IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received October 23, 2018, accepted November 19, 2018, date of publication November 26, 2018,
date of current version December 19, 2018.

Digital Object Identifier 10.1109/ACCESS.2018.2883341

A New Regional Localization Method for Indoor Sound Source Based on Convolutional Neural Networks

XIAOMENG ZHANG, HAO SUN, SHUOPENG WANG, AND JING XU
School of Artificial Intelligence, Hebei University of Technology, Tianjin 300130, China

Corresponding author: Hao Sun (sunhao@hebut.edu.cn)

This work was supported in part by the Opening Fund of the National Natural Science Foundation of China under Grant 61773151, Grant 61703135, and Grant 61503118 and in part by the Natural Science Foundation of Hebei Province under Grant F2017202119 and Grant F2016202327.

ABSTRACT At present, the sound source localization methods based on microphone arrays can be roughly classified into three categories: the controllable beamforming technology based on a maximum output power, the high-resolution spectrogram estimation technique, and the sound source localization technique based on time difference of sound. However, an existing localization technology in unstructured indoor environment lacks of localization accuracy and adaptability. In some practical situations, the location of sound source is limited to predefined areas. In this paper, we propose a research method of source region location system based on convolutional neural networks (CNNs). Based on the characteristics of weighted values of CNN, we realize the regional of indoor single sound sources transforming the sound source signals into grammar diagrams and then inputting them into the CNN. The whole process is based on the characteristics of weighted values of CNN. Finally, this paper completes the training and testing for CNN by using the Tensorflow framework. Simulation experiments on the test sets show the effectiveness of the proposed method.

INDEX TERMS Sound source localization, machine learning, spectrogram, CNN.

I. INTRODUCTION

With the development of voice processing technology and artificial intelligence technology, natural voice has gained much attention in the field of human-computer interaction in its friendly way. As an important medium of human-computer interaction, voice has significant advantages over other sources, such as omnidirectionality, high time resolution and the ability of spreading out of visibility. Using microphone arrays for sound source localization research has been a research hotspot in the field of signal processing [1]–[5]. It has already been applied in search, rescue, mobile robot, voice recognition, acoustic detection and intelligent vehicle [6]–[13] as well as other aspects.

At present, the traditional sound source localization methods based on microphone array can be roughly divided into three categories: the controllable beamforming technology based on a maximum output power, the high-resolution spectrogram estimation technology and the sound source localization technique based on time difference of sound [14]–[20]. Reference [21] used beam- forming algorithms for sound source localization. Although the array optimization improves the performance of the sound source localization system, it requires more microphones and higher computational complexity. Reference [22] used a music algorithm based on high-resolution spectral estimation and achieved a multi-target localization technique. However, this technique is only applicable to far-field models, and the reverberation generated by the reflection will also seriously interfere with the localization accuracy. Reference [23] proposed an iterative least square method, according to which, the GCC-PHAT method is used to find the sound localization. However, the sound source localization technique based on the time difference of sound takes the time-delay feature as an input variable. The obstruction and reflection of objects indoor such as tables and chairs, glass and walls still cause some delay deviation. With the popularity of pattern-based source localization algorithms, more and more researchers begin to investigate the sound source localization based on machine learning algorithms [24]–[27]. Reference [28] applied the BP neural network to the sensor array to improve the accuracy of localization in the indoor environment to certain extent. However, there still have some problems such as the overlong training time, insufficient localization accuracy and insufficient fitting ability of the model. Reference [29] proposed a sound source localization method based on a Gaussian mixture model, which used the phase difference between the array elements caused by the mismatch between the room acoustics and the microphone to locate. Reference [30] proposed to use the LS-SVM (Least Square Support Vector Machine) to identify TDOA for sound source localization, but the localization effect in bad environment is not ideal. Reference [31] proposed a RBF kernel support vector machine to construct a weighted minimum variance distortion-free response beamformer, which could effectively deal with the single-source localization problem in near-field. However, the computational complexity of the model is too high to apply in many real world applications. Ma *et al.* [32] proposed a machine-hearing framework which combines DNNs and head movements for robust localization of multiple sources in reverberant conditions. However, the self-noise generated by head rotation increases the number of the localization errors and this study merely adapted to the case where the sound source is stationary and the number of active sources is verified. In the same year, Vesperini et al developed completely data driven approach for Speaker Localization (SLOC) in multi-room environment, considering both the moving and the stationary conditions. The multi-room speaker localization algorithm is implemented by DNN-SLOC. Since the proposed algorithm consists of feature extraction and artificial neural network, the computational complexity of the algorithm is high. Besides, the localization of the microphone has a great influence on the accuracy of localization.

In recent years, deep-learning has achieved great success, which greatly promotes the development of machine learning and attaches great importance of researchers in relevant area as well as some high-tech companies all over the world. CNN are a typical kind of deep neural network [34], [35]. Its weight-sharing network structure makes it more similar to biological neural network, Which reduces the complexity of network model and the number of weights. Since this network structure has high invariance in translation, scaling, tilting and other forms of deformation, it has been widely used in image processing [36]–[39].As a visual representation of the time-frequency distribution of speech energy, the spectrogram itself contains speech features such as energy, pitch, and fundamental frequency. There are already some researchers using the spectrogram to combine image processing with speech processing, which made a good achievement [40]. Reference [41] proposed a speaker separation technique for the spectral Radon transform and discrete cosine transform. Reference [42] proposed a novel single channel speech dereverberation method using guided spectrogram filtering by considering a speech spectrogram as an image.

In this paper, we deal with the issue of sound source localization from the perspective of machine learning. With the observation in some practical situations, that the localization of the sound source of interest is only limited to some predefined areas [43], and the existing localization technology in unstructured indoor environment is lacking of location accuracy and adaptability, this paper proposes a research scheme for indoor sound source regional localization based on convolutional neural network. To our best knowledge, this is the first application to solve the indoor single source localization under the condition of convolutional neural network and spectrogram. The reason why we use convolutional neural networks is that compared to other existing machine learning methods; convolutional neural networks have translational invariance. Translation invariance refers to a mode that can be recognized by CNN regardless of its position at the input. This feature coincides with the need to identify a large number of repetitive local patterns in the spectrogram. Which also motivates us to apply CNN in this work. Besides, CNN uses weight sharing which means less training parameters, and can bring better robustness and generalization. Firstly, we convert the sound source signal collected by the microphone into a spectrogram to construct the location dataset and input it into the CNN for training, to realize the regional localization of the indoor single sound source. Then, we use Tensorboard to visualize the training and test results of the CNN, making the training process of the CNN more intuitive. Finally, we compare the proposed model with KNN (K-Nearest Neighbor), BP (Back Propagation) neural networks and SVM (Support Vector Machine). The simulation results verify the effectiveness of the proposed method.

II. BACKGROUND

A. BUILDING A SIGNAL MODEL

It is assumed that the propagation of sound satisfies the linear wave equation, and then the sound wave propagation channel between the sound source and the microphone can be considered as a linear system [44]. Actually, in a small room environment, considering the reflection of the room wall, the speech signal is multipath propagated in the room, which causes the amplitude attenuation of the received signal and the deterioration of the quality of the sound. This is the reverberation effect. Since reverberation affects the performance of a voice microphone array system, the room impulse response model must be conducted by multipath propagation. Suppose the signal received by the nth microphone is $x_i(t)$, which can be expressed as [45], [46]:

$$x_i(t) = s(t) * h_i(t) + v_i(t) \tag{1}$$

Where $s(t)$ is the sound source signal, $v_i(t)$ is the noise, $h_i(t)$ is the total impulse response, and '$*$' is the convolution operator. $h_i(t)$ is a function of the sound source spatial direction and the localization of the microphone, which is the output of the two-stage cascade filter of the room impulse response and the microphone channel response. The former includes all the characteristics of the acoustic path from the sound source to the microphone, including the direct path. $v_i(t)$ is the sum of acoustic multipath reflection interference and ambient noise. Usually, ambient noise is more significant than channel noise and is the main part of $v_i(t)$. Assuming that $s(t)$ is not related to $v_i(t)$, considering the impulse response of the direct path component, which can be showed as

$$x_n(t) = \frac{1}{r_n} s(t - \tau_n) * g_n(t) + v_n(t) \tag{2}$$

Where r_n represents the distance between the sound source and the microphone, τ_n represents the delay of the direct path, and $g_n(t)$ represents the corrected impulse response, which consists of the original response minus the direct path response.

The system uses four sensors and the sound source is located in a four-way microphone array. The system model in two-dimensional space is shown in Fig.1.

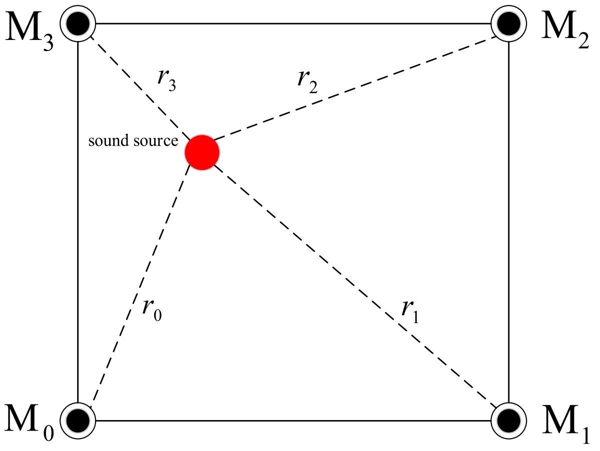

FIGURE 1. System model.

The system model in two-dimensional space is shown in Figure 1.This paper will convert the voice signals captured by the four microphones into grammar diagrams and use the genograms as the input of the CNN. After the CNN training, the classification results are obtained, so that the indoor sound source regional localization research is conducted.

B. SPECTROGRAM

Spectrogram [47] is a graph showing the change of speech spectrogram over time. It samples a two-dimensional plane to express three-dimensional information. The vertical axis represents frequency, the horizontal axis represents time, and the intensity or color of each point in the image represents the value of the energy. The deeper color is, the stronger speech energy will be at this point [48]. The spectrogram shows a large amount of information related to the speech features of the speech. It combines the characteristics of the spectrogram and the time-domain waveforms, clearly showing the change of speech spectrogram over time, or a dynamic spectrogram [49], [50]. The basic mathematical expression is

$$X(\omega, \tau) = \sum_{k=-\infty}^{\infty} \omega(k, \tau) x(k) e^{-j\omega k} \tag{3}$$

In the formula: the integral interval in equation (3) is $(-\infty, +\infty)$, which is the entire time axis, ω represents the angular frequency, j is the imaginary number, and $X(\omega, \tau)$ is a two-dimensional function which represents the Fourier transform of the windowed sound whose center point is located at τ, $\omega(k, \tau)$ is a window function of length N, and $x(k)$ represents the sound signal of the harmonic component number $k = 0, 1, \ldots, N - 1$.

In summary, the process of implementing the spectral diagram is shown in Fig.2.

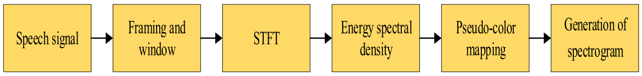

FIGURE 2. The realization of the language spectrogram.

C. CONVOLUTION NEURAL NETWORKS

CNN is a feed-forward neural network, which mainly includes an input layer, a feature extraction layer composed of one or more sets of convolution layer+pooling layers, a full connection layer, and an output layer, as shown in Fig.3. Multiple convolution layers and pooled layers are alternately combined to form the feature extraction stage. Finally, the final classification is obtained by integrating the output values of the pooled layers through the fully connected layer [51].

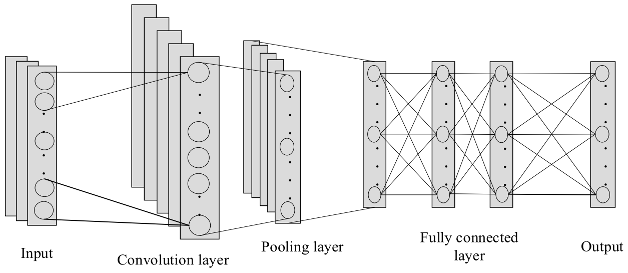

FIGURE 3. Convolution neural network.

III. BUILDING A LOCALIZATION DATASET

In this section, this paper introduces the research method of regional localization for sound source based on convolutional neural networks, and the specific flow of simulation experiment is given in detail.

A. DATA SAMPLE SELECTION

At the data sample selection stage, we need to locate the localization reference point in the area, then collect the sound information sent by the sound source at each reference point. As is shown in Fig.4, in the microphone array (M_0, M_1, M_2, M_3) in the two-dimensional coordinate system, the distance between adjacent micro- phones is $d = 10.2m$, and the microphone M_0 is the origin of the coordinates. Taking the microphone M_0 as the coordinate origin, the coordinates of the other three microphones are $M_1(d, 0)$, $M_2(d, d)$, $M_3(0, d)$. In the square matrix formed by the microphone

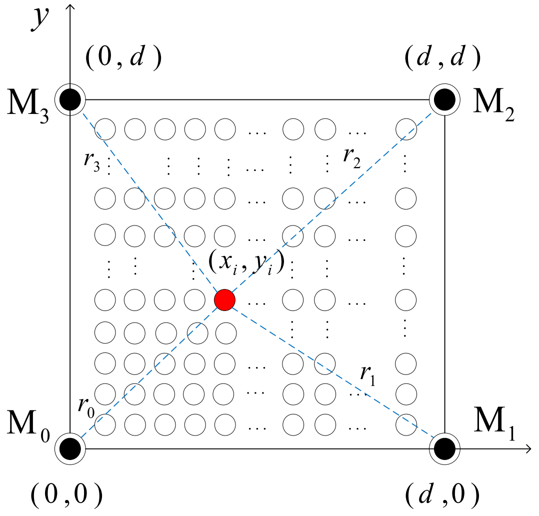

FIGURE 4. A two-dimensional model of microphone array.

(M_0, M_1, M_2, M_3), the uniform $100 \times 100 = 10000$ reference points are arranged, and the coordinates of the reference point at any localization of the sound source are (x_i, y_i). Taking an actual speech signal received by the microphone M_0 as a reference (The sound source signal is a ringtone of the mobile phone in the reality), we assume that the speech signal received by the microphone M_0 at each reference point are the same. We can obtain as follows [52]:

The distance from the sound source to the M_0 microphone:

$$r_0 = \sqrt{x_i^2 + y_i^2} \tag{4}$$

The distance from the sound source to the M_1 microphone:

$$r_1 = \sqrt{(x_i - 10.2)^2 + y_i^2} \tag{5}$$

The distance from the sound source to the M_2 microphone:

$$r_2 = \sqrt{(x_i - 10.2)^2 + (y_i - 10.2)^2} \tag{6}$$

The distance from the sound source to the M_3 microphone:

$$r_3 = \sqrt{x_i^2 + (y_i - 10.2)^2} \tag{7}$$

When the sound propagation speed $c = 340m/s$, the time difference between the microphones M_1, M_2, M_3 and the microphone M_0 is:

$$\Delta t_i = (r_i - r_0)/340 i = 1, 2, 3 \tag{8}$$

The length of the actual voice signal received by the microphone M_0 is $N = 100k$, then the number of backward moving points of the microphone M_1, M_2, M_3 relative to the M_0 microphone is:

$$\Delta d_i = \Delta t_i \times 100k \quad i = 1, 2, 3 \tag{9}$$

From formula (4-8), we get

$$\begin{cases} \Delta d_1 = \dfrac{\sqrt{(x_i - 10.2)^2 + y_i^2} - \sqrt{x_i^2 + y_i^2}}{340} \times 100k \\ \Delta d_2 = \dfrac{\sqrt{(x_i - 10.2)^2 + (y_i - 10.2)^2}}{340} \times 100k \\ \Delta d_3 \dfrac{r_3 = \sqrt{x_i^2 + (y_i - 10.2)^2}}{340} \times 100k \end{cases} \tag{10}$$

According to the obtained sound arrival time difference and the number of delay points of each microphone relative to M_0, we can obtain the sound signals collected by the four microphones at each reference point localization, for a total of 10000 data samples.

B. THE ESTABLISHMENT OF LOCALIZATION DATASET

The recognition of speech signals are mainly from the time domain and frequency domain, but the time domain signal cannot represent the frequency characteristics; the frequency domain cannot show the characteristics of changes with time. So we consider using spectrograms as inputs to CNN.

The experimental data is generated in the matlab environment. Since the simulation is in the indoor environment, the distance between the microphones and the sound source is different. Besides the received signal not only has a phase difference but also has amplitude attenuation caused by the propagation of the sound waves in the air. So we use formula (11):

$$t' = (rand \times 2 - 1) \times 5000 \tag{11}$$

Give the signal received by each microphone a random delay of 5000 points, then superimpose the delayed signal and the original signal to achieve the effect of reverberation, through the signal-to-noise ratio formula (12):

$$SNR = 10 \log_{10} \frac{\sum\limits_{n=0}^{N-1} s^2(n)}{\sum\limits_{n=0}^{N-1} d^2(n)} \tag{12}$$

Gaussian white noise is added to the speech signal. In the formula, $\sum\limits_{n=0}^{N-1} s^2(n)$ denotes the energy of the signal; $\sum\limits_{n=0}^{N-1} d^2(n)$ denotes the energy of the noise. Reference [53] clarifies that the energy of the point source signal is inversely proportional to the square of the distance from the sound source to the receiving point, according to the formula (13):

$$s = \frac{s_0}{4\pi \cdot d^2} \tag{13}$$

The attenuation model of the sound energy can be constructed, where s_0 is the energy of the signal at the sound source.

The sampling frequency of the sound signal is 100 kHz. Four spectrograms can be generated at each reference point, which are respectively obtained by the voice signal received by the microphone M_0, M_1, M_2, M_3. In this experiment, the Hanning window is selected for the window function, the frame length is chosen as 200, the frame shift size is 100, and the color is displayed. One of the speech waveforms corresponding to a set of speech signal waveforms is shown in the Fig.5.

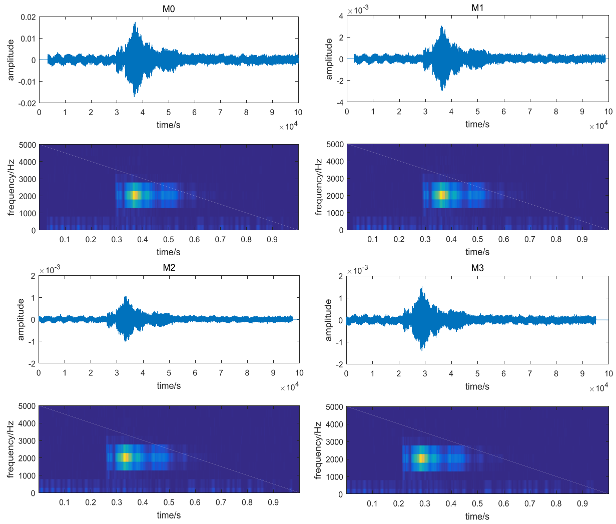

FIGURE 5. The corresponding waveform and language spectrogram of the four way microphone.

According to Figure 5, we can clearly see the waveform and spectral map corresponding to the four-way microphone, but the spectral map generated at each reference point corresponds to one data sample. In other words, the four sets of voice signals received by the four-way microphone are actually one sample. So we combine the four spectra into one, then we can get the speech signal received by the microphone M_0, M_1, M_2, M_3 at each reference point from the one spectra, the amplitude of the speech signal corresponds to the strength of the color in the spectrogram. As the picture shows:

The size of the picture is 875 × 656. In order to improve the training accuracy and facilitate the subsequent calculation, the white edges and coordinates around the language map are cropped out, and the size of the picture is adjusted to 100 × 100. Make a label for 10,000 spectrogram samples. One image corresponds to a label number. Divide the area to be located into nine blocks, as shown in Fig.7.

The database established in the experiment consists of 9 regional categories. There are 10000 samples in total, 1100 in each category. 90% of the samples of these spectrograms are used as training samples and 10% as test samples.

IV. APPLICATION OF CONVOLUTIONAL NEURAL NETWORKS IN LOCALIZATION OF SOUND SOURCE REGION

Due to the weight sharing characteristics of convolutional neural networks, the weight parameters of back- propagation errors that need to be trained are reduced, and the complexity of the network is reduced. The input of the network is a colorful image of 100 × 100 size with a sample size of 64.

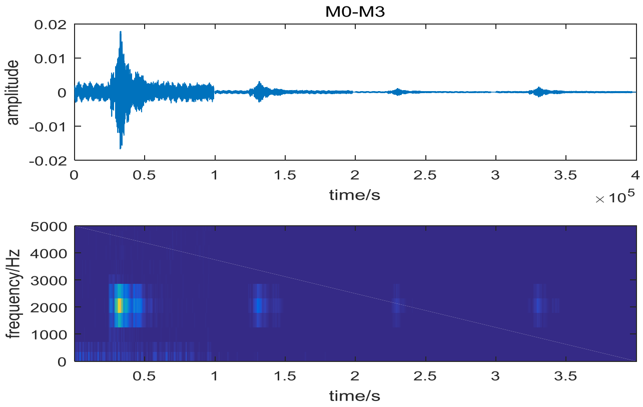

FIGURE 6. The corresponding waveform and language spectrogram of the four way microphone.

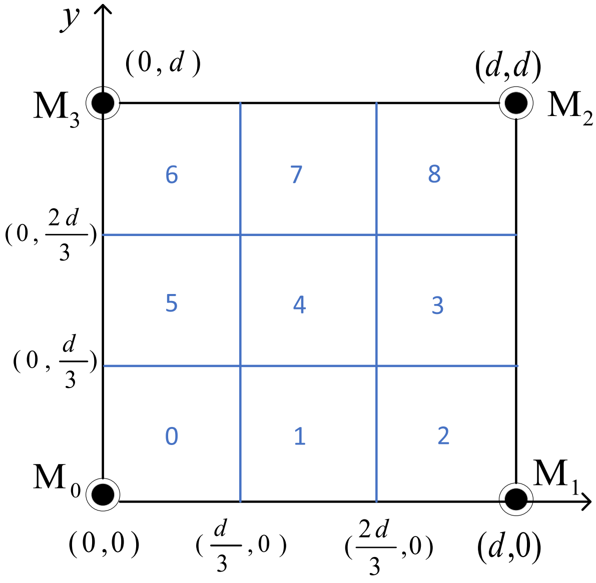

FIGURE 7. Localization partition map.

The network structure of the CNN for localization of the sound source are designed in this study is shown in Figure 8.

The CNN model of this experiment is composed of four convolution layers four pooled layers and three fully connected layers. The network structure is Conv1- (5 × 5,32) + P1 + Conv2(5 × 5,64) + P2 + Conv3(3 × 3,128) + P3 + Conv4(3 × 3,128) + P4 + FC1(1024) + FC2(512) +FC3 (9), where Conv represents a convolutional layer, P represents a pooled layer, and FC represents Fully connected layers, the size and number of convolution kernels are shown in parentheses respectively, and the number of neurons are shown in the brackets of the full connection layer. Here we analyze a convolutional layer, a pooling layer, and a full-connection layer.

The convolutional layer uses a training convolution to conduct the convolution operation of the input data, that is, feature extraction. Use the Relu activation function. Each plane of the convolution layer is determined

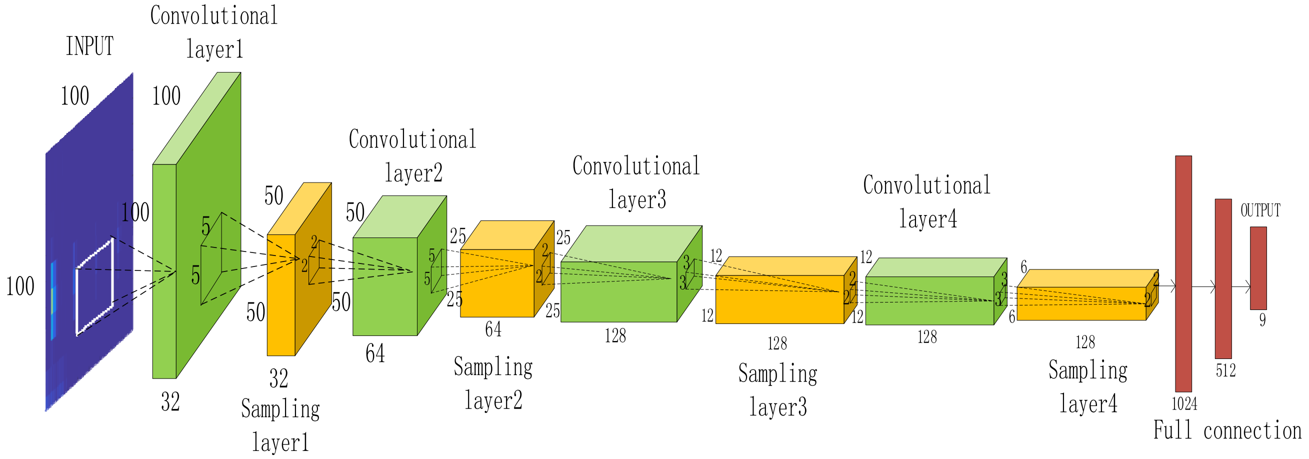

FIGURE 8. The network structure of CNN.

by formula (14):

$$X_j^l = f(\sum_{i \in M_j} X_j^{l-1} * k_{ij}^l + b_j^l) \tag{14}$$

Where M_j denotes the set of selected feature maps, l denotes the current layer number, f denotes an activation function, k_{ij}^l denotes a convolution kernel corresponding to different input feature maps, and b_j^l denotes an additive bias corresponding to the output feature map.

The resulting feature map is used as the input to the next pooling layer to reduce the dimension. Dimensionality reduction has three effects on the system: it makes the features more compact and highlights salient features. It reduces the training parameters of the system. The n-size pooling layer can reduce n^2-fold parameters and increase the system's robustness. Each plane is determined by formula (15):

$$X_j^l = f(\beta_j^l down(X_j^{l-1}) + b_j^l) \tag{15}$$

Where $down(.)$ denotes a down sampling function, l denotes the current layer number, f denotes an activation function, β_j^l is the multiplicative offset corresponding to the output feature map, and b_j^l denotes an additive bias corresponding to the output feature map.

The full-connected layer reduces the input two-dimensional feature matrix to a one-dimensional feature vector to facilitate the output layer for classification processing. The output layer is classified according to the one-dimensional vector of the output of the full connected layer above. This column uses the softmax cross-entropy loss.

The CNN localization process in the indoor sound source area includes the training process and the classification process. The training process includes the forward propagation process and the back propagation process. The classification process uses the parameter model obtained from the training for the test sample to perform the operation and obtain the localization result. The specific process is shown in Fig.9.

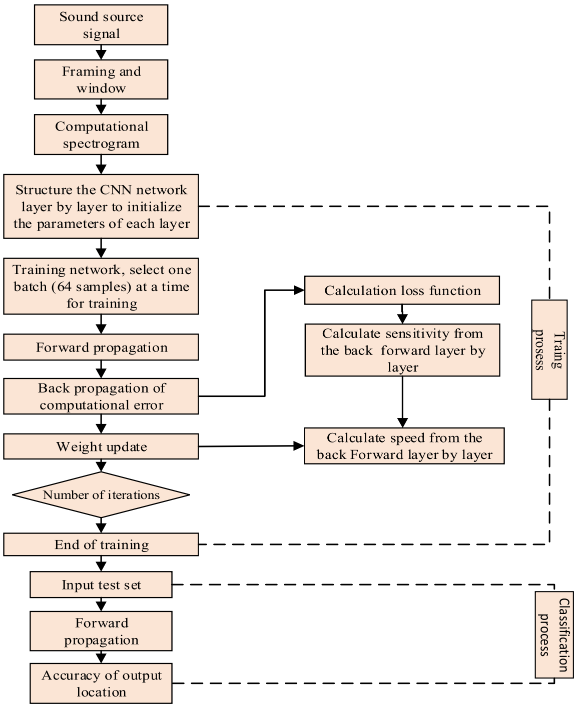
Sound source signal
Framing and window
Computational spectrogram
Structure the CNN network layer by layer to initialize the parameters of each layer
Training network, select one batch (64 samples) at a time for training
Forward propagation
Back propagation of computational error
Weight update
Number of iterations
End of training
Calculation loss function
Calculate sensitivity from the back forward layer by layer
Calculate speed from the back Forward layer by layer
Traing prosess
Input test set
Forward propagation
Accuracy of output location
Classification process

FIGURE 9. The algorithm flow of CNN in the localization of the indoor sound source region.

V. SIMULATION ANALYSIS

A. SIMULATION EXPERIMENTAL RESULTS

The CNN solve the weight through iterative operation. After multiple iterations, the ideal parameters are obtained. Since the noise in the actual environment is unpredictable, when acoustic features change, the validity of the collected training data varies. In this case, we need to ensure that the CNN is still accurate and robust even when the environment changes after training. So we collected five sets of training data, SNR = −5dB, 0dB 5dB, 10dB and 15dB. The same parameter configuration is adapted during training: the learning rate is 0.01, batch_size = 64. The experimental results of

TABLE 1. The experimental results with iterations at SNR = −5dB.

Number of iterations	Training accuracy/%	Testing accuracy/%	Training time/s	Testing time/s
1	80.3	79.1	15.9	9.3
50	90.1	89.9	260.3	9.6
100	93.2	92.8	779.4	9.8
200	97.3	97.1	1550.4	9.3
500	97.3	97.2	3976.2	9.1
1000	97.3	97.2	7720.8	9.5

TABLE 2. The experimental results with iterations at SNR = 0dB.

Number of iterations	Training accuracy/%	Testing accuracy/%	Training time/s	Testing time/s
1	82.3	81.1	15.5	9.0
50	91.4	90.6	262.4	9.6
100	94.3	93.6	780.3	9.6
200	97.9	97.7	1552.1	9.4
500	97.8	97.7	3970.2	9.2
1000	97.8	97.7	7731.2	9.6

TABLE 3. The experimental results with iterations at SNR = 5dB.

Number of iterations	Training accuracy/%	Testing accuracy/%	Training time/s	Testing time/s
1	85.2	83.1	15.1	9.6
50	95.4	93.9	259.3	9.2
100	98.9	97.8	774.8	9.9
200	99.2	98.1	1549.9	9.3
500	99.3	98.1	3984.1	9.0
1000	99.3	98.1	7718.3	9.4

TABLE 4. The experimental results with iterations at SNR = 10dB.

Number of iterations	Training accuracy/%	Testing accuracy/%	Training time/s	Testing time/s
1	81.2	85.4	15.2	8.9
50	91.0	94.8	264.3	9.4
100	98.9	98.4	777.3	9.3
200	99.1	98.4	1556.6	9.2
500	99.1	98.4	3985.3	9.9
1000	99.2	98.4	7721.7	9.0

TABLE 5. The experimental results with iterations at SNR = 15dB.

Number of iterations	Training accuracy/%	Testing accuracy/%	Training time/s	Testing time/s
1	86.2	85.6	15.3	9.1
50	97.1	96.2	266.5	9.3
100	99.0	98.5	778.8	9.8
200	99.2	98.5	1560.1	9.6
500	99.2	98.5	3987.8	9.3
1000	99.2	98.5	7731.4	9.7

different iteration times under different signal-to-noise ratios are shown in Table 1-5:

According to Table 1-5, in the case of a small number of iterations, network learning is not sufficient, and the training model is not ideal, so the classification effect of training is poor. As the number of iterations increases, the network parameters are continuously optimized, the classification accuracy rate increases, and the number of iterations tends to be stable when it reaches 200. Moreover, it shows that the CNN have strong robustness by comparing the experimental results under different signal-to-noise ratios, and its final average localization accuracy is about 98%.

Tensorboard is a visual tool embedded in Tensorflow. It can display the various drawing data in process by reading the event log. Since the accuracy of convolutional neural networks training results are nearly the same under different SNR, this paper uses the Tensorboard tool to visualize the results of 1000 training and testing when SNR = 10, as shown in Fig.10 and Fig.11.

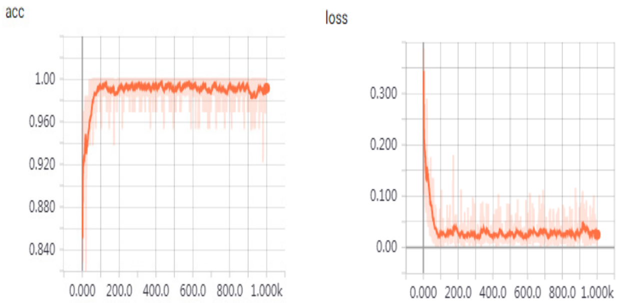

FIGURE 10. The accuracy and loss function on the training set.

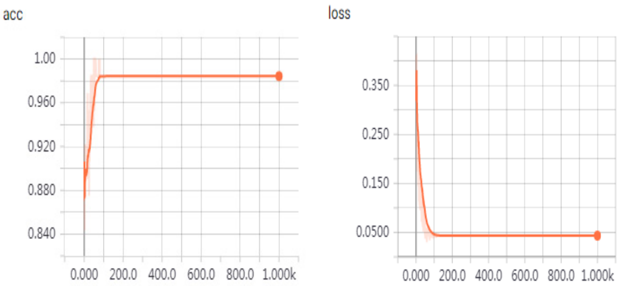

FIGURE 11. The accuracy and loss function on the testing set.

It records the classification accuracy of the model on the training set and the test set after each iteration and the loss function of the model. It can be seen from the figure that the loss function decreases with iterations, and the classification accuracy increases with iterations. In the end, both changes have stabilized. In the course of 10000 iterations, the accuracy of the model on the training data reaches a maximum of 1, and the accuracy on the test data eventually tends to be about 98.4%.

B. COMPARATIVE EXPERIMENT

In order to verify the performance of CNN in sound source localization, this paper compares it with KNN (K-Nearest Neighbor), BP (Back Propagation) neural network and SVM (Support Vector Machine), using sklearn to test these three machine learning methods [54]. For KNN, we use KNeighborsClassifier; for SVM, we use SVC; for BP neural networks, we use MLPClassifier. Then, the localization dataset of the spectrogram sample is trained and tested at a signal-to-noise ratio of -5dB, 0dB, 5dB, 10dB, and 15dB respectively. The experimental results are shown in Fig.12-Fig.16.

Since KNN has no parameter training process, it makes direct classification decisions based on the distribution of training data. Therefore, it is impossible to reproduce the relationship between the accuracy of the KNN method and the time in the line graph. We use Table 6 to show the experimental results based on KNN.

TABLE 6. The experimental results on KNN method.

SNR/dB	Accuracy/%	Time/s
-5	54.6	570.4
0	70.1	561.2
5	89.7	553.6
10	90.3	560.8
15	90.6	563.0

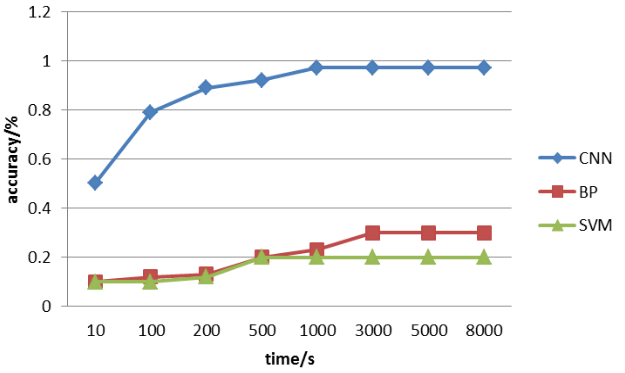

FIGURE 12. Comparison of localization accuracy at SNR = −5.

The experimental results show that the accuracy of indoor sound source localization based on convolutional neural network is significantly higher than that of other three machine learning methods, and the convergence speed of the network is also faster. According to Figure 12 - Figure 16, we can get the training duration of CNN algorithm to be stable at about 200s and achieve satisfactory training accuracy. The training time required for KNN is about 560s, the consumption of BP and SVM take about 3000s, and so from the perspective of computational complexity, CNN is also significantly better than the other three machine learning methods. For traditional machine learning methods such as KNN, BP, SVM, etc., complex feature engineering is usually required. Firstly, perform the deep exploratory data analysis on the data set, then a simple process of reducing the dimension. Finally, carefully chose the best features to pass to

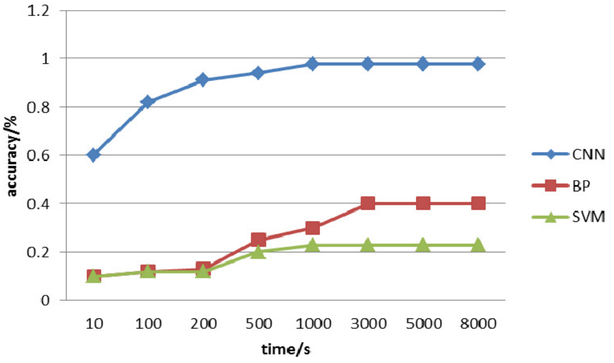

FIGURE 13. Comparison of localization accuracy at SNR = 0.

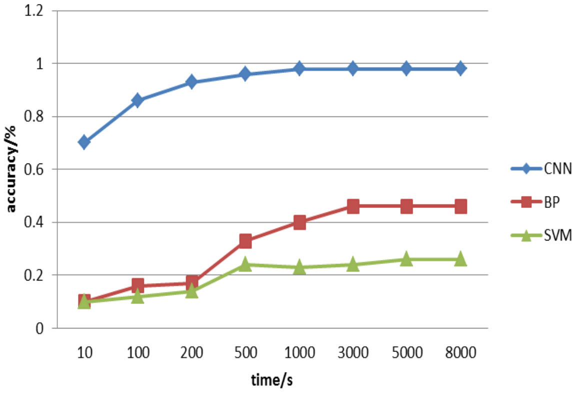

FIGURE 14. Comparison of localization accuracy at SNR = 5.

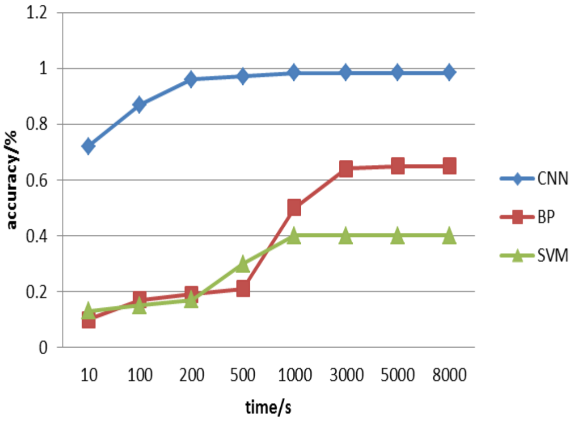

FIGURE 15. Comparison of localization accuracy at SNR = 10.

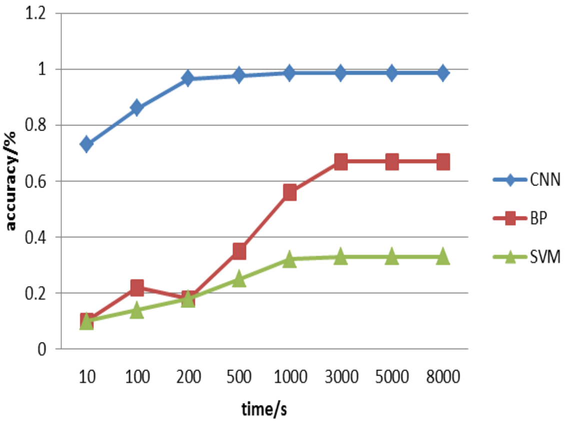

FIGURE 16. Comparison of localization accuracy at SNR = 15.

the machine learning algorithm. When using CNN, we don't need to separate the two processes of feature extraction and classification training. Because the most effective features are automatically extracted during training, especially the images of multi-dimensional input vectors which can be directly input into the network. The complexity of data reconstruction during feature extraction and classification is avoided. Therefore, the CNN algorithm is significantly better than the other three machine learning methods.

VI. CONCLUSIONS

In this paper, a method of indoor sound source localization based on convolutional neural network is designed. We bring up the idea to convert the sound source signal into a spectral map and input it into the convolutional neural network to realize the regional localization of the single sound source for the first time. The method proposed in this paper solves the problems of low localization accuracy, high dependence on model and high computational complexity in the traditional unstructured space. The simulation experiment proves that the convolutional neural network has good robustness and good generalization ability for speech signals with different SNR. Currently our research focuses on the application of convolutional neural networks in sound source localization, and the accuracy of the test also verifies the accuracy and effectiveness of the algorithm. As future work, the first thing we are going to do is to further verify the adaptability of the convolutional neural network algorithm in indoor sound source localization for different experimental environments and experimental conditions. Secondly, this experiment only involves the use of a single sound source for indoor area positioning. For multiple sound source localization, it still needs further study. Finally, we may analyze CNN of different structures and methods combined with other neural networks to achieve higher accuracy and faster convergence.

REFERENCES

[1] J. Cao, J. Liu, J. Wang, and X. Lai, "Acoustic vector sensor: Reviews and future perspectives," *IET Signal Process.*, vol. 11, no. 1, pp. 1–9, 2016.

[2] H. C. Song and C. Cho, "Array invariant-based source localization in shallow water using a sparse vertical array," *J. Acoust. Soc. Amer.*, vol. 141, no. 1, p. 183, 2017.

[3] L. Wei, M. Li, D. Yang, F. Niu, and W. Zeng, "Reconstruction of sound source signal by analytical passive TR in the environment with airflow," *J. Sound Vibrat.*, vol. 392, pp. 77–90, Mar. 2017.

[4] S. He and H. Chen, "Closed-form DOA estimation using first-order differential microphone arrays via joint temporal-spectral-spatial processing," *IEEE Sensors J.*, vol. 17, no. 4, pp. 1046–1060, Feb. 2017.

[5] C. H. Lee, "Location-aware speakers for the virtual reality environments," *IEEE Access*, vol. 5, pp. 2636–2640, 2017.

[6] K. Hoshiba *et al.*, "Design of UAV-embedded microphone array system for sound source localization in outdoor environments," *Sensors*, vol. 17, no. 11, p. 2535, 2017.

[7] J. Even, J. Furrer, Y. Morales, C. T. Ishi, and N. Hagita, "Probabilistic 3-D mapping of sound-emitting structures based on acoustic ray casting," *IEEE Trans. Robot.*, vol. 33, no. 2, pp. 333–345, Apr. 2017.

[8] D. A. Hambrook, M. Ilievski, M. Mosadeghzad, and M. Tata, "A Bayesian computational basis for auditory selective attention using head rotation and the interaural time-difference cue," *PLoS ONE*, vol. 12, no. 10, p. e0186104, 2017.

[9] D. Salvati, C. Drioli, and G. L. Foresti, "Incoherent frequency fusion for broadband steered response power algorithms in noisy environments," *IEEE Signal Process. Lett.*, vol. 21, no. 5, pp. 581–585, May 2014.

[10] R.-H. Zhang, Z.-C. He, H.-W. Wang, F. You, and K.-N. Li, "Study on self-tuning tyre friction control for developing main-servo loop integrated chassis control system," *IEEE Access*, vol. 5, pp. 6649–6660, 2017.

[11] X. Sun, H. Zhang, W. Meng, R. Zhang, K. Li, and T. Peng, "Primary resonance analysis and vibration suppression for the harmonically excited nonlinear suspension system using a pair of symmetric viscoelastic buffers," *Nonlinear Dyn.*, vol. 94, no. 2, pp. 1243–1265, 2018, doi: 10.1007/s11071-018-4421-9.

[12] J. Li, Q. H. Lin, and K. Wang, "Performance analysis for focused beamformers in passive underwater acoustic localization," *IEEE Access*, vol. 6, pp. 18200–18208, 2018.

[13] H. Xiong, X. Zhu, and R. Zhang, "Energy recovery strategy numerical simulation for dual axle drive pure electric vehicle based on motor loss model and big data calculation," *Complexity*, vol. 2018, Aug. 2018, Art. no. 4071743, doi: 10.1155/2018/4071743.

[14] S. Tervo and A. Politis, "Direction of arrival estimation of reflections from room impulse responses using a spherical microphone array," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 23, no. 10, pp. 1539–1551, Oct. 2015.

[15] R. Porteous, Z. Prime, C. J. Doolan, D. J. Moreau, and V. Valeau, "Three-dimensional beamforming of dipolar aeroacoustic sources," *J. Sound Vibrat.*, vol. 355, pp. 117–134, Oct. 2015.

[16] L. Wang, T.-K. Hon, J. D. Reiss, and A. Cavallaro, "Self-localization of ad-hoc arrays using time difference of arrivals," *IEEE Trans. Signal Process.*, vol. 64, no. 4, pp. 1018–1033, Feb. 2016.

[17] X. Cui, K. Yu, and S. Lu, "Direction finding for transient acoustic source based on biased TDOA measurement," *IEEE Trans. Instrum. Meas.*, vol. 65, no. 11, pp. 2442–2453, Nov. 2016.

[18] B. Zhang, Y. Hu, H. Wang, and Z. Zhuang, "Underwater source localization using TDOA and FDOA measurements with unknown propagation speed and sensor parameter errors," *IEEE Access*, vol. 6, pp. 36645–36661, 2018.

[19] J. H. Chang and C. H. Jeong, "A measure based on beamforming power for evaluation of sound field reproduction performance," *Appl. Sci.*, vol. 3, no. 3, p. 249, 2017.

[20] M. Zhu, H. Yao, X. Wu, Z. Lu, X. Zhu, and Q. Huang, "Gaussian filter for TDOA based sound source localization in multimedia surveillance," *Multimedia Tools Appl.*, vol. 77, no. 3, pp. 3369–3385, 2017.

[21] Z. Liu, R.-L. Chen, P.-X. Teng, and Y.-C. Yang, "Sound source localization system based on planar microphone array," *Tech. Acoust.*, vol. 30, no. 2, pp. 123–128, 2011.

[22] L. Kumar, A. Tripathy, and R. M. Hegde, "Robust multi-source localization over planar arrays using MUSIC-group delay spectrogram," *IEEE Trans. Signal Process.*, vol. 62, no. 17, pp. 4627–4636, Sep. 2014.

[23] C.-D. Kee, G.-H. Kim, and T.-J. Lee, "Real-time sound localization system for reverberant and noisy environment," *J. Korean Soc. Aeronaut. Space Sci.*, vol. 38, no. 3, pp. 258–263, 2010.

[24] X. Xiao, S. Zhao, X. Zhong, D. L. Jones, E. S. Chng, and H. Li, "A learning-based approach to direction of arrival estimation in noisy and reverberant environments," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Apr. 2015, pp. 2814–2818.

[25] M. Kovandžić, V. Nikolić, A. Al-Noori, I. Cirić, and M. Simonović, "Near field acoustic localization under unfavorable conditions using feedforward neural network for processing time difference of arrival," *Expert Syst. Appl. Int. J.*, vol. 71, pp. 138–146, Apr. 2016.

[26] R. Lefort, G. Real, and A. Drémeau, "Direct regressions for underwater acoustic source localization in fluctuating oceans," *Appl. Acoust.*, vol. 116, pp. 303–310, Jan. 2017.

[27] Y. Sun, J. Chen, C. Yuen, and S. Rahardja, "Indoor sound source localization with probabilistic neural network," *IEEE Trans. Ind. Electron.*, vol. 65, no. 8, pp. 6403–6413, Aug. 2017.

[28] T. Fu, Z. Zhang, Y. Liu, and J. Leng, "Development of an artificial neural network for source localization using a fiber optic acoustic emission sensor array," *Struct. Health Monitor.*, vol. 14, no. 2, pp. 168–177, 2015.

[29] J.-S. Hu, C.-C. Cheng, and W.-H. Liu, "Robust speaker's location detection in a vehicle environment using GMM models," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 36, no. 2, pp. 403–412, Apr. 2006.

[30] H. Chen and W. Ser, "Acoustic source localization using LS-SVMs without calibration of microphone arrays," in *Proc. IEEE Int. Symp. Circuits, Syst. (ISCAS)*, May 2009, pp. 1863–1866.

[31] D. Salvati, C. Drioli, and G. L. Foresti, "A weighted MVDR beamformer based on SVM learning for sound source localization," *Pattern Recognit. Lett.*, vol. 84, pp. 15–21, Dec. 2016.

[32] N. Ma, T. May, and G. J. Brown, "Exploiting deep neural networks and head movements for robust binaural localization of multiple sources in reverberant environments," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 25, no. 12, pp. 2444–2453, Dec. 2017.

[33] F. Vesperini, V. Vecchiotti, E. Principi, S. Squartini, and F. Piazza, "Localizing speakers in multiple rooms by using deep neural networks," *Comput. Speech, Lang.*, vol. 49, pp. 83–106, May 2018.

[34] X. Zhang, Y. Qiao, F. Meng, C. Fan, and M. Zhang, "Identification of maize leaf diseases using improved deep convolutional neural networks," *IEEE Access*, vol. 6, pp. 30370–30377, 2018.

[35] S. J. Lee, T. Chen, L. Yu, and C.-H. Lai, "Image classification based on the boost convolutional neural network," *IEEE Access*, vol. 6, pp. 12755–12768, 2018.

[36] C. Gu, H. Du, S. Cai, and X. Chen, "Joint multiple image parametric transformation estimation via convolutional neural networks," *IEEE Access*, vol. 6, pp. 18822–18831, 2018.

[37] W. Quan, K. Wang, D.-M. Yan, and X. Zhang, "Distinguishing between natural and computer-generated images using convolutional neural networks," *IEEE Trans. Inf. Forensics Security*, vol. 13, no. 11, pp. 2772–2787, Nov. 2018.

[38] M. Aykanat, Ö. Kılıç, B. Kurt, and S. Saryal, "Classification of lung sounds using convolutional neural networks," *EURASIP J. Image Video Process.*, vol. 2017, no. 1, p. 65, 2017.

[39] C. Fan, Y. Zhang, L. Feng, and Q. Jiang, "No reference image quality assessment based on multi-expert convolutional neural networks," *IEEE Access*, vol. 6, pp. 8934–8943, 2018.

[40] K. Asahi and A. Ogawa, "Reduction of noise in speech signals through image processing using the spectrogram," *IEEJ Trans. Electron. Inf. Syst.*, vol. 126, no. 12, pp. 1483–1489, 2006.

[41] P. K. Ajmera, D. V. Jadhav, and R. S. Holambe, "Text-independent speaker identification using Radon and discrete cosine transforms based features from speech spectrogram," *Pattern Recognit.*, vol. 44, no. 10, pp. 2749–2759, 2011.

[42] C. Zheng, Z.-H. Tan, R. Peng, and X. Li, "Guided spectrogram filtering for speech dereverberation," *Appl. Acoust.*, vol. 134, no. 5, pp. 154–159, 2018.

[43] Y. Li and H. Chen, "Reverberation robust feature extraction for sound source localization using a small-sized microphone array," *IEEE Sensors J.*, vol. 17, no. 19, pp. 6331–6339, Oct. 2017.

[44] L. J. Ziomek, *Fundamentals of Acoustic Field Theory and Space-Time Signal Processing*. Boca Raton, FL, USA: CRC Press, 1995.

[45] L. Wang, T.-K. Hon, J. D. Reiss, and A. Cavallaro, "An iterative approach to source counting and localization using two distant microphones," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 24, no. 6, pp. 1079–1093, Jun. 2016.

[46] Y. Sun, J. Chen, C. Yuen, and S. Rahardja, "Indoor sound source localization with probabilistic neural network," *IEEE Trans. Ind. Electron.*, vol. 65, no. 8, pp. 6403–6413, Aug. 2017.

[47] C. Zhang, C. Yu, and J. H. L. Hansen, "An investigation of deep learning frameworks for speaker verification anti-spoofing," *IEEE J. Sel. Topics Signal Process.*, vol. 11, no. 4, pp. 684–694, Jun. 2017.

[48] J. Cadore, F. J. Valverde-Albacete, A. Gallardo-Antolín, and C. Peláez-Moreno, "Auditory-inspired morphological processing of speech spectrograms: Applications in automatic speech recognition and speech enhancement," *Cognit. Comput.*, vol. 5, no. 4, pp. 426–441, 2013.

[49] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digit. Signal Process.*, vol. 10, nos. 1–3, pp. 19–41, 2000.

[50] W.-K. Lu and Q. Zhang, "Deconvolutive short-time Fourier transform spectrogram," *IEEE Signal Process. Lett.*, vol. 16, no. 7, pp. 576–579, Jul. 2009.

[51] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, p. 436, 2015.

[52] R. Kaune, J. Hörst, and W. Koch, "Accuracy analysis for TDOA localization in sensor networks," in *Proc. Int. Conf. Inf. Fusion*, Jul. 2011, pp. 1–8.

[53] X. Sheng and Y.-H. Hu, "Maximum likelihood multiple-source localization using acoustic energy measurements with wireless sensor networks," *IEEE Trans. Signal Process.*, vol. 53, no. 1, pp. 44–53, Jan. 2005.

[54] I. Ahmed, P. Witbooi, and A. Christoffels, "Prediction of human-*Bacillus anthracis* protein_protein interactions using multi-layer neural network," *Bioinformatics*, pp. 1–6, Jun 2018, doi: 10.1093/bioinformatics/bty504.

XIAOMENG ZHANG received the B.Eng. degree from the City College, Hebei University of Technology, Tianjin, China, in 2016, where she is currently pursuing the M.S. degree with the School of Artificial Intelligence. Her current research interests include sound source localization and machine learning.

HAO SUN received the Ph.D. degree from the Hebei University of Technology in 2012. He is currently an Associate Professor with the Hebei University of Technology. He has published more than 30 papers and has obtained more than 10 patents and software copyrights. His main research interests include auditory robot and intelligent robotics. Besides, he has participated in the formulation of one national standard. He is currently a Youth Member of the Rehabilitation Engineering Branch of the China Biomedical Engineering Society and a Youth Committee Member of the Rehabilitation Engineering Committee of the China Rehabilitation Aids Association.

SHUOPENG WANG received the M.S. degree from the Guilin University of Electronic Technology in 2011. He is currently pursuing the Ph.D. degree with the School of Artificial Intelligence, Hebei University of Technology, Tianjin, China. His current research interests are mobile robot and sound source localization system.

JING XU received the B.Eng. degree from the Tianjin University of Technology and Education in 2016. She is currently pursuing the M.S. degree with the School of Artificial Intelligence, Hebei University of Technology, Tianjin, China. Her current research interests include mobile robot and sound source localization system.

• • •