

Research Article

Mechanical Fault Sound Source Localization Estimation in a Multisource Strong Reverberation Environment

Yaohua Deng (), Xiali Liu (), Zilin Zhang (), and Daolong Zeng ()

School of Electromechanical Engineering, Guangdong University of Technology, Guangzhou 510006, China

Correspondence should be addressed to Xiali Liu; lxl@gdut.edu.cn

Received 31 August 2023; Revised 2 January 2024; Accepted 2 February 2024; Published 24 February 2024

Academic Editor: Felix Albu

Copyright © 2024 Yaohua Deng et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Aiming at the sound source localization of mechanical faults in a strong reverberation scenario with multiple sound sources, this paper investigates a mechanical fault source localization method using the U-net deep convolutional neural network. The method utilizes the SRP-PHAT algorithm to calculate the response power spectra of the collected multichannel fault signals. Through the utilization of the U-net neural network, the response power spectra containing spurious peaks are transformed into "clean" estimated source distribution maps. By employing interpolation search, the estimated source distribution maps are processed to obtain location estimations for multiple fault sources. To validate the effectiveness of the proposed method, this paper constructs an experimental dataset using mechanical fault data from electromechanical equipment relays and conducts sound source localization experiments. The experimental results show that the U-net network under 0.2 s/0.5 s/0.7 s reverberation time can effectively eliminate spurious peak interference in the response power spectrum. As the signal-to-noise ratio decreases, it can still distinguish the sound sources with a distance of 0.2 m. In the context of multifault source localization, the method is capable of simultaneously locating the positions of four fault sources, with an average localization error of less than 0.02 m. The method in this paper effectively eliminates spurious peaks in the response power spectra under conditions of multisource strong reverberation. It accurately locates multiple mechanical fault sources, thereby significantly enhancing the efficiency of mechanical fault detection.

1. Introduction

Over the past decade, researchers have been actively exploring the use of acoustic features, classification, and clustering algorithms to predict or detect the state of machinery. They have also leveraged microphone arrays to capture spatial information for fault source localization [1, 2]. Many researchers have developed various sound source locators, such as Schober et al., who developed a functional sound source locator based on stochastic computing (SC) [3]. However, due to various factors such as noise interference, reverberation, the nature of source signals, and the number of sources, the practical efficacy of source localization systems is influenced. As a result, the application of source localization technology in mechanical fault detection still faces numerous challenges.

Traditional source localization algorithms are primarily built upon signal models and array signal processing techniques and are roughly divided into three categories: time difference of arrival- (TDOA-) based source localization algorithms [4, 5], signal subspace-based source localization algorithms [6, 7], and beamforming-based source localization algorithms [8, 9]. The TDOA-based source localization algorithms offer a low computational complexity and ease of implementation, rendering it highly practical for scenarios demanding real-time performance. However, it is sensitive to array hardware errors and exhibits reduced resilience to noise and reverberation. The signal subspacebased source localization algorithms allow for achieving ultra-high-resolution multisource localization. Nevertheless, they also exhibit certain limitations. (1) They require prior knowledge of the number of sources, making them unsuitable for scenarios with an unknown number of sources.

(2) They impose several restrictions on sources and noise, demanding that sources be uncorrelated and noise adhere to Gaussian signal assumptions. The beamforming-based source localization algorithms utilize synthesized beams to visualize the target region. The direction with the highest response power corresponds to the direction of the source. They are more robust than the TDOA-based source localization algorithm, offering increased resilience. DiBiase et al. [10] combined the characteristics of TDOA-based and beamforming-based source localization algorithms and proposed a method called steered response power-phase transform (SRP-PHAT). This approach involves applying the PHAT weighting to the sound signals received by each microphone, forming a directional beam. Subsequently, the beam scans various spatial search grids and computes response power spectra. The location of the spectral peak is used to estimate the source's position. Compared to adaptive beamforming, SRP-PHAT does not require prior knowledge of source and noise, making it one of the mainstream source localization algorithms in recent years. Shi et al. [11] proposed a low-frequency noise sources localization method based on the virtual SMA extrapolation method and resolved the localization problem of a small-aperture SMA with lowfrequency noise sources.

In addition to the array signal techniques mentioned above for source localization, many researchers have introduced neural networks into sound source localization, such as CNN [12], CRNN [6], and AE [13]. These approaches process the signals collected by microphone arrays through feature extraction modules to obtain input features. These are then fed into neural networks for estimating source positions or arrival directions. For example, Chakrabarty and Habets [14] proposed using the phase spectrum obtained from multichannel data after STFT transformation as the input to a CNN network and learning the phase relationship between neighboring channels through three consecutive convolutional layers to achieve sound source azimuth estimation, which shows excellent robustness in reverberant environments. Salvati et al. [15] proposed inputting weights of narrowband response power components from each beamforming into a neural network. Through a CNN, they automatically learned the weighted vectors of these components, achieving higher-precision source localization. Adavanne et al. [16], on the other hand, utilized a CRNN network for multitask configuration. This method applies to various microphone array structures and demonstrates strong robustness in scenarios involving reverberation and low signal-to-noise ratios. Senocak et al. [17] proposed a cross-modal alignment task as a joint task with sound source localization to better learn the interaction between audio and visual modalities. This approach aims to achieve high localization performance through robust crossmodal semantic understanding. Park et al. [18] proposed to localize sound sources in visual scenes with a self-supervised approach. Using a less strict decision boundary in contrastive learning can alleviate the effect of noisy correspondences in sound source localization.

In summary, source localization methods incorporating neural networks can effectively address the performance degradation issues in source localization systems under complex scenarios involving reverberation, noise, and multiple sources. To address the challenge of spurious peaks in the response power spectra of the SRP-PHAT algorithm in multisource strong reverberation environments, it is worth noting that source localization is akin to locating the direction of sources within audio signals. The U-net deep neural network has demonstrated remarkable performance in object localization within images [19]. Thus, we propose to introduce the U-net network. Firstly, the SRP-PHAT algorithm is utilized to generate response power spectra. Subsequently, the U-net network is employed for multifault sound source localization.

This paper is organized as follows. Section 1 discusses the application of sound source localization technology in mechanical fault detection and reviews the research progress of various methods. Section 2 provides a detailed introduction to multisource localization method we proposed, which combines the U-net and SRP-PHAT algorithm. Section 3 focuses on the training and experimental analysis of the U-net network. Finally, Section 4 summarizes the key findings and conclusions of the paper.

2. U-Net-Based Multifault Sound Source Localization Method

The multifault sound source localization method proposed in this paper combines U-net with the SRP-PHAT algorithm. It treats sound source localization as a pixel-level classification task, where each pixel corresponds to a discrete spatial grid region, and the size of the pixel value represents the response power size corresponding to that grid, based on which the power size can be used to determine the actual sound source. In detail, the SRP-PHAT algorithm is used as a feature extractor of the spatial information of sound sources. The response power spectrum with pseudopeaks is output as a feature map, which is converted into a "clean" sound source distribution map by a series of convolution, pooling, and upsampling operations through the U-net network, combined with an interpolation search method to realize multisource localization.

The overall framework is shown in Figure 1, which is divided into two stages: U-net network training and sound source localization. In the training stage, for each sample in the dataset, the multichannel audio data are processed through the SRP-PHAT algorithm to compute the response power spectrum, which serves as the input feature. Simultaneously, the source distribution map is calculated using the source positions and sound power levels, and this map is used as the ground truth label for the samples. The U-net network is then trained. In the sound source localization stage, the array-collected signals undergo the SRP-PHAT algorithm to calculate the response power spectrum, which is then fed into the pretrained U-net model. This model predicts the source distribution map. By employing an interpolation search technique on this map, the specific positions of multiple sources are determined, thus completing the localization process.



FIGURE 1: Multifault sound source localization using U-net and SRP-PHAT.

2.1. Controlled Response Power Spectrum Calculation Based on SRP-PHAT. The process of calculating the response power spectrum using SRP-PHAT is depicted in Figure 2. The search space is divided into discrete spatial grids based on the desired resolution, and all these grids are potential candidate positions for sources. Firstly, each microphone's received audio signal is weighted differently to form a directional beam. Secondly, this beam is utilized to scan each spatial grid in the search space, and the controllable response power of that grid is computed. This process results in the response power spectrum of the source plane.

Assuming there are M microphones in the microphone array, let us denote $R_{mn}(\tau)$ as the phase-transform weighted generalized inter-correlation function between the signals received by microphones m and n. Here, $\Delta \tau_{mn}(l)$ represents the time delay difference, also known as the propagation time delay, of the sound signal from grid l to microphones mand n. This delay accounts for the microphone array's steering delay at grid l. The response power at grid l can be represented as follows:

$$P(l) = \sum_{m=1}^{M} \sum_{n=m+1}^{M} R_{mn} (\Delta \tau_{mn}(l)).$$
(1)

Compared to the conventional controllable response power-based source localization algorithms, the SRP-PHAT algorithm introduces phase-transformed weighting to the generalized cross-correlation function to calculate controllable response power. This phase transformation weighting removes the amplitude information from the cross-power spectrum, retaining only the phase information. This approach can weaken the irrelevant peaks in the generalized cross-correlation function, leading to sharper spectral peaks. Consequently, it reduces the sensitivity of the SRP-PHAT localization algorithm to noise and reverberation components.

Let $x_m(t)$ and $x_n(t)$ represent the received signals from microphones m and n. Similarly, let $X_m(\omega)$ and $X_n(\omega)$ denote the Fourier transforms of $x_m(t)$ and $x_n(t)$, and $(\cdot)^*$ indicates the complex conjugate transpose. With these definitions in mind, the phase-transformed weighted generalized cross-correlation function can be represented as

$$R_{mn}(\tau) = \int_{-\infty}^{+\infty} \frac{X_m(\omega) X_n^*(\omega) e^{j\omega\tau}}{\left|X_m(\omega) X_n^*(\omega)\right|} \mathrm{d}\omega.$$
(2)

By substituting (2) into (1), the controlled response power spectrum of the SRP-PHAT algorithm can be obtained as

$$P(l) = \sum_{m=1}^{M} \sum_{n=m+1}^{M} \int_{-\infty}^{+\infty} \frac{X_m(\omega) X_n^*(\omega) e^{j\omega\Delta\tau_{mn}(l)}}{|X_m(\omega) X_n^*(\omega)|} d\omega.$$
(3)

In the resulting response power spectrum, each grid point's value corresponds to the microphone array's response power at that location. However, the response power spectrum often contains numerous false peaks due to reverberation and noise. When the number of sound sources is unknown, directly performing peak searching on the



FIGURE 2: SRP-PHAT scanning process.

response power spectrum to locate sources might result in misjudgements of the source count due to the influence of reverberation and noise, leading to inaccurate localization results. In addition, the presence of noise also causes the main flap of the source to widen, which may result in aliasing when the source positions are close together, making it difficult to distinguish the positions of two neighboring sources.

2.2. Construction of U-Net Network Architecture. The structure of the U-net network constructed in sound source localization is shown in Figure 3. The U-net neural network is designed with an encoder part comprising four power spectrum feature extraction units, progressively extracting features from the response power spectrum reducing its spatial dimension. The decoder part of the network consists of four source distribution map restoration units, systematically reconstructing the source distribution maps. To prevent potential overfitting caused by excessive skip connections during training, which might lead to the network becoming overly sensitive to false peaks and noise in the response power spectrum, this paper does not connect all the corresponding layers of the encoder network and the decoder network as in the original U-net network. Still, it only connects the feature reduction units 1 and 2 of the decoder network with the corresponding layers of the encoder network.

2.2.1. Design of Response Power Spectrum Feature Extraction Unit. In the encoder part of the U-net network, the input SRP-PHAT response power spectrum is initially subjected to a series of convolutional layers for feature extraction and pooling layers for downsampling. Hierarchical computations yield feature maps composed of multiple channels. The structure of the response power spectrum feature extraction unit is shown in Figure 4. It commences with two dilated convolutional layers for feature extraction, followed by a batch normalization layer and a ReLU activation function layer after each convolutional layer. Finally, a max-pooling layer is employed for downsampling. Each feature extraction unit consists of two dilated convolutional layers for feature extraction, followed by a max-pooling layer for downsampling. Each feature restoration unit includes a transpose convolutional layer for upsampling and two regular convolutional layers for feature fusion. Skip connections are incorporated in feature restoration units 1 and 2. These connections involve channelwise concatenation with the corresponding layers in the encoder, enhancing the ability to restore fine-grained details of the source distribution maps.

To effectively capture spatial structural information within the response power spectrum, the convolutional process in the encoder network of the U-net employs dilated convolution kernels of size 3×3 with a dilation rate K = 2. This choice allows the convolution operation to extend beyond neighboring elements, covering more considerable distances. Figure 5 illustrates convolution operations with different dilation rates. The dark grid represents the distribution of the dilated convolution kernel. Convolution with a dilation rate of K = 1 is equivalent to standard convolution. However, using a dilation rate K > 1 creates holes in the input image, allowing the convolution kernel to capture a broader receptive field. This process creates "holes" or "dilation" in the input image, enabling the convolutional kernel to encompass a broader receptive field.

As shown in Figure 3, the max-pooling layer is used to reduce the resolution of the feature map. With a pooling size of 2×2 , the input response power spectrum size is reduced by half with each passing feature extraction unit resolution. When the input response power spectrum is initially sized 100×100 , after passing through four feature extraction units, its size decreases to 6×6 , while the number of channels increases to 512.

2.2.2. Design of Source Distribution Map Feature Restoration Unit. In the decoder part of the U-net network, the feature maps obtained after feature extraction and dimensionality reduction through the encoder network are gradually restored into source distribution maps via a series of upsampling layers and convolutional layers. In addition,



FIGURE 3: The U-net network architecture.



FIGURE 4: The structure of the feature extraction unit.



FIGURE 5: Dilated convolution.

skip connection operations are incorporated into the decoder network's first and second feature restoration units, as shown in Figure 6. The process starts with a transpose convolutional layer for upsampling. After upsampling, the feature map is connected to the corresponding layer in the encoder through skip connections, and then two consecutive convolutional layers are used for feature fusion.

Skip connections can encourage the decoder network to reuse the high-level contextual information from the input response power spectrum and better restore the details in the source distribution maps but also alleviate the vanishing gradient problem commonly encountered in deep neural networks. This makes it easier for gradients to propagate through the network. Therefore, skip connections are used in the decoder part to stitch all channels of the shallow characteristics of the encoder and all channels of the decoder corresponding to the network layer to improve the ability of the U-net network to restore the sound source distribution maps.

2.3. Source Localization Based on Source Distribution Map Interpolation Search. The sound source distribution map reflects the location of multiple sound sources and sound power. The sound source distribution map of each grid sound power is inversely proportional to the distance from the grid center point to the sound source. In the presence of multiple sources on the plane, the power level in each grid results from the summation of the values contributed by each source in that grid. During the U-net network training, it is necessary to construct the corresponding source distribution maps as labels for training samples based on the source positions and power information. The construction process of the source distribution map is shown in Figure 7. Firstly, according to the size of the response power spectrum, build a distribution map containing L grids, calculate the sound power of each grid in turn, and build a sound source distribution map according to the position of the sound source and sound power information.

Using $r_{\text{grid}}^{l} = [x_{\text{grid}}^{l}, y_{\text{grid}}^{l}]$ to represent the position of the center point of the *l*th grid, $r_{s}^{m} = [x_{s}^{m}, y_{s}^{m}]$ and q_{s}^{m} to represent the position and power of the *m*th source in *M* sources, R(l,m) to represent the distance between the *l*th grid center point and the *m*th source, and $\zeta(\cdot)$ as a function to calculate the power attenuation coefficient, the power $B(r_{\text{grid}}^{l})$ of each grid in the source distribution map can be expressed as

$$B(r_{\text{grid}}^{l}) = \sum_{m=0}^{M-1} |q_{s}^{m}|^{2} \cdot \zeta(R(l,m)),$$
(4)

where $\zeta(R) = \varepsilon/R^N + \varepsilon, \varepsilon = (\Delta x)^N/2, N$ is a predefined constant value.

By changing the value of N, the speed of attenuation of sound power can be changed with the increase of R, and the speed of attenuation affects the width of the source distribution map's main lobe. Retaining a moderate main lobe width can improve the distribution map's ability to describe sound sources that are not at the center point of the grid and avoid the extreme sparsity of the sound source distribution map.



FIGURE 6: The structure of the feature restoration unit.

The position and sound power of sound sources in all samples are constructed into a sound source distribution map by the above method, which can form a label set for training U-net networks. The following example uses a response power spectrum and source distribution map for a sample with two sources to illustrate the principle and advantages of using source distribution maps as labels. Figure 8 shows the spatial representation and plane mapping of input features and labels for a sample with two sources, where the "+" mark points of the input features represent the actual locations of the two sources.

It can be seen from Figure 8(a) that the spectral peaks where the two sound sources are located in the response power spectrum are aliased with each other. Some false peaks are near the grid where the sound source is located. The amplitude of these false peaks is close to the true sound source position spectral peak, and if the local maximum search is performed directly on the response power spectrum, the location of the real sound source cannot be located. From Figure 8(b), it can be seen that the amplitude distribution trend of the power spectrum is spread around the center point of the two sound sources, and the center point of the diffusion corresponds to the true position of the sound source. The constructed sound source distribution map is based on this feature, accurately delineating the position of each sound source by analyzing the overall spatial distribution trend of power. It employs (4) to simulate the diffusion of response power in the response power spectrum. From Figures 8(c) and 8(d), it can be seen that the sound source distribution map only retains the true main lobe of the sound source. This approach prevents false peaks from causing misjudgments of the number of sound sources and errors in the localization results during subsequent local maximum searches.

Let \hat{B}_T represent the estimated sound source distribution map after the response power spectrum *P* is input to the Unet network, and the interpolation search of the estimated sound source distribution map can obtain the location of the



FIGURE 7: The construction process of the sound source distribution map.



FIGURE 8: Comparison of U-net network input features and labels: (a) input feature spatial map; (b) input feature plane mapping; (c) label space map; (d) label plane mapping.

sound source, and the implementation process is as follows: first, record the global maximum value of the estimated sound source distribution map as $B = \max(\hat{B}(r_{grid}^l))$, and then record all other local maximums. B_{\min} is a predefined static threshold, and the number of local maxima that meet the above conditions is the estimated number of sound sources. The center point coordinates of the mesh in which these local maximums are located are used as a rough estimate of the location of the sound source. After obtaining a rough estimate of the sound source location, interpolation searches are conducted around that grid to determine the precise position of each sound source. The specific process of interpolation search is shown in Figure 9. A subregion is selected for each local maximum, with the local maximum as the center. This subregion consists of *h* grids. A partition ratio *k* is chosen to divide each grid of the subregion into k^2 subgrids proportionally. The entire subregion is divided into $I = h * k^2$ grids, and then assume that each subdivision grid is the true position of the sound source, and calculate the sound power value $B(r_{\text{grid}}^i)$ of all subdivided grids when the sound source is in this grid, according to (4). By calculating the error between $B(r_{\text{grid}}^i)$ and the estimated distribution map's grid power value $\hat{B}(r_{\text{grid}}^i)$, an error distribution map is obtained, and the



FIGURE 9: Interpolation search (+, local maximum; •, true sound source location).

subdivision grid with the smallest error value is the sound source position \hat{r}_s^m estimated by the localization algorithm. This operation is performed for each local maximum to obtain the estimated positions of all sources.

3. U-Net Network Training and Simulation Experiment Analysis

To train the U-net network, this paper uses the mirrorsource model of the Pyroomacoustics library [20] to simulate sound propagation in a room and generate microphone array received signals. Using the CHZ02-S-112LA2 relay test object, a simulation dataset was constructed to train the U-net network and simulate the location of fault sound sources.

3.1. Simulation Dataset Construction and Model Training. Taking multirelay mechanical fault detection as an example, the basic principle is to install multiple relays to be inspected on the vibrating plate at a specific interval and drive the relay vibration on the vibrating plate through the electromagnetic exciter. The fault relay vibrates under force, causing irregular movement of its armature due to loosening. As the armature moves, it collides with the relay shell, producing passive sound waves that propagate through the relay shell. These sound waves are then captured by the microphone array positioned directly above the relay and converted into multichannel audio signals. Analyzing the audio signal, multiple faulty sound sources are located, and the relay corresponding to the fault sound source location is judged as a fault relay.

To ensure the diversity of the simulation dataset, the relay fault acoustic signal without noise and reverberation is first collected, and its diversity is ensured by setting different numbers of sound sources, different reverberation times, and different signal-to-noise ratios, and the maximum number of sound sources in the training sample is set to 4.

The collection of fault acoustic signals is carried out in an anechoic chamber, the recording environment is shown in Figure 10, the relay fixture is fixed on the exciter, the inner wall of the anechoic chamber has a layer of sound-absorbing cotton to suppress reverberation, and the outside has a layer of soundproof cotton to isolate noise. All 500 real relays vibrate continuously at 25 Hz during the recording, and the audio sampling frequency is 48 kHz.



FIGURE 10: Collection of relay fault sound signals.

Figure 11(a) shows the time-domain waveform of the relay fault sound signal. It can be observed that the fault sound signal exhibits distinct segmented characteristics, as it is not continuously emitting sound but rather has certain time intervals between each sound emission. The spectrum obtained by performing a Fast Fourier Transform (FFT) on the fault signal is shown in Figure 11(b). It can be seen that the fault sound signal possesses a continuous frequency spectrum, and based on the relationship between center frequency and signal bandwidth, the fault sound signal should be considered as a broadband signal.

Perform frame-by-frame processing on the audio data, dividing the audio into consecutive time segments. Apply a Fast Fourier Transform (FFT) to each audio data segment to convert it from the time domain to the frequency domain. Then, concatenate the spectra of all segments along the time axis to obtain a two-dimensional matrix, which results in the spectrogram of the fault sound signal (Figure 12). Different colors represent the energy level of the signal segments at different frequencies. Darker areas in the spectrogram indicate higher energy levels at those frequency points.

The audio data samples are generated in the room depicted in Figure 13. The simulated room size is set to $3 \text{ m} \times 4 \text{ m} \times 3 \text{ m}$, the 8-element circular array with a radius of 0.2 m is placed on a plane with a vertical height of 1 m from the ground, the coordinates of the center point of the array are (1.5 m, 2 m, 1 m), and the microphone sampling rate is set to 48 kHz.

First, a set of 2000 random coordinates is generated on the sound source plane {x, y, z | 1.5 < x < 2.5, 2 < y < 4, z = 0.5} (unit m). Each time, 1~4 copies are selected from 500 real relay fault audio recordings to serve as the sound source signal. Then, the same number of coordinates in the coordinate set is randomly selected, and the sound source is placed, the signal-to-noise ratio is randomly set between 0 dB and 25 dB, and the reverberation time is randomly set between 0.2 s and 0.7 s. The multichannel audio data collected by the microphone array are used as samples, and 2000 samples are generated when the number of sound sources is 1 to 4, and a total of 8000 samples are generated to form the training set.



FIGURE 11: The waveform and spectrum of fault sound signal: (a) time-domain waveform; (b) the spectrum.



FIGURE 12: Spectrogram of fault sound signal.



FIGURE 13: Simulated room schematic.

The training process adopts the deep learning framework TensorFlow 2.6.0 and is implemented on the Windows 11 operating system. The training environment includes an Intel Core i5-11600 processor and NVIDIA GeForce RTX 3060 graphics card. The training samples are input to the network model for training, the batch size is set to 20, the number of training iterations is 500, the network initialization learning rate is 0.0005, and the Adam algorithm is used to optimize the network parameters.

Before model training, the sample data are preprocessed, the audio frame with a sample point of 2048 is intercepted from the multichannel audio data, the response power spectrum is calculated by SRP-PHAT, the grid resolution is set to 0.02 m, the scanning range is the sound source plane range, and the output dimension is 100×100 response power spectrum; to maintain the consistency of the sample data, all the response power spectrum needs to be normalized. According to the position and sound power of all sound sources, an actual sound source distribution map B_T with dimensions of 100×100 is constructed, and the normalized sound source distribution map is used as the training label for the samples.

In U-net network training, the predicted sound source distribution map can be regarded as a classification task for each pixel value, and the pixel mean squared error is used to represent the error between the actual source distribution map and the reconstructed sound source distribution map.

 $B_T(r_{\text{grid}}^l)$ and $\hat{B}_T(r_{\text{grid}}^l)$ represent the pixel values of the *l*th grid in the true source distribution map B_T and the predicted sound source distribution map \hat{B}_T , and *N* represents the total number of pixel points in the sound source distribution image, and the loss function can be defined as

$$L(B_T, \widehat{B}_T) = \frac{1}{N} \sum_{l} \left(B_T(r_{\text{grid}}^l) - \widehat{B}_T(r_{\text{grid}}^l) \right)^2.$$
(5)

3.2. Simulation Experiment Analysis

3.2.1. Positioning Accuracy Experiment under Different Reverberation times. To visually observe the effect of the U-net network in eliminating false peaks in the response power spectrum, localization simulation experiments were carried out under different conditions: SNR = 10 dB, $RT_{60} = \{0.2 \text{ s}, 0.5 \text{ s}, 0.7 \text{ s}\}$.

Figures 14(a)-14(c) and Figures 14(d)-14(f) compare the U-net input feature response power spectrum and the Unet network's output sound source distribution map. The "+" in the response power spectrum indicates the actual positions of the sound sources. The comparison between Unet input and output reveals that reverberation led to varying degrees of distortion in the response power spectrum. In the case of a reverberation time of 0.7 s, the false peak is more obvious, and the trained U-net network accurately distinguishes the false peak and the spectral peak where the real sound source is located and accurately finds all the sound sources. From the localization results in Table 1, the localization error in all cases is less than 0.02 m, which proves that the sound source localization algorithm has high positioning accuracy and localization stability.

3.2.2. Analysis of Spatial Resolution under Different Signalto-Noise Ratios (SNRs). Spatial resolution refers to the minimum separation between two adjacent sound sources that a sound source localization method can distinguish. Two sound source signals are placed in the sound source plane, and the algorithm proposed in this paper is used for localization. Then, the sound source spacing is continuously reduced with a step size of 0.01 m, and the localization algorithm can identify the minimum spacing between the two sound source positions as the experimental result. Table 2 presents the spatial resolution of the algorithm in an environment with SNRs of {0 dB, 5 dB, 10 dB, 15 dB, 20 dB, 25 dB and an RT₆₀ of 0.7 s. It can be seen from the results that the algorithm's spatial resolution is affected by noise. As the signal-to-noise ratio increases, the algorithm's spatial resolution also improves. At an SNR of 25 dB, the proposed method in this paper can distinguish two sound sources spaced 0.08 m apart. When the SNR decreases to 0 dB, the method can distinguish sound sources spaced 0.2 m apart.

3.2.3. Multisource Localization Simulation Experiment. To test the effectiveness of the proposed method for different numbers of sound sources, a simulation experiment for multisource localization is conducted. The accuracy of localization and root mean square error (RMSE) are used as evaluation metrics.

The simulation parameters are as follows: SNR = 10 dB, RT₆₀ = 0.5 s, the grid resolution is set to 0.02 m, and the distance error threshold v is set to 0.02 m. During calculating accuracy, the sound power threshold is set to 0.6 when interpolating the search. Five sets of tests were carried out under each environmental characteristic configuration. Each group contains 120 samples, and the average of the results from these five groups was used as the experimental results. Based on the previously obtained algorithm spatial resolution, when the number of sound sources is greater than 1, the distance between any two sound sources is constrained to be greater than 0.2 m.

Table 3 shows the test results of the proposed algorithm's localization accuracy and root mean square error (RMSE) for different numbers of sound sources ranging from 1 to 5. It can be seen that when the number of sound sources is less

than or equal to the maximum number of sound sources (4) in the training sample, the localization accuracy is greater than 98%, and the RMSE is less than 0.014 m. When the number of sound sources increases to 5, because the U-net training set does not contain samples of 5 sound sources, the positioning performance decreases slightly, but it still maintains high positioning accuracy and small positioning error, which proves that the proposed method can accurately and stably locate the position of multiple sound sources.

3.3. Experiment on Localization of Multiple Faulty Relay Sound Sources

3.3.1. Experimental Environment and Parameter Settings. The experiment was carried out in a $4 \text{ m} \times 6 \text{ m} \times 4 \text{ m}$ room, and the overall experimental environment is shown in Figure 15. The exciter was placed on the ground, and a $0.22 \text{ m} \times 0.22 \text{ m}$ vibrating disk was attached to the top pole of the exciter to fix the relays. An 8-element microphone array was chosen in a circular arrangement with a radius of r = 0.2 m, and the microphone array was placed 0.5 m directly above the vibrating plate and was parallel to the vibrating plate. The plane where the vibrating disk was located was the sound source plane scanned by the beam during the sound source positioning process, and the sound source surface grid was established at the origin of the point vertically mapped to the sound source surface by the center of the array, and the four relays with mechanical faults were installed in the four positions of the vibration plate. The exciter was activated to cause the relays to vibrate and emit faulty relay sounds, which were then captured as audio signals for the experiments.

The relay vibration sound device comprises an SA-SG signal generator, SA-PA power amplifier, and SA-JZ electromagnetic exciter. The signal generator sends a sine wave electrical signal, which is amplified by the power amplifier and input to the electromagnetic exciter; the exciter drives some fixed relays to vibrate. The microphone array uses MSM261S4030H0R omnidirectional digital microphones. The audio capture card operates at frequencies of up to 160 MHz, with sample rates supporting 8 K, 16 K, 22.05 K, 24 K, 32 K, 44.1 K, and 48 K.Additionally, it simultaneously acquires 8 microphone signals by combining the left and right channels.

According to the spatial resolution results of Subsection 3.2.2, when the signal-to-noise ratio is reduced to 0 dB, sound sources that are 0.2 m apart can be resolved, so the interval between adjacent relays is set to 0.2 m, and the installation location is shown in Figure 16.

3.3.2. Analysis of Experimental Results. Fault relays were installed at all four locations shown in Figure 16, and the trained U-net network model was loaded for 50 sound source localizations. The average of the 50 localization results was used as the experimental result. When performing sound source localization, the beam scanning plane grid resolution is set to 0.01 m, and the scanning range is $\{x, y | -0.5 < x < 0.5, -0.5 < y < 0.5\}$ (in m). The audio



FIGURE 14: U-net input and output under different reverberation times: (a) U-net input $(RT_{60} = 0.2 s)$; (b) U-net input $(RT_{60} = 0.5 s)$; (c) U-net input $(RT_{60} = 0.7 s)$; (d) U-net output $(RT_{60} = 0.2 s)$; (e) U-net output $(RT_{60} = 0.5 s)$; (f) U-net output $(RT_{60} = 0.7 s)$.

Reverberation time (s)	Sound source	True sound source location (m)	Localization results (m)	Localization distance error (m)
0.2	Fault source 1	(-0.3, 0.3)	(-0.312, 0.308)	0.014
	Fault source 2	(0.3, 0.3)	(0.290, 0.310)	0.014
	Fault source 3	(0.5, 0.2)	(0.494, 0.210)	0.012
0.5	Fault source 1	(-0.3, 0.3)	(-0.294, 0.310)	0.012
	Fault source 2	(0.3, 0.3)	(0.316, 0.310)	0.018
	Fault source 3	(0.5, 0.2)	(0.488, 0.206)	0.013
0.7	Fault source 1	(-0.3, 0.3)	(-0.310, 0.288)	0.016
	Fault source 2	(0.3, 0.3)	(0.316, 0.308)	0.016
	Fault source 3	(0.5, 0.2)	(0.518, 0.194)	0.019

TABLE 1: Localization results under different reverberation times.

SNR (dB)	0	5	10	15	20	25
Spatial resolution (m)	0.20	0.15	0.12	0.11	0.09	0.08
	TABLE 3: 1	Multisource localiz	ation experin	nent results.		
			Numbe	er of sound sour	ces	
	1	2		3	4	5
Localization accuracy (%)	98.3	98.6		98.0	98.5	95.2
RMSE (m)	0.0127	0.0129		0.0134	0.0123	0.0169

TABLE 2: Spatial resolution of localization algorithm under different SNRs.



FIGURE 15: Experimental platform for fault sound source localization.



FIGURE 16: Schematic diagram of relay installation positions.

Shock and Vibration

Sound source	True sound source location (m)	Localization results (m)	Localization distance error (m)
Fault source 1	(-0.1, -0.1)	(-0.084, -0.107)	0.017
Fault source 2	(0.1, -0.1)	(0.114, -0.105)	0.015
Fault source 3	(-0.1, 0.1)	(-0.119, 0.103)	0.019
Fault source 4	(0.1, 0.1)	(0.093, 0.115)	0.016

TABLE 4: The localization results.



FIGURE 17: Response power spectrum and U-net outputs for four fault sound sources: (a) response power spectrum; (b) U-net output.

sampling frequency is set to 48 kHz, 2048 sampling points are taken each time for positioning, and the shaker vibrates at a frequency of 25 Hz.

Table 4 presents the localization results and localization errors. Figure 17(a) is the response power spectrum of the four fault sound source positioning tests, and Figure 17(b) is the U-net output sound source distribution diagram, in which the " * " mark point is the location of the sound source positioned by the algorithm in this paper; it can be seen that the multifault sound source localization method proposed in this paper identifies the location of all fault relays.

Based on the results in Table 4, the average localization error is less than 0.02 m, which is significantly smaller than the spacing between the relays. In practical relay fault detection scenarios, if four relays simultaneously have a mechanical failure and abnormal sound, the algorithm presented in this paper would correctly identify all malfunctioning relays. The experimental results prove that this algorithm can simultaneously localize sound sources from four malfunctioning relays in a real-world environment.

4. Conclusion

In this paper, the estimation method of mechanical fault sound source localization under strong reverberation of multiple sound sources is studied, the SRP-PHAT algorithm is used to calculate the response power spectrum, and a Unet network is utilized to transform the response power spectrum with spurious peaks into a "clean" estimated sound source distribution map. The accurate location of fault sound sources is realized through interpolation search. The research employs the SRP-PHAT algorithm to perform crosscorrelation and phase transformation weighting on multichannel audio signals, creating directional beams. These beams scan the entire sound source plane and calculate the response power spectrum of the sound source plane. Then, the U-net encoder network and decoder network for sound source distribution prediction are constructed, the power spectrum feature extraction unit and the sound source distribution map feature reduction unit are designed, and the interpolation search method based on the sound source distribution map is studied to estimate the precise location of each fault sound source.

The experimental dataset was constructed from the mechanical fault data of the relay of electromechanical equipment to train the U-net network. The experimental results show that the reverberation time increases from 0.2 s to 0.7 s, and the U-net network can still effectively eliminate the pseudopeak interference of the response power spectrum; when the signal-to-noise ratio is reduced from 25 dB to 0 dB, the spatial resolution increases from 0.08 m to 0.2 m; when the relay multifault sound source is located, the position of four fault sound sources can be located at the same time, and the average positioning error is less than 0.02 m. Compared with the traditional threshold processing and smoothing processing methods, the method proposed in this study can eliminate false peaks and accurately locate multifault sound sources without affecting the real sound source signal, which provides a new method for solving the problem of sound source localization in the scene of strong reverberation of multiple sound sources.

Due to experimental limitations, this study only used an 8-element circular microphone array to collect audio signals and conducted localization experiments with only 4 fault sources. However, in the actual mechanical fault detection, the microphone array is diverse, and there are more fault sound sources, so we will focus on exploring different microphone arrays, increase the number of sound sources, and different signal types for our research. Additionally, we plan to conduct experiments on sound source localization in multisource strong reverberation scenarios using various algorithms to compare, such as MUSIC, CNN-DOA, and DAMAS, enhance localization accuracy, and increase their adaptability.

Data Availability

The data used to support the findings of this study are included within the article.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (52175457), Basic and Applied Basic Research Foundation of Guangdong Province (2022B1515120053), Guangdong Science and Technology Plan Project (2023A0505050151), Jieyang Science and Technology Innovation Plan Project (231031156996165), and Heyuan Science and Technology Plan Project (230511101473496).

References

- [1] D. Zhang, E. Stewart, and M. Entezami, "Intelligent acousticbased fault diagnosis of roller bearings using a deep graph convolutional network," *Measurement*, vol. 156, 2020.
- [2] F. Huda, A. Anggriawan, and M. Rusli, "The using of sound signal and simple microphone to detect damages in induction motor," *IOP Conference Series: Materials Science and Engineering*, vol. 539, no. 1, Article ID 012034, 2019.
- [3] P. Schober, S. N. Estiri, S. Aygun, A. H. Jalilvand, M. H. Najafi, and N. TaheriNejad, "Stochastic computing design and implementation of a sound source localization system," *IEEE journal on emerging and selected topics in circuits and systems*, vol. 13, no. 1, pp. 295–311, 2023.
- [4] X. Qu and L. Xie, "An efficient convex constrained weighted least squares source localization algorithm based on TDOA measurements," *Signal Processing*, vol. 119, pp. 142–152, 2016.
- [5] J. Lim, Y. Pyeon, and M. Cheong, "GCC-PHAT based time delay estimation using BPD," *The Journal of Korean Institute* of Communications and Information Sciences, vol. 42, no. 9, pp. 1857–1862, 2017.
- [6] L. Perotin, R. Serizel, E. Vincent, and A. Guerin, "CRNNbased multiple DoA estimation using acoustic intensity features for Ambisonics recordings," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 1, pp. 22–33, 2019.
- [7] S. Y. Jia, M. Lu, H. Z. Ding, M. Chen, and L. Y. Zhao, "A modified wideband DOA estimation algorithm for focusing signal Subspace," *Engineering Computers*, vol. 48, no. 6, pp. 175–181, 2022.
- [8] X. Qian, Q. Zhang, G. Guan, and W. Xue, "Deep audio-visual beamforming for speaker localization," *IEEE Signal Processing Letters*, vol. 29, pp. 1132–1136, 2022.

- [9] J. S. Lim, M. J. Cheong, and S. Kim, "Improved generalized cross correlation-phase transform based time delay estimation by frequency domain autocorrelation," *The Journal of the Acoustical Society of Korea*, vol. 37, no. 5, pp. 271–275, 2018.
- [10] J. H. DiBiase, H. F. Silverman, and M. S. Brandstein, "Robust localization in reverberant rooms," *Microphone arrays: Signal Processing Techniques and Applications*, vol. 58, no. 7, pp. 157–180, 2001.
- [11] S. Shi, B. Yang, Q. Guo, Y. Li, and C. Gui, "Low-frequency sound source localization and identification with spherical microphone arrays extrapolation method," *Frontiers in Physics*, vol. 11, no. 4, 2023.
- [12] S. Chakrabarty and E. A. P. Habets, "Multi-speaker localization using convolutional neural network trained with noise," *Workshop on Machine Learning for Audio Processing*, vol. 1712, no. 04276, pp. 156–161, 2017.
- [13] P. Jin, B. Wang, and L. Li, "Semi-supervised underwater acoustic source localization based on residual convolutional autoencoder," *EURASIP Journal on Applied Signal Processing*, vol. 107, no. 1, 2022.
- [14] S. Chakrabarty and E. A. P. Habets, "Multi-speaker DOA estimation using deep convolutional networks trained with noise signals," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 1, pp. 8–21, 2019.
- [15] D. Salvati, C. Drioli, and G. L. Foresti, "Exploiting CNNs for improving acoustic source localization in noisy and reverberant conditions," *IEEE Transactions on Emerging Topics* in Computational Intelligence, vol. 2, no. 2, pp. 103–116, 2018.
- [16] S. Adavanne, A. Politis, J. Nikunen, and T. Virtanen, "Sound event localization and detection of overlapping sources using convolutional recurrent neural networks," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 1, pp. 34–48, 2019.
- [17] A. Senocak, H. Ryu, and J. Kim, "Sound source localization is all about cross-modal alignment," in *Proceedings of the IEEE International Conference on Computer Vision*, New York, NY, USA, December 2023.
- [18] S. Park, A. Senocak, and J. S. Chung, "MarginNCE: robust sound localization with a negative margin," in *Proceedings of the 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, Rhodes Island, Greece, August 2023.
- [19] O. Ronneberger, P. Fischer, and T. Brox, "U-net: convolutional networks for biomedical image segmentation," *Medical Image Computing and Computer-Assisted Intervention-MICCAI 2015: 18th International Conference*, Springer, pp. 234–241, Munich, Germany, 2015.
- [20] R. Scheibler, E. Bezzam, and I. Dokmanić, "Pyroomacoustics: a python package for audio room simulation and array processing algorithms," in *Proceedings of the IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 351–355, Calgary, Canada, April 2018.