Three-Dimensional Sound Field Reconstruction and Sound Power Estimation by Stereo Vision and Beamforming Technology

Yetian Cai 1, Xiaqin Liu 1,*, Yanjiao Xiong 1,2 and Xing Wu 1,3

Abstract: The size of the sound field reconstruction area has an important influence on the beamforming sound source localization method and determines the speed of reconstruction. To reduce the sound field reconstruction area, stereo vision technology is introduced to continuously obtain the three-dimensional surface of the target and reconstruct the sound field on it. The fusion method can quickly locate the three-dimensional position of the sound source, and the computational complexity of this method is mathematically analyzed. The sound power level can be estimated dynamically by the sound intensity scaling method based on beamforming and the depth information of the sound source. Experimental results in a hemi-anechoic chamber show that this method can quickly identify the three-dimensional position of the moving source. When the depth of the moving sound source changes, the estimated sound power is more stable than the sound pressure on the microphone.

Keywords: moving sound source; beamforming; 3D surface; sound state

1. Introduction

Beamforming is a method of processing array signals and has mature applications in many fields, such as sonar, mobile communications, medical imaging, and radio astronomy [1–3]. Davids and Billingsley applied this method to sound source localization [4]. The beamforming method based on the acoustic array can locate steady-state and transient sound sources at medium and long distances with medium and high frequencies [5]. The beamforming method has been widely used in mechanical structures such as airplanes [6], trains [7], and automobiles [8].

The acoustic array-based beamforming is to locate the sound source by reconstructing the sound field in the 2D or 3D space [9]. The usual method is to assume the depth Z of the plane where the sound source is located, then perform gridding on this plane (focus reconstruction surface) and calculate the sound field output of each grid point (reconstruction point) to complete the sound field reconstruction of the plane. This continuously iterates planes of different depths to achieve the sound field reconstruction in the spatial region. The grid point (reconstruction point) where the maximum output is located is the position of the sound source. This method is limited by the number of reconstruction points and requires huge computation.

Introducing the three-dimensional information of the object to be recognized and reducing the sound field reconstruction range is an effective way to reduce the amount of computation and achieve rapid positioning of the three-dimensional sound source. Döbler et al. [10] used the car’s three-dimensional model as the sound field reconstruction...
area, which avoided the sound field reconstruction of the area outside the car body, and reduced the range. The above methods have the advantages of fewer reconstruction points and fast speed for sound source recognition, but the model information needs to be input into the computer in advance. Limited by the static characteristics of the model, it is difficult to identify the sound source in motion. In addition to obtaining structural information by importing the model, Legg [11] uses structured light scanning technology to enhance acoustic beamforming. In the experiment, the structure information of a fixed plexiglass plate is obtained and the sound source inserted into it is located, and this proves that the acoustic imaging error is an order of magnitude lower than that of a 2D plane.

Some scholars have tried to combine beamforming with vision. For example, Bub et al. [12] and Virginia et al. [13] used a camera to dynamically track the face of the speaker, and only searched in the face area to improve the recognition accuracy of beamforming. Jennings et al. [14] used visually guided beamformer to improve listening in complex acoustic environments. José Novoa et al. [15] used visual tracking on the robot and combined beamforming to locate the voice source. O’Donovan et al. [16] developed a real-time audio camera to more intuitively check the relationship between acoustic features and architectural details. In recent years, the rapid development of binocular cameras has made this an effective way to obtain 3D information in space. It has the advantages of fast measuring 3D information, high accuracy, mature technology, and continuous output of scene depth data [17,18]. Combining the binocular camera and beamforming technology can obtain the three-dimensional surface information of the operating mechanical structure, and complete the surface sound field reconstruction of the mechanism.

The radiated sound power is an important parameter to describe the acoustic characteristics [19,20]. For example, mechanical failure will produce noise, and accurate measurement of its radiation noise can obtain the degree of failure. The sound intensity scaling method estimates sound power, which is accurate and fast under the condition of obtaining the depth of the sound source [21]. This method can be combined with the three-dimensional positioning of the mobile sound source to monitor the sound power of the moving sound source.

The major contributions of this study are as follows:
1. The introduction of binocular camera technology continuously obtains the three-dimensional surface of the operating mechanism and performs sound field reconstruction on it, which reduces the number of reconstruction points and improves the positioning speed. Avoiding the import of 3D models and static restrictions realizes continuous 3D positioning of mobile sound sources;
2. Combining the three-dimensional positioning results with the sound intensity scaling method completes the sound power monitoring of moving sound sources with depth changes.
3. Through mathematical analysis, the computation amount between conventional beamforming and fusion methods is compared. In the experiment, the average time consumption of the two methods to reconstruct the sound field was recorded, which proved that the amount of computation can be effectively reduced.

2. Methods
2.1. Focused Beamforming in Three-Dimensional Space

Generally, conventional beamforming (CBF) is to search on a reconstruction surface with a preset depth, and the Z coordinate of the reconstruction point is the same as the Z coordinate of the sound source. The three-dimensional spatial focused beamforming searches in a spatial region composed of multiple reconstruction planes of different depths, as shown in Figure 1.

A planar array containing M microphones is used to locate the point sound source S in the space. The coordinate of the m-th microphone is \((x_m, y_m, z_m)\) \((m = 1, 2, 3 \ldots, M)\). The signal strength of the point sound source S received by different microphones attenuates
as the distance increases [9]. The sound pressure signals measured by the \( m \)-th microphone is \( P_m(\omega) \). The first microphone \( (x_1, y_1, z_1) \) is taken as the reference. Point \( f(x_f, y_f, z_f) \) is a reconstruction point in the sound pressure reconstruction area in three-dimensional space. \( r_f \) is the position vector of the reference microphone to the reconstruction point \( f \), and \( r_{mf} \) is the position vector of the \( m \)-th microphone to the reconstruction point \( f \), then the delay time \( \tau_m(x_f, y_f, z_f) \) of the \( m \)-th microphone to the reference microphone is as follows [9]:

\[
\tau_m(x_f, y_f, z_f) = \frac{|r_f| - |r_{mf}|}{c} \tag{1}
\]

In formula (1), \( | \cdot | \) is the vector modulus, and \( c \) is the speed of sound.

According to the principle of delay-and-sum beamforming, the sum of the phase delay \( \tau_m(x_f, y_f, z_f) \) of each microphone signal is compensated. The delay-and-sum beamformer output of point \( f \) is as follows [9]:

\[
B(x_f, y_f, z_f, \omega) = \frac{1}{M} \sum_{m=1}^{M} w_m P_m(\omega) e^{-j\omega \tau_m(x_f, y_f, z_f)} \tag{2}
\]

where \( w_m \) is the weighting coefficient of the \( m \)-th microphone [9].

Then, the sound pressure reconstruction points in all 3D reconstruction areas are calculated. The complex output of the sound pressure at each point \( B(x_f, y_f, z_f, \omega) \) constitutes the acoustic imaging map. The reconstruction point with the maximum sound pressure output \( B_{\text{max}}(x_f, y_f, z_f, \omega) \) is the sound source position.

![Figure 1. Principle of space beamforming.](image)

### 2.2. Binocular Stereo Vision

After the long-term development of stereo vision technology, there are many mature solutions already. This study uses the Intel® RealSense™ depth camera (Manufactured by Inter Corporation, Santa Clara, CA, USA), which can directly operate on its built-in chip and output a 3D point cloud of the scene. The point cloud is presented as a depth image on the screen. The camera uses active infrared stereo vision technology, and the depth perception is realized by two image sensors and an infrared projector, as shown in Figure 2.

The infrared projector projects invisible structured infrared light to improve the depth accuracy of the scene. Two image sensors simulate human eyes, and obtain two images of
the measured area at the same time from different positions, and obtain the spatial three-dimensional information of each pixel through the stereo vision algorithm [22, 23].

![Image](image_url)

**Figure 2.** Principle of active infrared stereo vision technology.

2.3. Sound Intensity Scaling Theory Based on Beamforming Technology

The output of beamforming reflects the sound pressure contribution of the sound source on the plane of the microphone array. Jorgen Hald [21] scales the beamforming output result by constructing a scaling factor to ensure that the integral of the scaled output in the main lobe area is equal to the sound power radiated by the sound source on the side of the microphone array.

The scaled beamforming result has sound intensity meaning, and defined as “scaling sound intensity”, the coefficient is defined as “sound intensity scaling factor” $\alpha$ [21], and it satisfies:

$$\begin{align*}
B_i(R, \varphi) &= \alpha \cdot |B(R, \varphi)|^2 \\
W_h &= \int_{0}^{R_1} \int_{0}^{2\pi} \alpha \cdot |B(R, \varphi)|^2 Rd\varphi
\end{align*} \tag{3}$$

In Formula (3), the $(R, \varphi)$ polar coordinate system is used to describe each focus grid point on the sound source calculation plane, $R$ is the distance from the focus grid point to the intersection of the sound source calculation plane and the Z-axis, and $\varphi$ is the azimuth angle. $B_i(R, \varphi)$ is the sound intensity after scaling the beamforming output result, and $W_h$ is the sound power of the point source on the side of the microphone array.

By constructing a fixed sound source recognition model and mathematical derivation, the sound intensity scaling factor $\alpha$ is obtained by:

$$\alpha = \frac{1}{16\pi} \frac{(kD)^2}{\rho c} \approx \frac{5.88}{\rho c} \left( \frac{D}{\lambda} \right)^2 \tag{4}$$

$D$ is the array diameter, m. $\rho$ is the air density, kg/m³. $\lambda$ is the wavelength, m. $\alpha$ is the sound intensity scaling factor.

The integral of the scaled sound intensity in the main lobe area is the sound power of the side hemisphere of the sound array, and the sound power $W_h$ is [21]:

$$W_h = \alpha \cdot 32\pi \left( \frac{|P_r|L}{kD} \right)^2 \cdot 0.419 \tag{5}$$

$L$ is the sound source depth, m. $|P_r|$ is the effective amplitude of the microphone signal. $k$ is the wave number in the propagation direction of the plane wave.

According to Equation (5), it is necessary to obtain the depth of the sound source to calculate the sound power of the moving sound source. In this paper, the stereo vision
method is introduced to obtain the three-dimensional positioning of the sound source, then the sound power can be estimated.

3. The Sound Source Monitoring Combing Stereo Vision and Beamforming

The stereo vision system and microphone array are used to collect optical and acoustic information at the same time in the space where the sound source is located. The target area of the sound source obtained by the former is used as the search domain for the beamforming sound source 3D localization, thereby narrowing the searching range and reducing the amount of computation. Specific steps are as follows:

1. Use binocular cameras and microphone arrays to acquire images and sound signals at the same time.
2. Segment the sound signals corresponding to the frame rate of the depth image, that is, the time matching of the image and the sound.
3. Segment the target from the depth image, and obtain the 3D information of the target surface.
4. Convert the 3D information of the target surface to the acoustic array coordinate system to complete the spatial matching of the two systems.
5. The 3D information of the target surface after coordinate conversion is used as the beam-forming reconstruction surface, and the focused beamforming is performed to reconstruct the sound field of the target 3D surface. The sound pressure reconstruction point where the maximum output is located is the three-dimensional coordinate of the sound source.
6. Import the depth of the three-dimensional coordinates into the sound intensity scaling method to obtain the sound power of the hemisphere on the side of the sound array.

3.1. Three-Dimensional Coordinates Extraction of Target Machinery Surface

Stereo vision can obtain the depth image of the entire field of view including the target, and extracting the three-dimensional coordinates of the target machinery surface is one of the key points of this method.

The threshold method is used to segment the depth image to obtain the pixel set of the target machinery surface. The threshold method compares the color value of the image pixel with a given threshold and divides the pixel points into two types, the background, and the target. This method has high computational efficiency and stable performance. Assuming that \( f(x, y) \) is the original image and the threshold is \( T \). The image can be divided into two parts by the following formula:

\[
g(x, y) = \begin{cases} 
    b_0 & f(x, y) < T \\
    b_1 & f(x, y) \geq T 
\end{cases}
\] (6)

This method uses \( b_0 = 0 \) (black) and \( b_1 = 1 \) (white) to realize image binarization. In this study, the black area is the target. All pixels in the black area are coordinated, and the depth of these pixels can be obtained.

3.2. Coordinate System Transformation

The camera coordinate system takes the stereo camera as the origin, and the coordinate system for beamforming usually takes the center of the array as the origin. Spatial matching is needed to coincide the two systems. Coordinate system transformation is realized through rotation and translation between two coordinate systems.

Before the coordinate system transformation, each pixel in the depth image output by the stereo camera needs to be transformed into the camera coordinate system. According to the imaging principle of the camera, the spatial transformation relationship between the pixel coordinates and the camera coordinate system \((o, u, v, w)\) can be established, as shown in Equation (7):
\[
\begin{bmatrix}
 x_1 \\
 y_1
\end{bmatrix} = 2l \begin{bmatrix}
 \tan \frac{a_0}{2}/U & 0 \\
 0 & \tan \frac{\beta_0}{2}/V
\end{bmatrix} \begin{bmatrix}
 x \\
 y
\end{bmatrix} - l \begin{bmatrix}
 \tan \frac{a_0}{2} \\
 \tan \frac{\beta_0}{2}
\end{bmatrix}
\tag{7}
\]

where \((x_1, y_1, l)\) is the spatial coordinate of the pixel in the camera coordinate system, and \(l\) is the depth of the pixel. \(U \times V\) is the resolution of the output depth image. \(a_0\) and \(\beta_0\) are the horizontal view angle and the vertical view angle of the camera, respectively.

The camera coordinate system \((o, u, v, w)\) and the array coordinate system \((Oz, Xz, Yz, Zz)\) are different spatial coordinate systems. The use of beamforming to reconstruct the sound field needs to be carried out in the array coordinate system, and their projection relationship is shown in Figure 3.

**Figure 3.** Relationship between array and stereo vision coordinate system.

Assuming that point \((x_1, y_1, l)\) in the camera coordinate system, the corresponding coordinates in the array coordinate system is \((X_{bf}, Y_{bf}, Z_{bf})\). The relationship between the two coordinate systems:

\[
\begin{bmatrix}
 X_{bf} \\
 Y_{bf} \\
 Z_{bf}
\end{bmatrix} = M_R \begin{bmatrix}
 x_1 \\
 y_1 \\
 l
\end{bmatrix} = \begin{bmatrix}
 X_{zu} \\
 Y_{zu} \\
 Z_{zu}
\end{bmatrix}
\tag{8}
\]

\[
M_R = R_u(\alpha)R_v(\beta)R_w(\theta)
= \begin{bmatrix}
 1 & 0 & 0 & \cos (\beta) & 0 & -\sin (\beta) \\
 0 & \cos (\alpha) & \sin (\alpha) & 0 & 1 & 0 \\
 0 & -\sin (\alpha) & \cos (\alpha) & 0 & 0 & \cos (\beta)
\end{bmatrix} \begin{bmatrix}
 \cos (\theta) & \sin (\theta) & 0 \\
 -\sin (\theta) & \cos (\theta) & 0 \\
 0 & 0 & 1
\end{bmatrix}
\tag{9}
\]

where \(M_R\) is the three-dimensional rotation matrix [24], which makes the camera coordinate system rotate around the \(w\) axis by \(\theta\), rotate around the \(v\) axis by \(\beta\), and rotate around the \(u\) axis by \(\alpha\). Then the two space coordinate systems are parallel to each other. The vector \([X_{zu}, Y_{zu}, Z_{zu}]^T\) is the translation between the origins of the two coordinate systems.

After coordinate system transformation, the target area in the depth map obtained by stereo vision is converted to the corresponding 3D point cloud. The point cloud is transformed into the microphone array coordinate system. The reconstruction points of the three-dimensional sound source position are replaced with the 3D point cloud, and the beamforming calculation is performed on the surface of the target to achieve the sound field reconstruction on the 3D surface.
3.3. Time Matching

The signal acquisition is recorded by the visual acquisition system and the acoustic acquisition system, and the matching of the two signals is completed through time synchronization. In this study, simultaneous acquisition is achieved by setting an external trigger. Then according to the frame rate $fps$ captured by the camera, the shooting time $t_i$ of the $i$-th frame can be obtained. Next, the acoustic signal in the time $[t_i - \frac{1}{2} fps, t_i + \frac{1}{2} fps]$ is used to reconstruct the sound field on the $i$-th frame.

When the $fps$ is high, it can be considered that the reconstruction result of using each frame of depth image is the instantaneous sound field at that moment.

4. Computational Complexity of 3D Sound Field Reconstruction

The three-dimensional surface of the target machinery is obtained by stereo vision technology, and the sound field is reconstructed on this surface. Therefore, it is possible to avoid the reconstruction of the sound field in the redundant area outside the organization area and reduce the computational complexity. At the same time, however, it also increases computation in some steps. The following is the analysis process of the amount of computation during positioning.

The computations for beamforming when reconstructing the sound field mainly comes from two aspects: (1) the number of reconstruction points, (2) the amount of computations inside each reconstruction point. Through the binocular camera and extracting the operating mechanism area, the number of reconstruction points can be reduced. But the method in this paper also increases the amount of computations for 3D information extraction and internal coordinate conversion of each reconstructed point.

In CBF, $L$ is the number of data points of the acoustic signal, the reconstruction surface contains $N$ reconstruction points, and the number of microphones is $M$. According to the beamforming principle in Section 2.1, the main computations for a single reconstruction point include:

1. Calculating the distance vector from each microphone to the reconstructed point, which includes $3M$ times of square/multiplication/division, $M$ times of root, and $5M$ times of addition/subtraction;
2. In formula (1) $M$ times the multiplication/division and addition/subtraction are required to calculate the time delay;
3. Equation (2) performs compensation summation and normalized sound pressure complex output for each channel, which requires about $14 \times M \times L$ times of multiplication/division, and $M$ times the natural number exponential operations, and $M$ times of addition/subtraction calculations.

4.1. The Computation Increasement of the Combination Method

When combing the depth image of stereo vision, the amount of the increased computation mainly includes:

1. Target region extraction: using the threshold method to segment the image to extract target information. The segment uses logical operations.
2. Coordinate system transformation: (1) Image coordinate computations increases 2 tangent operations. (2) Projection and translation increase about 19 times of multiplication or division, 1 time of Root, and 4 times of addition or subtraction. (3) Calculating the rotation matrix $M_R$ increases 6 times of sine or cosine. (4) Each reconstructed point projected to the array coordinate system by rotation and translation (Equation (8)) increases 12 times of addition or subtraction and 43 times the multiplication or division (MD).

In the above calculations, roots, trigonometric functions, and exponents are complex calculations. The root operation uses the Newton iteration method to calculate the solution, and the trigonometric functions and exponents use the Taylor series to find their numerical solutions. Addition, subtraction, and logical calculations are simple
calculations. There is a huge gap between complex calculations and simple calculations on the computer, so simple calculations can be ignored.

Table 1 shows the calculation times of different calculation types for a reconstruction point of CBF, and the total amount of computations added by the combined method. R is the total number of reconstruction points in the acquired target’s three-dimensional surface.

Table 1. Computational complexity of single reconstruction point and the increase of combining stereo vision.

| Calculation Type          | Single Reconstruction Point Computations | Increase Computations |
|---------------------------|-----------------------------------------|-----------------------|
| Index                     | M                                       | 0                     |
| multiplication or division| M (4 + 14 L)                             | 62 R                  |
| Square root               | M                                       | R                     |
| Trig (sin/cos)            | 0                                       | 6                     |
| Trig(tangent)             | 0                                       | 2                     |
| Index                     | M                                       | 0                     |

4.2. Comparative Analysis

To facilitate the comparison between a single CBF reconstruction point and all the increased computations in the combined method, complex calculations are converted into multiplication and division types through numerical calculation principles.

In computer, the Taylor series expansion method is used to calculate the approximate value of exponential and trigonometric functions. Exponent expansion to \( n \) items \((n > 2)\) requires \((n - 1)^2 + 1\) MD operations. The expansion of the trigonometric function sine and cosine to \( n \) terms requires approximately \(2n^2\) times of MD. The tangent is composed of the ratio of sine and cosine, and it requires approximately \(4n^2 + 1\) times of MD [25].

Root calculations use the Newton–Raphson method to continuously iterate to solve: In this type of calculation, in addition to a division operation to find the initial value, approximately \(3n + 1\) times of MD are required for \(n\) iterations [25].

Generally, the accuracy of the numerical solution is 10 decimals: the index needs to expand about 15 items, which converts into 197 times the MD; the trigonometric function needs to expand about 8 items, which is converted into 128 times the MD; tangent is converted into 257 times the MD; when the Newton–Raphson method is used for calculation, when the solution accuracy is 10, about 5 to 6 iterations, and 6 iterations are approximately equivalent to 19 MD.

A reconstruction point in beamforming requires approximately times of multiplication/division operations. The overall increase of the algorithm in this paper is about \(T_{add}\) times the MD, then:

\[
\frac{T_{add}}{T_1} = \frac{62R + 19R + 1282}{197M + M(4 + 14L) + 19M} = \frac{81R + 1282}{M(220 + 14L)}
\]  

(10)

The difference in computation amount mainly depends on the number of reconstruction points R, the number of sound data points L and the number of array channels M of the measured target area.

For example, the number of points in the beamforming scan reconstruction area is 100 \(\times\) 100, \(M = 16, L = 25,600\) Hz/30 \(fps = 853\). If the number of reconstruction points after extracting the target is half of the scanning area \(R = 5000\), the increased calculation amount is only equivalent to 2.1 points. The total calculation amount is reduced by nearly 1/2.
5. Experiment and Results

5.1. Experimental Setup

The purpose of the experiment is to locate the sound source on the experimental platform. The experiment was carried out in a hemi-anechoic room. The experimental bench contains a slowly rotating swing arm driven by a motor and a sound source. The rotation speed of the swing arm under the combined action of the motor and gravity gradually changes from the slow speed of climbing to the fast speed of descending. A Bluetooth speaker with a diameter of 10 cm is installed at the end of the swing arm as the sound source to be located on the experimental bench. The speaker plays a single-frequency sound with a frequency steadily.

The depth images are obtained by Intel® RealSense™ D415 depth camera (Manufactured by Inter Corporation, Santa Clara, CA, USA). As shown in this section and the next, the camera can output a normal image obtained by a red, green and blue (RGB) sensor and a depth image obtained by stereo vision technology at the same time. The key specifications of the depth camera are presented in Table 2.

Table 2. The key specifications of Intel RealSense D415.

| Parameters                       | Value   |
|----------------------------------|---------|
| Depth Stream Output Resolution   | Up to 1280 × 720 |
| Depth Stream Output Frame Rate (Max) | 30 fps |
| Depth Field of View (FOV)        | 69.4 × 42.5° |
| Operating Range (Min-Max)        | 0.16–10 m |
| Distortion                       | ≤1.5%   |
| Baseline (h)                     | 55 mm   |
| Subpixel (p)                     | 0.08    |

The resolution used by the camera in this experiment is 1280 × 720 (rx = 1280, θ_HFOV = 69.4°) and 640 × 480 (rx = 640, θ_HFOV = 65°), and the shooting frame rate is set to 10 fps.

The depth accuracy of the reconstructed point required for beamforming depends on the camera accuracy, and the absolute error e of the measured value depth and the linear relationship of the distance [23].

\[
e_{\text{max}} = \frac{2p_{\text{subpixel}} \tan \left( \frac{\theta_{\text{HFOV}}}{2} \right)}{r_x h_b} \times z^2 + 2.895774 \times 10^{-6} \times z^2 \tag{11}
\]

A spiral array composed of 24 microphones was used in the experiment and the diameter of the array was 0.8 m. The layout of the microphone array is shown in Figure 4. The microphones were MPA416. The sound signal was collected through the data acquisition instrument SCM05 of the LMS company with a sampling rate of 51.2 kHz. The rising edge of an external direct current (DC) signal was used as a trigger so that the image and sound signals were collected at the same time.

The experiment used two layouts to test the effectiveness of the combined method. In layout 1, the plane of the microphone array was parallel to the plane of the test bench and the distance between them was 1.63 m. The camera and microphone array were placed on the lower right of the test bench. The resolution of the camera was 1280 × 720, and the speaker played a single frequency sound of 1 kHz. In layout 2, the center of the microphone array was 1.99 m away from the center of the test bench. The angle between the plane of the moving sound source in the operating mechanism and the array plane was 30° to ensure that the sound source changed in the XYZ direction. The resolution of the camera was 640 × 480, and the speaker played a single frequency sound of 3 kHz. (To simplify the computation of the program, the camera was placed in the same plane of the microphone array at a horizontal distance of 0.16 m from the center of the array). Figures 5 and 6 show the anechoic room experimental layout.
5.2. Localization Result

In the experiment of layout 1, both the video and the acoustic signals were recorded as the sound source rotating from below to above the experimental platform. The signals were also recorded from the right to the left of the experimental platform in layout 2.

Figure 7 shows the two groups of source location images (a, b, c) at different positions in layout 1. Each group of images contains the normal image captured by the optical camera, the environment depth image taken by the stereo camera, and the acoustic beamforming result on the surface of the experimental platform. The set of images (a) is obtained at the start position of layout 1 in this experiment, the sound source localization is obtained by the combination method and the result is (599, 454, 1644). The radiated
sound power level at this position is 49.33 dB. The sound source localization results of the middle position (b) is (663, 384, 1632) and the sound power is 50.96 dB. The sound source localization results of the end position (c) is (726, 58, 1672) and the sound power is 50.50 dB.

![Figure 7](image1.png)  
Figure 7. Sound source localization results of layout 1 (1000 Hz) in the experiment. (a) Images of the Start position. (b) Images of the Middle position. (c) Images of the End position.

Figure 8 also shows the three groups of source location images (a, b, c) at different positions in layout 2. The set of images (a) is obtained at the start position in this experiment, the sound source localization is obtained by the combination method and the result is (497, 197, 1703). The radiated sound power level at this position is 82.84 dB. The sound source localization results of middle position (b) is (306, 21, 1874) and the sound power is 81.76 dB. The sound source localization results of end position (c) is (65, 158, 2060) and the sound power is 81.38 dB. Table 3 shows the positioning results and the corresponding sound power level of two layouts. In the Table, x, y is the pixel coordinate, and z is obtained by indexing their values in the depth image.

Since the experimental platform was separated from the background, the experimental platform area could be easily distinguished from the depth image. In the experiment of layout 2, the actual number of reconstructed points in beamforming was about 22,000 points, and the number of pixels in this area accounted for about 8% of the total area. The combination method in this article added about 2.06 points of calculation in total. At the same time, in this experiment, the average time for the combined method to reconstruct a sound field was recorded as 0.039 s. The average time to reconstruct a 2D sound field using traditional reconstruction methods was 0.523 s. It can be seen that the combined method can effectively reduce the amount of computation.
Figure 8. Sound source localization results of layout 2 (3000 Hz) in the experiment. (a) Images of the Start position. (b) Images of the Middle position. (c) Images of the End position.

The obtained sound source depth was combined with the sound intensity scaling method to achieve the real-time acquisition of sound power. From these two layouts and settings, the position of the camera and array could also be adjusted to select a more suitable resolution in order to obtain better results according to actual needs.

Table 3. Sound power of the localized sound source at different positions.

| Sound Source Localization \( (x, y, z) \) | Radiated Sound Power Level (dB) | Sound Source Localization \( (x, y, z) \) | Radiated Sound Power Level (dB) |
|-----------------------------------------|-------------------------------|-----------------------------------------|-------------------------------|
| Layout 1 (1000 Hz)                      |                               | Layout 2 (3000 Hz)                      |                               |
| 599, 454, 1644                          | 49.33                         | 497, 197, 1703                          | 82.40                         |
| 608, 447, 1644                          | 49.75                         | 480, 186, 1708                          | 82.08                         |
| 623, 428, 1642                          | 50.16                         | 464, 189, 1724                          | 81.94                         |
| 639, 414, 1634                          | 50.50                         | 457, 102, 1763                          | 81.79                         |
| 658, 392, 1639                          | 50.87                         | 414, 65, 1771                           | 81.91                         |
| 670, 377, 1628                          | 51.09                         | 398, 54, 1774                           | 81.79                         |
| 684, 364, 1632                          | 51.26                         | 365, 38, 1827                           | 81.58                         |
| 695, 357, 1618                          | 51.17                         | 344, 28, 1852                           | 81.36                         |
| 708, 342, 1621                          | 51.04                         | 327, 23, 1884                           | 81.71                         |
| 717, 327, 1623                          | 50.79                         | 306, 21, 1874                           | 81.76                         |
| 741, 238, 1634                          | 49.20                         | 287, 23, 1903                           | 82.08                         |
5.3. Sound Power Results

In the experiment, the sound field on the surface of the experimental platform at each frame was reconstructed. Figures 9–11 show the monitoring results of moving sound sources during operation. According to the rotation angle of the swing arm, the movement speed of the sound source is from slow to fast.

The change of the calculated depth of sound source with time-angle is shown in Figure 9. The x-coordinate in these figures is the time and the corresponding angle of the swing arm calculated from the localization of the proposed method. The depth of the sound source varies from 1598 to 1665 mm in layout 1. The sound source depth value is stable at about 1630 mm, which is consistent with the layout.

The depth of the sound source varies from 1651 to 2060 mm in layout 2, the sound source depth changes from far to near, and then from near to far. Although the swing arm shakes randomly and the camera has imaging errors, it can still be seen that this trend is consistent with the trend of the theoretical depth curve of the sound source.

Figure 9. Sound source depth versus time-angle.

Figure 10 shows the curve of the sound pressure level (SPL) of the channel at the center of the array versus time-angle. The SPL of layout 1 is within 27.36–29.43 dB, and the range of layout 2 is 53.42–59.48 dB. In layout 1, since the sound source depth is relatively stable, the value of SPL is also relatively stable. Due to the direction of the source and the reflection from the floor, when the Bluetooth speaker at the end of the swing arm rotates to certain angles, the sound pressure level of some channels in the array changes significantly in layout 2 (e.g., the sound pressure level at 1.2 s). However, according to the sound pressure level fitting curve, it can be seen that the overall trend is decreasing with the increase of the distance.
Figure 10. Sound pressure level on one of the microphones versus time-angle.

By combining the sound source depth change and the sound intensity scaling method, the sound power level versus time-angle curve in Figure 11 is obtained. The sound power level range of layout 1 is stable at 81.02–82.48 dB. The sound power level is within 81.02–82.48 dB in layout 2, and is more stable than the sound pressure level. The sound intensity scaling method is to scale the entire array signal, and the influence of abnormal signals of certain microphones is reduced on the sound intensity of the entire array.

Figure 11. Calculated sound power level versus time-angle.
5.4. Discussion

The accuracy of the three-dimensional surface of the target is the basis for the accurate reconstruction of the 3D sound field, so the measurement error of the stereo camera affects the reconstruction area. The error between the point cloud and the real surface of the target includes the offsets on horizontal, vertical and depth directions. The depth of each pixel is calculated from the horizontal pixels of the two cameras, and this ensures that the horizontal error is smaller than the depth error. Although the accuracy of modern stereo cameras has reached a very high level, these errors may still need to be considered in some scenes with high accuracy requirements. At the same time, the error relationship between the depth error and the actual distance is given in Equation (11). It is necessary to pay attention to these key camera parameters.

Moreover, since the reconstruction of the three-dimensional sound field is limited on the surface of the target, the sound source inside the target cannot be located accurately. A manual adjustment on the depth of the surface may have to be undertaken.

It is worth noting that the recognition area (both depth and field of view) of the combined method is limited by both the stereo camera and the beamforming systems. Taking the effective recognition depth in the experiment as an example, the recognizable depth of the camera is 0.2~10 m, and the recognition depth of the microphone array is 0.3~5 m, so the depth recognition range of the entire system is 0.3~5 m.

This study uses conventional beamforming (CBF) to eventually locate the sound source. At present, many advanced algorithms have been proposed based on CBF to achieve higher precision in the positioning or better interference suppression. The CBF in the proposed fusion method can also be replaced with those methods to take advantage of them.

Furthermore, the sound intensity scaling method used to estimate sound power in this study is derived mathematically based on the physical propagation model of the point sound source. It is not suitable for accurate sound power estimation of multiple sound sources or directional sound sources.

In the practice, the camera can be built into the array to avoid manual measurement of errors between the camera and the array. At the same time, it will bring an advantage that no additional visual and acoustic registration is required when reconstructing the three-dimensional sound field distribution.

6. Conclusions

Aimed at the current difficulties in spatial localization and monitoring of the sound power of mobile sound sources, this paper studies the sound source localization method combining a binocular camera and beamforming. This method scans scene information through binocular cameras and outputs depth images, and obtains the structural information of the target area through image segmentation. As the range of the reconstructed sound field is limited on the three-dimensional surface of the target, the combination method can not only realize the three-dimensional localization of the sound source with less computation, but also achieve the sound field reconstruction of the moving target by updating the three-dimensional information of the target continuously. Locating the moving sound source spatially provides the depth variation of the source, and the sound power of the source can be estimated by the sound intensity scaling method. The sound power level is more stable than sound pressure level to monitor the condition of the sound source.

It was proved by a mathematical derivation and experiment that the use of stereo technology to reduce the number of reconstruction points can effectively improve computational efficiency. With mature technology and a high precision of stereo vision, the accuracy of sound source focusing and reconstruction is improved, which is conducive to obtaining accurate source positioning and sound power value.
Author Contributions: Conceptualization, Y.C. and X.L.; methodology, Y.C., X.L. and Y.X.; software, Y.C. and Y.X.; validation, Y.C., X.L. and Y.X.; writing—review and editing, X.L.; supervision, X.W.; project administration, X.W.; funding acquisition, X.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by National Natural Science Foundation of China (51875272); Science and Technology Major Project of Yunnan Province (202002AC080001).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

Conflicts of Interest: The authors declare no conflict of interest.

References
1. Li, J.; Bai, Y.; Zhang, Y.; Qua, F.; Wei, Y.; Wang, J. Cross power spectral density based beamforming for underwater acoustic communications. Ocean Eng. 2020, 216, 107786, doi:10.1016/j.oceaneng.2020.107786.
2. Asim, F.E.; Antreich, F.; Cavalcante, C.C.; de Almeida, A.L.F.; Nossek, J.A. Channel parameter estimation for millimeter-wave cellular systems with hybrid beamforming. Signal Process. 2020, 176, 107715, doi:10.1016/j.sigpro.2020.107715.
3. Wiacek, A.; Olywemi, E.; Myers, K.; Mullen, L.; Bell, M.A.L. Coherence-based beamforming increases the diagnostic certainty of distinguishing fluid from solid masses in breast ultrasound exams. Ultrasound Med. Biol. 2020, 46, 1380–1394, doi:10.1016/j.ultrasmedbio.2020.01.016.
4. Michel, U. History of acoustic beamforming. In Proceedings of the 1st Berlin Beamforming Conference, Berlin, Germany, 22–23 November 2006.
5. Chiarotti, P.; Martarelli, M.; Castellini, P. Acoustic beamforming for noise source localization—Reviews, methodology and applications. Mech. Syst. Signal Process. 2019, 120, 422–448, doi:10.1016/j.ymssp.2018.09.019.
6. Merino-Martínez, R.; Snellen, M.; Simons, D.G. Functional beamforming applied to imaging of flyover noise on landing aircraft. J. Aircr. 2016, 53, 1830–1843, doi:10.2514/1.c033691.
7. Barsikow, B. Experiences with various configurations of microphone arrays used to locate sound sources on railway trains operated by the DB AG. J. Sound Vib. 1996, 193, 283–293, doi:10.1006/jsvi.1996.0269.
8. Padois, T.; Berry, A. Application of acoustic imaging techniques on snowmobile pass-by noise. J. Acoust. Soc. Am. 2017, 141, EL134–EL139, doi:10.1121/1.4976138.
9. Christensen J J; Hald J. Beamforming. B&K Tech. Rev. 2004, 1, 1–31.
10. Döbler, D.; Ocker, J.; Puhle, C. On 3D-beamforming in the wind tunnel. In Proceedings of the 6th Berlin Beamforming Conference, Berlin, Germany, 29 February–1 March 2016.
11. Legg, M.; Bradley, S. Automatic 3D scanning surface generation for microphone array acoustic imaging. Appl. Acoust. 2014, 76, 230–237, doi:10.1016/j.apacoust.2013.08.008.
12. Bub, U.; Hunke, M.; Waibel, A. Knowing who to listen to in speech recognition: visually guided beamforming. In Proceedings of the 1995 International Conference on Acoustics, Speech, and Signal Processing, Washington, DC, USA, 9–12 May 1995; Volume 841, pp. 848–851.
13. Best, V.; Roverud, E.; Streeter, T.; Mason, C.R.; Kidd, G. The benefit of a visually guided beamformer in a dynamic speech task. Trends Hear. 2017, 21, doi:10.1177/2331216517722304.
14. Jennings, T.R.; Kidd, G. A visually guided beamformer to aid listening in complex acoustic environments. Proc. Meet. Acoust. 2018, 33, 050005, doi:10.1121/2.0000972.
15. Novoa, J.; Mahu, R.; Diaz, A.; Wuth, J.; Stern, R.; Yoma, N.B. Weighted delay-and-sum beamforming guided by visual tracking for human-robot interaction. arXiv 2019, arXiv:1906.07298.
16. Donovan, A.O.; Duraiswami, R.; Zotkin, D. Imaging concert hall acoustics using visual and audio cameras. In Proceedings of the 2008 IEEE International Conference on Acoustics, Speech and Signal Processing, Hannover, Germany, 31 March–4 April 2008; pp. 5284–5287.
17. Campero, A.; Baldoncini, M.; Villalonga, J.F.; Abarca-olivias, J. Three-dimensional microscopic surgical videos: A novel and low-cost system. World Neurosurg. 2019, 132, 188–196, doi:10.1016/j.wneu.2019.08.139.
18. Fu, K.; Xie, Y.; Jing, H.; Zhu, J. Fast spatial-temporal stereo matching for 3D face reconstruction under speckle pattern projection. Image Vis. Comput. 2019, 85, 36–45, doi:10.1016/j.imavis.2019.02.007.
19. Jones, C.B.; Goates, C.B.; Blotter, J.D.; Sommerfeldt, S.D. Experimental validation of determining sound power using acoustic radiation modes and a laser vibrometer. Appl. Acoust. 2020, 164, 107254, doi:10.1016/j.apacoust.2020.107254.
20. Zhong, J.; Tao, J.; Niu, F.; Qiu, X. Effects of a finite size reflecting disk in sound power measurements. Appl. Acoust. 2018, 140, 24–29, doi:10.1016/j.apacoust.2018.05.007.
21. Hald, J. Estimation of partial area sound power data with beamforming. In Proceedings of the 2005 Congress and Exposition on Noise Control Engineering, Rio de Janeiro, Brazil, 7–10 August 2005.
22. Zhao, S.; Kang, F.; Li, J. Displacement monitoring for slope stability evaluation based on binocular vision systems. *Optik* 2018, 171, 658–671, doi:10.1016/j.ijleo.2018.06.097.

23. Liu, S.; Gao, D.; Wang, P.; Guo, X.; Xu, J.; Liu, D.-X. A depth-based weighted point cloud registration for indoor scene. *Sensors* 2018, 18, 3608, doi:10.3390/s18113608.

24. Sawada, T.; Zaidi, Q. Rotational-symmetry in a 3D scene and its 2D image. *J. Math. Psychol.* 2018, 87, 108–125, doi:10.1016/j.jmp.2018.10.001.

25. Faires, J.D.; Burden, R.L. *Numerical Analysis*; Brooks Cole: Belmont, CA, USA, 2004.