step three: Using the vector decomposition method, multiply the probabilities of the n directions by the angle between the data directions of the two acquisitions. Then we let the vector sum in the two adjacent directions, and finally get the probability of the center direction, the expression is as follows:

$$N_{\mathrm{i}} = \begin{cases} P_{\mathrm{j}} \cos\frac{\pi}{n} + P_{\mathrm{j+1}} \cos\frac{\pi}{n} & i \neq \frac{\pi}{n} \\ P_{\mathrm{n}} \cos\frac{\pi}{n} + P_1 \cos\frac{\pi}{n} & i = \frac{\pi}{n} \end{cases} \tag{6}$$

Where j is 1, 2, 3...n, P_n is the probability of direction n and N_i is the probability of oblique direction i.

Step four: Based on the high-resolution algorithm, we add a new filter condition. According to the empirical algorithm, when the probability of an oblique direction is greater than a confidence probability N_v, we can think that it is correct, that means, N_i and N_v are compared. If $N_i > N_v$, the judgment is correct.

When we are making the oblique direction test set, we know the label corresponds to each data, that means, the direction of the sound source. The purpose of setting the confidence probability is to give a value range based on the real position, if we finally obtained the probability which is within this range by the high resolution algorithm. We can determine that the classification is correct. A schematic diagram of the algorithm is shown in Figure 12.

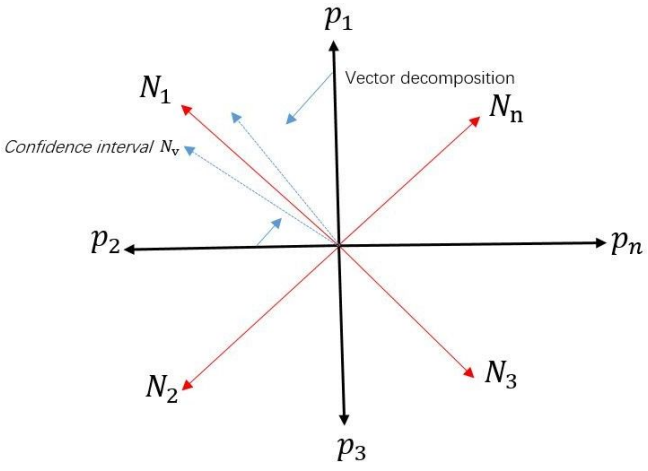

Figure 12: Schematic diagram of high resolution algorithm

In order to verify the accuracy of the high-resolution algorithm, we designed a set of experiments: Taking n as 4 as an example and collecting four oblique data. Finally we obtained the results by the above algorithm are shown in Table 9.

Table 9: The effect of confidence probability on the algorithm.

confidence probability	Recognition accuracy
0.3	99.98%
0.4	77.69%
0.5	49.99%

It can be seen from Table 9 that when the confidence probability is equal to 0.3 or 0.4, the result of the direction discrimination both can reach a higher probability. Thus the results verify our previous guess that there is a certain relevancy between the directions of the sound source points and also validates the practicality of our algorithm.

In this article, we use the dual-channel raw data as the input feature of the neural network to design a dual microphone model that simulates the human ear, At the same time, the high-resolution directional model doubles the recognition accuracy of the sound source direction. The model based on the convolutional neural network shows good performance, and the accuracy on the test set reaches more than 0.9. When the confidence probability is set to 0.3, the recognition accuracy of the high-resolution directional model reaches 0.99. The dual microphone model that simulates the human ear has a wide range of application prospects in the project. The reduction in the number of microphones can reduce the consumption of materials and cost of the enterprise, and the dual microphone model can also achieve the recognition accuracy of the microphone array. The simulation of the human ear enhances the credibility of the dual microphone system, making it more widely used. Among them, the high-resolution directional model is a practical application of simulating the human ear, which makes the recognition accuracy of the direction of sound source much improved.

6 Conclusions

In this article, a two-microphone sound direction recognition model is proposed. Through this article, we can have the following benefits:

(1) The theoretically untrustworthy dual microphone model can be used to identify the direction of the sound, thus achieving some of the functions of the human ear.
(2) We can use the convolutional neural network to identify the sound direction. When the input is two-channel raw data, we can get a very good result.
(3) The anti-noise and anti-reverberation of this model are not strong. If it is carried out in an ordinary room, the audio data must be pre-processed to remove noise interference.
(4) When the size of the selected data is larger, the improvement of the test accuracy is not obvious, but we found that the training accuracy can be improved to almost 1 during the training.

Finally, we performed high-resolution directional experiments based on probability distributions, which are no longer limited to four directions. A high resolution algorithm is designed to achieve higher resolution accuracy. In the next job, it also enhances the model's noise immunity and reverberation capabilities, allowing it to be used in complex environments.

Acknowledgement: The authors would like to acknowledge funding received from the National Natural Science Foundation of China (Grant No. 61877002,51405005).

References

[1] X. Alameda-Pineda and R. Horaud, A geometric approach to sound source localization from time-delay estimates, *IEEE/ACM Transactions on Audio Speech and Language Processing*, **22**(2014), no.6, 1082-1095.

[2] Y. Azenkot and I. Gertner, The least squares estimation of time delay between two signals with unknown relative phase shift, *IEEE Transactions on Acoustics Speech and Signal Processing*, **33**(2014), no.6, 1082-1095.

[3] A. Canclini, F. Antonacci and A. Sarti, Acoustic source localization with distributed asynchronous microphone networks, *IEEE Transactions on Audio Speech and Language Processing*, **21**(2013), no.2, 308-309.

[4] J. P. Dmochowski, J. Benesty and S. Affes. A generalized steered response power method for computationally viable source localization, *IEEE Transactions on Audio Speech and Language Processing*, **15**(2007), no.8, 2510-2526.

[5] A. Elmar, and D. Soffker, Learning from interaction with the environment using a situation-operator calculus with application to mobile robots, *IEEE International Conference on Systems, Man and Cybernetics*, (2004), 3839-3844.

[6] H. He, L. Wu, J. Lu, X. Qiu and J. Chen, Time difference of arrival estimation exploiting multichannel spatio-temporal prediction, *IEEE Transactions on Audio, Speech, and Language Processing*, **21**(2013), no.3, 463-475.

[7] A. Iozsa, Adaptive Beamforming applied for signals estimated with direction-of-arrival algorithms from the ESPRIT family, *International Symposium on Electronics and Telecommunications, Timisoara, Romania*, (2012), 397-400.

[8] B. Ivancevic, K. Jambrosic and A. Petosic, Binaural hearing computer models in multisource environments, *International Conference on Applied Electromagnetics and Communications*, (2005), 471-474.

[9] H. Liu and J .Zhang. A binaural sound source localization model based on time-delay compensation and interaural coherence, *IEEE International Conference on Acoustics, Speech and Signal Processing, Florence, Italy*, (2014), 1424-1428.

[10] H. Li, T. Zhao, N. Li, Q. Cai and J. Du, Feature matching of multi-view 3D models based on hash binary encoding, *Neural Network World*, **27**(2017), no.1, 95-105.

[11] B. Mungamuru and P. Aarabi, Enhanced sound localization, *IEEE Transactions on Systems Man & Cybernetics*, **34**(2014), no.3, 1526-1540.

[12] M. Matassoni and S. Piergiorgio, Efficient time delay estimation based on cross-power spectrum phase, *European signal processing conference*, (2006), 1-5.

[13] V. P. Minotto, C. R. Jung and B. Lee, Simultaneous-speaker voice activity detection and localization using mid-fusion of SVM and HMMs, *IEEE Transactions on Multimedia*, **16**(2014), no.4, 1032-1044.

[14] H. Niu and P. Gerstoft. Source localization in an ocean waveguide using supervised machine learning, *Journal of the Acoustical Society of America*, **142**(2017), no.3, 1176-1188.

[15] D. Pavlidi, A. Griffin and M. Puigt, Real-time multiple sound source localization and counting using a circular microphone array, *IEEE Transactions on Audio Speech and Language Processing*, **21**(2013), no.10, 2193-2206.

[16] A. Saxena and A.Y. N g, Learning sound location from a single microphone, *International Conference on Robotics and Automation, Kobe, Japan*, (2019), 4310-4315.

[17] J. Velasco, J. Macias-Guarasa and D. Pizarro, Proposal and validation of an analytical generative model of SRP-PHAT power maps in reverberant scenarios, *Signal Processing*, **119**(2013),209-228.

[18] F. Vesperini, P. Vecchiotti and E. Principi, Localizing speakers in multiple rooms by using deep neural networks, *Computer Speech and Language*, **49**(2018), 83-106.

[19] Z. Xing, W. Xue and L. Chang, Sensor array based predicted spatial multi-signal classification method for target localization and tracking, *Chinese Journal of Scientific Instrument*, **33**(2012), no.5, 970-975.

[20] D. Yook, D. Lee D and Y. Cho, Fast sound source localization using two-level search space clustering, *IEEE Transactions on Cybernetics*, **46**(2016), no.1, 20-26.

[21] Z. X. Yao, K. Y. Jiang and R. Guo, An improved bearing estimation algorithm using acoustic vector sensor array based on rotational invariance technique, *Transactions of Beijing Institute of Technology*, **32**(2012), no.5, 513-516+521.