sensors

MDPI

Article

An Outlier Detection Method Based on Mahalanobis Distance for Source Localization

Qingli Yan [1,2,*], Jianfeng Chen [1] and Lieven De Strycker [2]

[1] School of Marine Science and Technology, Northwestern Polytechnical University, Xi'an 710072, China; chenjf@nwpu.edu.cn

[2] KU Leuven, ESAT-DRAMCO, Ghent Technology Campus, 9000 Ghent, Belgium; lieven.destrycker@kuleuven.be

* Correspondence: gongchyy@163.com; Tel.: +86-136-5919-3864

Received: 20 June 2018; Accepted: 5 July 2018; Published: 7 July 2018

Abstract: This paper addresses the problem of localization accuracy degradation caused by outliers of the angle of arrival (AOA). The problem of outlier detection of the AOA is converted into the detection of the estimated source position sets, which are obtained by the proposed division and greedy replacement method. The Mahalanobis distance based on robust mean and covariance matrix estimation method is then introduced to identify the outliers from the position sets. Finally, the weighted least squares method based on the reliable probabilities and distances is proposed for source localization. The simulation and experimental results show that the proposed method outperforms representative methods when unreliable AOAs are present.

Keywords: angle of arrival; source localization; outlier detection; Mahalanobis distance; unreliable nodes

1. Introduction

The source localization techniques based on the angle of arrival (AOA) estimate the target position using a set of estimated bearings. Various methods have been proposed to solve the localization problem [1–3], among which the closed-form method of the pseudolinear estimator (PLE) was proposed under the assumption that AOA errors are small [4]. Although the PLE method is easy to implement and is efficient to compute, it is sensitive to outliers (AOAs with large errors), and such outliers (also referred to as unreliable AOAs) may exist in many practical applications of distributed node networks, which typically consist of a large number of small, low-cost sensor nodes. When nodes are deployed in harsh and unattended environments, animal attack or other forms of interference may occur. Moreover, low-cost nodes have limited amounts of power, computational, and memory capacity, and these limitations may also cause outliers. Other factors, such as node failures, data loss, and non-line of sight (NLOS) propagation [5,6], can lead to unreliable measurements. As a result, the estimated AOAs at each node will deviate significantly from the true values. Such outliers have been found to be detrimental to the PLE [7,8]. Thus, it is important to identify these erroneous data to improve the localization performance or perform a repair of the data.

To reduce the error induced by outliers in node networks, several hybrid localization methods have been proposed by combining the AOA with the time difference of arrival (TDOA) and received signal strength (RSS) to identify and mitigate the NLOS error [9]. The expectation maximization (EM) method is introduced to identify unreliable AOAs caused by NLOS [10]. The intersection points (IPs)-based method [11] calculates the source position by taking the centroid of the set of intersections obtained by pairs of bearing lines; however, this method cannot significantly improve the localization performance, even eliminating the IPs obtained by two bearing lines close to parallel. The proposed unreliable AOA detection method in [7] can improve the localization accuracy; however, many

threshold parameters need to be set. The steered-response power phase transform (SRP-PHAT) [12] source localization approaches have demonstrated robustness when operating in reverberant and noisy environments. Regardless, these methods require a considerably higher amount of information to be transmitted to the central processing node and cannot be applicable to large-range localization scenarios (e.g., hundreds or thousands of meters). In this work, every node is equipped with a microphone array to estimate the AOA and then transmits the estimated AOA to the central node. This method does not require time synchronization in different nodes. Note that AOA estimation methods under an environment with complex environmental noise are outside the scope of this paper; interested readers are referred to [13–15]. These robust AOA estimation methods are proposed under the assumption that only small portions of snapshots are contaminated; they can perform well for continuous source signals or impulsive interference noise, which only have influence on limited snapshots. However, with other causes that can last a period of time, such as sensor failures and non-line of sight (NLOS) propagation, all snapshots for one node are unreliable; thus, outliers that may deteriorate the localization performance are still present even when these methods are applied to estimate AOAs under complex noise. Therefore, the outlier detection for the AOA is still necessary to improve the localization accuracy.

Here, we propose a robust localization method when outliers are present. A large number of positions can be obtained by different node combinations. The maximum number of estimated positions is $N \times (N - 1)/2$ for an N-node network. However, the estimated positions are sensitive to the bearing lines and their differences. Deleting the outliers from all intersections alone cannot significantly improve the localization performance [11]. To increase the estimated position reliability and improve the detection accuracy, we propose the division and greedy replacement (DIG) method to obtain different estimated position sets by changing one node at one time. The robust estimation method of the mean and covariance matrix for estimated position sets is then addressed to provide the information for outlier detection. The Mahalanobis distance (MD) [16] is finally proposed to identify the outliers from the estimation position sets. Finally, the weighted least squares (WLS) method based on detected reliable probabilities and distances is used to estimate the source position. The proposed method is easy to implement and can be easily extended to a three-dimensional (3D) source localization method. The main contributions of this paper can be summarized as follows:

- The division and greedy replacement (DIG) method is developed to estimate the target positions.
- The Mahalanobis distance based on robust estimation of mean and covariance matrix is proposed to detect the outliers from estimated source positions.
- An improved WLS localization method based on reliable probabilities and distances is introduced.
- Outdoor experiments are conducted to verify the proposed method.

The remainder of this paper is organized as follows. Section 2 describes the AOA localization method and addresses the existing problem. The unreliable node detection method is proposed in Section 3. Simulations and experimental results are presented in Sections 4 and 5, respectively. Finally, our work is summarized in Section 6.

2. AOA-Based Localization Method and Problem Statement

2.1. The Pseudolinear Estimator (PLE)

We consider N nodes equipped with a microphone array for each one, with known positions $\mathbf{s}_k = [x_k, y_k]^{\mathrm{T}}$ (k = 1, 2, ... , N), are deployed in an area of interest to estimate the location of a single source $\mathbf{p} = [x, y]^{\mathrm{T}}$ as shown in Figure 1. Under the Gaussian background noise assumption, the estimated angle $\hat{\theta}_k$ of k-th node can be given by the following:

$$\hat{\theta}_k = \theta_k + \eta_k, \tag{1}$$

where

$$\theta_k = \arctan(\frac{y - y_k}{x - x_k}), \tag{2}$$

and η_k is the zero mean Gaussian noise with variance σ_i^2.

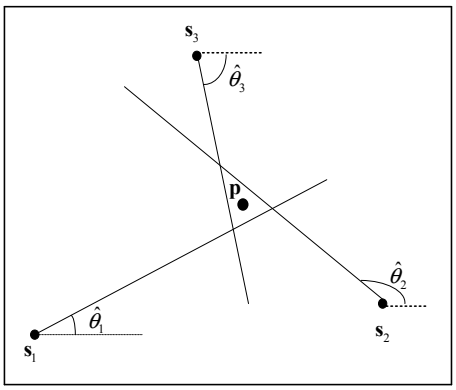

Figure 1. Illustration of angle of arrival (AOA) localization with three nodes deployed in the test area.

The set of measurements from N nodes can be written as follows:

$$\hat{\boldsymbol{\theta}} = \boldsymbol{\theta} + \boldsymbol{\eta}, \tag{3}$$

where $\hat{\boldsymbol{\theta}} = [\ \hat{\theta}_1 \quad \hat{\theta}_2 \quad \dots \quad \hat{\theta}_N\ ]^{\mathrm{T}}$, $\boldsymbol{\theta} = [\ \theta_1 \quad \theta_2 \quad \dots \quad \theta_N\ ]^{\mathrm{T}}$, and $\boldsymbol{\eta} = [\ \eta_1 \quad \eta_2 \quad \dots \quad \eta_N\ ]^{\mathrm{T}}$. Thus, the pseudolinear estimator (PLE), also known as the orthogonal vectors (OV) estimator, can be used to estimate the source position and is given by the following [4]:

$$\mathbf{Ap} = \mathbf{B} + \mathbf{e}, \tag{4}$$

where the estimated source position is as follows:

$$\hat{\mathbf{p}} = (\mathbf{A}^{\mathrm{T}}\mathbf{A})^{-1}\mathbf{A}^{\mathrm{T}}\mathbf{B}, \tag{5}$$

where the k-th row of matrix $\mathbf{A}$ and $\mathbf{B}$ is $\mathbf{A}(k,:) = [\ \sin\hat{\theta}_k \quad \cos\hat{\theta}_k\ ]$, $\mathbf{B}(k,:) = x_k \sin\hat{\theta}_k - y_k \cos\hat{\theta}_k$, $k = 1, 2, \dots, N$, and

$$\mathbf{e} = [\ r_1 \sin\eta_1 \quad r_2 \sin\eta_2 \quad \dots \quad r_N \sin\eta_N\ ]^{\mathrm{T}}, \tag{6}$$

where r_k is the distance between the source and node $\mathbf{s}_k$.

2.2. Problem Formulation

The PLE is easy to implement, even for large-scale data. However, the PLE is sensitive to unreliable measurements (i.e., outliers). In this section, we use theoretical analysis and simulation results to illustrate this problem.

If the measurement error is sufficiently small, then we have $\sin\eta_k \approx \eta_k$. Thus, the approximation of the residuals of Equation (6) can be expressed as $e_k \approx r_k\eta_k$. The estimated error of the source position can be expressed as follows [8]:

$$\begin{aligned} \Delta\mathbf{p} &= \hat{\mathbf{p}} - \mathbf{p} \\ &= (\mathbf{A}^{\mathrm{T}}\mathbf{A})^{-1}\mathbf{A}^{\mathrm{T}}\mathbf{B} - (\mathbf{A}^{\mathrm{T}}\mathbf{A})^{-1}\mathbf{A}^{\mathrm{T}}\mathbf{Ap} \\ &= (\mathbf{A}^{\mathrm{T}}\mathbf{A})^{-1}\mathbf{A}^{\mathrm{T}}(-\mathbf{e}) \end{aligned}. \tag{7}$$

The covariance matrix of Equation (7) can be obtained by the following:

$$\mathrm{cov}(\Delta \mathbf{p}) = E(\Delta \mathbf{p} \Delta \mathbf{p}^{\mathrm{T}}). \tag{8}$$

Thus, the mean-square error (MSE) is given by the following:

$$MSE = \mathrm{tr}[\mathrm{cov}(\Delta \mathbf{p})]. \tag{9}$$

Submitting Equations (6) and (7) into Equation (10), we have the following:

$$\begin{aligned} MSE \quad &= \frac{1}{\sum\limits_{i,j \in s} \sin^2(\theta_i - \theta_j)} \\ &\cdot \sum_{i \in s} \left\{ \sigma_{e_i}^2 \left[(f_{11} \sin\theta_i - f_{12} \cos\theta_i)^2 + (f_{21} \sin\theta_i - f_{22} \cos\theta_i)^2 \right] \right\}, \end{aligned} \tag{10}$$

where S is defined as all the combinations if $\{i, j\}$ with $j > i$. $f_{11} = \sum\limits_{i \in s} \sin^2\theta_i$, $f_{22} = \sum\limits_{i \in s} \cos^2\theta_i$ and $f_{21} = f_{12} = \sum\limits_{i \in s} \cos\theta_i \sin\theta_i$ and $\sigma_{e_i}^2 = E(\mathbf{e}\mathbf{e}^{\mathrm{T}})$.

We can see from Equation (10) that the MSE is affected by the relative geometry between the source and the nodes, the number of nodes, and the AOA measurement errors. To illustrate the effect of outliers, we conducted several simulations to analyze the characteristics of the localization error for different source positions. The source is assumed to be located at the gridded points, ranging from -10 m to 10 m in a 20×20 m^2 grid with a resolution of 0.5 m. Four nodes—s_1, s_2, s_3, and s_4—are randomly deployed in the test area, as shown in Figure 1. The root-mean-square error (RMSE) of the PLE of 500 trials for every target position is used as the performance metrics.

The RMSEs for different source positions are shown in Figure 2a when $\sigma_1 = \sigma_2 = \sigma_3 = \sigma_4 = 1^\circ$. It is clear that the localization errors are relatively lower when the source is surrounded by the nodes compared to the outside source. The conclusion follows the analysis based on the Cramer–Rao lower bound (CRLB) in [17].

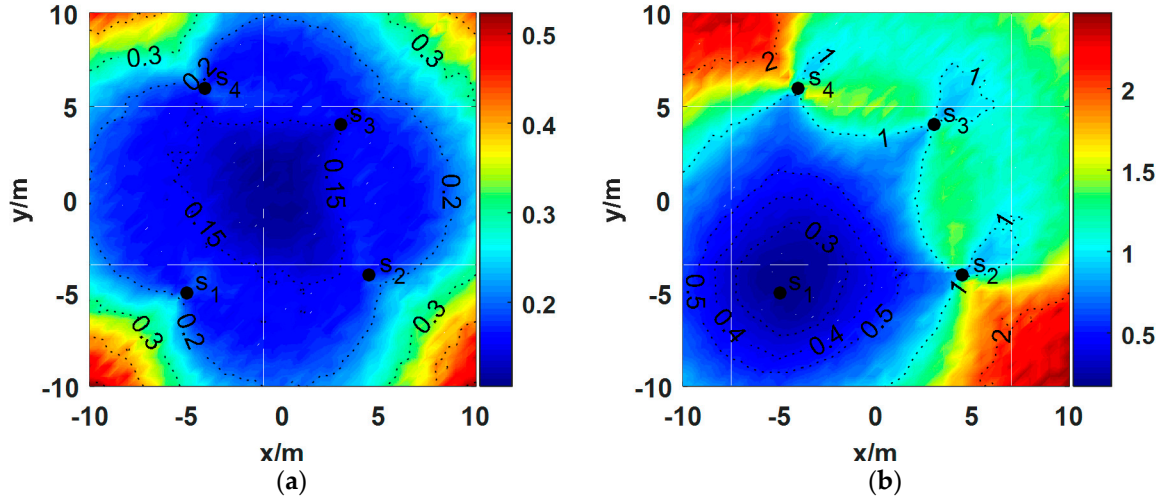

Figure 2. Root-mean-square error (RMSE) for different source positions when four nodes are deployed in a 20×20 m^2 test area: (**a**)$\sigma_1 = \sigma_2 = \sigma_3 = \sigma_4 = 1^\circ$; (**b**) $\sigma_1 = 10^\circ, , \sigma_2 = \sigma_3 = \sigma_4 = 1^\circ$. The black dots denote the nodes, and the numbers on the dotted line contours are the values of the RMSEs.

Assume that the unreliable node $\mathbf{s}_1$ is subject to a large noise with zero means $\sigma_1 = 10^\circ$ and $\sigma_2 = \sigma_3 = \sigma_4 = 1^\circ$. The resulting RMSEs are plotted in Figure 2b. When the source is close to the unreliable node, the localization accuracy is not significantly deteriorated. However, the RMSEs are significantly increased when the source is far from the unreliable nodes. From Equation (10), we can

see that for the same AOA estimation error σ_i, the MSE is mainly influenced by the distance r_k between the source and the node $\mathbf{s}_k$.

To demonstrate the importance of detecting unreliable nodes, the RMSEs of the estimated positions obtained from the four nodes with one being unreliable are compared with the RMSEs when only three reliable observations are used for the source located at $\mathbf{p} = [3, 0]^{\mathrm{T}}$ m. As shown in Table 1, the localization errors obtained using only three reliable nodes are significantly lower than those obtained with four nodes, one of which is unreliable. Therefore, it is necessary to detect the unreliable nodes and then remove them to improve the localization accuracy.

Table 1. RMSE for different numbers of nodes.

Unreliable Node	s_1	s_2	s_3	s_4
RMSE (m): $N = 4$	1.2201	0.3168	0.3146	0.8060
RMSE (m): $N = 3$	0.2155	0.1642	0.1695	0.1780

3. The DIG_MD Method

We know that at least two nonparallel bearing lines are required to estimate an IP, and the maximum number of IPs for N nodes is $N(N-1)/2$. Regardless of the parallel cases of two bearing lines, all IPs are expected to be close to each other and surround the true source position when they are only subjected to low-level environment noise. In contrast, the bearings corrupted by large noise will lead the IPs to be far from the source position. As shown in Figure 3, the IPs obtained from s_1 are obviously far from the other IPs. Therefore, we can identify the unreliable AOAs by detecting outliers from the estimated target positions. However, there are too many intersections to calculate for large-scale node networks if only two bearings are used. Moreover, the IPs are also easily affected by the errors of either one and by the angular distance [11]. For example, it is easy to cause a false alarm if s_6 is determined to be unreliable when $\mathbf{p}_{16}$ and $\mathbf{p}_{36}$ are detected as outliers. To solve these problems, the division and greedy replacement (DIG) method is proposed here to improve the stability of the estimated positions. The two-dimensional (2D) outlier detection method is then used to find the unreliable bearings. Finally, the WLS based on detected reliable probabilities and distances from initial position to nodes is used to perform the localization. The procedure is given as follows:

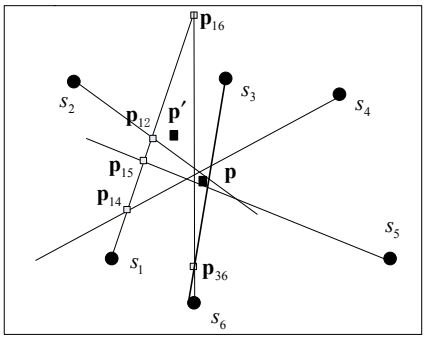

Figure 3. Illustration of the intersection points (IPs) distribution when one bearing is unreliable, where the rectangle denotes the IP and $\mathbf{p}_{ij}$ is the intersection point obtained by $\mathbf{s}_i$ and $\mathbf{s}_j$.

3.1. The Division and Greedy Replacement (DIG) Method

In order to detect outliers from the AOAs, based on estimated source positions, a set of position estimations are needed, which should be calculated by a fixed number of nodes only with one independent variable. Thus, every estimated position corresponds to the unique different node. In this paper, we propose to divide all nodes into two sets, and the greedy replacement is then used to obtain different combinations of a fixed number of nodes with one difference.

(1) Division: In this section, the two separated set are defined as the reference node set (Ω_{ref}) and the replacement node set (Ω_{rep}), with sizes m and $N-m$, respectively. Here, to provide an easier explanation, we assume that the reference nodes are indexed from 1 to m. Thus, Ω_{ref} and Ω_{rep} can be denoted as $\Omega_{ref} = \{\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_m\}$ ($m \geq 3$) and $\Omega_{rep} = \{\mathbf{s}_{m+1}, \mathbf{s}_{m+2}, \ldots, \mathbf{s}_N\}$, respectively. Algorithm 1 presents the selection method of reference nodes:

Algorithm 1. Selection method of reference nodes.

(1) Estimate the initial source position $\mathbf{p}'$ by the PLE based on all measurements;
(2) Calculate the distances from $\mathbf{p}'$ to all nodes;
(3) Select m nodes that have short distances and can form a convex polygon with the target inside.

The performance analysis in [17] shows that the nodes that are close to the target are dominant in the localization results and that the localization error for the target inside a convex polygon composed of multiple nodes is smaller than that of an outside one. So, we propose to use the nodes that can comprise a convex polygon with the target inside and have short distances to the source as reference nodes, as shown in step 3; thus, no fewer than three nodes should be selected as the reference nodes. As the true position is unknown, an initial obtained from all the measurements can be used to evaluate the distances stated as step 1 and step 2. As shown in Figure 3, $\mathbf{p}'$ is the initial position calculated by six measurements; s_1, s_2, and s_3 are closest to $\mathbf{p}'$, and $\mathbf{p}'$ is inside the convex polygon formed by the three nodes. Thus, s_1, s_2, and s_3 are selected as reference nodes (i.e., $\Omega_{ref} = \{\mathbf{s}_1, \mathbf{s}_2, \mathbf{s}_3\}$ when m = 3). The detail of the division method can be summarized as follows.

To identify the unreliable bearings by detecting outliers from a set of estimated positions, we obtain the positions by changing only one node at a time. As noted above, the localization error is sensitive to the bearing error of the nodes relatively far from the source. Thus, we design the greedy replacement method by using each node in Ω_{rep} to replace one of those in Ω_{ref}. The procedure is given by Algorithm 2. Every node in Ω_{ref} is replaced by $(N-m)$ nodes from Ω_{rep}. Next, m sets, including $(N-m)$ positions in each set, can be obtained, and every point is calculated by m nodes with $(m-1)$ same nodes from Ω_{ref}. For this method, the position sets can be calculated with cost $(-m^2+mN)$. In contrast, the cost is $[N(N-1)/2]$ if all IPs are estimated. In general, the DIG method is computationally simpler than the IP method.

Algorithm 2. Greedy Replacement Method.

For k = 1:m
For j = m + 1:N
$\Omega = \Omega_{ref} \cup \{\mathbf{s}_j\} - \{\mathbf{s}_k\}$
$\mathbf{p}_{k,j} = (\mathbf{A}(\Omega)^{\mathrm{T}}\mathbf{A}(\Omega))^{-1}\mathbf{A}(\Omega)^{\mathrm{T}}\mathbf{B}(\Omega)$
$\mathbf{P}_k(j,:) = \mathbf{p}_{k,j}^{\mathrm{T}}$
end
end

In this paper, $X \cup Y$ and $X - Y$ denote the union and difference of sets X and Y, respectively; $\mathbf{A}(\Omega)$ represents the matrix $\mathbf{A}$ in Equation (4) calculated based on the nodes from set Ω; and $\mathbf{P}_k(j,:)$ is the j-th row of $\mathbf{P}_k$. Each element $\mathbf{p}_{k,j}$ of $\mathbf{P}_k$ is the estimated position using $\mathbf{s}_j, j \in \{m+1, \ldots, N\}$ to replace $s_k, k \in \{1, \ldots, m\}$.

3.2. Outlier Detection Method for Estimated Target Position Sets

All the position elements in $\mathbf{P}_k = [\mathbf{p}_{k,m+1}, \mathbf{p}_{k,m+2}, \ldots, \mathbf{p}_{k,n}]^{\mathrm{T}}, k = 1, 2, \ldots, m$ should be close to each other under the assumption that all the nodes are reliable. The outlier positions should be obtained from the unreliable nodes. For the 2D source localization problem, the elements in $\mathbf{P}_k$ are identically distributed 2D random vectors with mean $\boldsymbol{\mu}_k$ and a positive-definite covariance matrix $\boldsymbol{\Sigma}_k$. To identify

the unreliable nodes in set Ω_{rep}, the square of the Mahalanobis distance (MD) [18–20], which can be formulated as in Equation (11), is proposed to detect outliers from the position matrix in $\mathbf{P}_k$ as follows:

$$d_{k,j}^2(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) = (\mathbf{p}_{k,j} - \boldsymbol{\mu}_k)^{\mathrm{T}} \boldsymbol{\Sigma}_k^{-1} (\mathbf{p}_{k,j} - \boldsymbol{\mu}_k) \tag{11}$$

In the field of data statistics, MD is typically used to characterize how far a particular datum is from the center. A point with a distance greater than a predetermined threshold is assumed to be an outlier. The outlier detection problem in this work is a 2D data detection problem. Therefore, the robust estimated method of $\boldsymbol{\Sigma}_k$ and $\boldsymbol{\mu}_k$ is important for robust outlier detection. The outlier detection method for one position set $\mathbf{P}_k$ is given by Algorithm 3.

To better explain Algorithm 3, let us recall the Gnanadesikan Kettenring (GK) estimator first [18], which provides a reasonable relationship between variance and covariance. Assume that V is the covariance matrix of L-dimensional random vector $\mathbf{x}$ and $\sigma(\cdot)$ represents the standard deviation; thus, we have

$$\sigma(\mathbf{c}^{\mathrm{T}}\mathbf{x})^2 = \mathbf{c}^{\mathrm{T}}\mathbf{V}\mathbf{c} \tag{12}$$

for all $\mathbf{c} \in R^L$. The GK estimator can be formulated as the following:

$$\mathrm{cov}(\mathbf{x}, \mathbf{y}) = \frac{1}{4}\Big(\sigma(\mathbf{x}+\mathbf{y})^2 - \sigma(\mathbf{x}-\mathbf{y})^2\Big), \tag{13}$$

where $\mathbf{x}$ and $\mathbf{y}$ are a pair of random vectors.

Algorithm 3. Outlier detection method from $\mathbf{P}_k$.

Step 1. Let $\mathbf{D} = \mathrm{diag}(\sigma(\mathbf{P}_k(:,1)), \sigma(\mathbf{P}_k(:,2)))$, and $\mathbf{M}_k = \mathbf{P}_k\mathbf{D}^{-1}$.

Step 2. Compute the correlation matrix $\boldsymbol{\Psi}$, $\boldsymbol{\Psi}_{11} = \boldsymbol{\Psi}_{22} = 1$, and $\boldsymbol{\Psi}_{12} = \boldsymbol{\Psi}_{21} = \frac{1}{4}[\sigma(\mathbf{M}_k(:,1) + \mathbf{M}_k(:,2))^2 - \sigma(\mathbf{M}_k(:,1) - \mathbf{M}_k(:,2))^2]$.

Step 3. Compute the matrix $\mathbf{E}$ whose columns are the eigenvectors of $\boldsymbol{\Psi}$, and $\boldsymbol{\Psi} = \mathbf{E}\boldsymbol{\Lambda}\mathbf{E}^{\mathrm{T}}$, where $\Lambda = \mathrm{diag}(\lambda_1, \lambda_2)$, and λ_i are the eigenvalues.

Step 4. Let $\mathbf{G} = \mathbf{D}\mathbf{E}$ and $\mathbf{Z}_k(j,:) = \left(\mathbf{G}^{-1}\mathbf{p}_{k,j}\right)^{\mathrm{T}}$, and $\mathbf{v} = [\mu(\mathbf{Z}_k(:,1)), \mu(\mathbf{Z}_k(:,2))]^{\mathrm{T}}$, and define $\boldsymbol{\Sigma}_k \leftarrow c_1\boldsymbol{\Sigma}_k{}'$, and $\boldsymbol{\mu}_k \leftarrow c_2\boldsymbol{\mu}_k{}'$, where $\boldsymbol{\Sigma}_k{}' = \mathbf{G}\boldsymbol{\Gamma}\mathbf{G}^{\mathrm{T}}$, $\boldsymbol{\mu}_k{}' = \mathbf{G}\mathbf{v}$ and $\boldsymbol{\Gamma} = \mathrm{diag}(\sigma(\mathbf{Z}_k(:,1))^2, \sigma(\mathbf{Z}_k(:,2))^2)$,

Step 5. Calculate the square of MD $d_{k,j}^2$ based on Equation (11) with a threshold of $d_{k0}^2 = \chi_2^2(\alpha)$.

Step 6. If $d_{s_{k,j}} > d_{k0}$, $\mathbf{p}_{k,j}$ is an outlier. Thus, the unreliable probability for s_j is 1/m; otherwise, it is 0.

In Algorithm 3, $med(\cdot)$ represents the median value, $\chi_p^2(\alpha)$ is the α-quantile of the chi-squared distribution with p degrees of freedom, $\mathrm{diag}(\cdot)$ is the diagonal matrix, and $\sigma(\cdot)$ and $\mu(\cdot)$ denote the univariate standard deviation and average value, respectively. c_1 and c_2 are a constant. $\boldsymbol{\Sigma}_k{}'$ and $\boldsymbol{\mu}_k{}'$ are the estimations of $\boldsymbol{\Sigma}_k$ and $\boldsymbol{\mu}_k$.

Steps 1–4 in Algorithm 3 provide a method to obtain the positive-definite and approximately equal-variant covariance matrix $\boldsymbol{\Sigma}_k$ for high-dimensional scatter datasets with much shorter computing times [19]. The first step in Algorithm 3 makes the position vector scale-equivariant for different dimensions. Then, the GK estimator is used to calculate the covariance matrix $\boldsymbol{\Psi}$ in step 2. However, $\boldsymbol{\Psi}$ is symmetric but not necessarily positive semidefinite, it cannot satisfy the requirement of positive definiteness of $\boldsymbol{\Sigma}_k$ [20]. Considering the fact that, the eigenvalues of a covariance matrix can be seen as the variances along the directions of respective eigenvectors, the eigenvalue decomposition is performed to find eigenvalues and eigenvectors in step 3. A modification is then made in step 4 by using the positive robust variances calculated by Equation (12) to replace the eigenvalues, which may be negative [21], to obtain the positive diagonal covariance matrix $\boldsymbol{\Gamma}$. Then, $\boldsymbol{\Gamma}$ is used to estimate the positive-definite covariance matrix $\boldsymbol{\Sigma}_k{}'$ instead of Λ. It has been proven in [22] that there exist constant c_1 and c_2, such that the true $\boldsymbol{\Sigma}_k$ and $\boldsymbol{\mu}_k$ can be approximated by the estimations $\boldsymbol{\Sigma}_k{}'$ and $\boldsymbol{\mu}_k{}'$, that is

$\mathbf{\Sigma}_k \leftarrow c_1\mathbf{\Sigma}_k'$, and $\boldsymbol{\mu}_k \leftarrow c_2\boldsymbol{\mu}_k'$. For the classical fast minimum covariance determinant (FASTMCD) method [23], c_1 is defined as follows:

$$c_1 = \frac{med(d_{s_{k,m+1}}, \ldots, d_{s_{k,N}})}{\chi_2^2(0.5)}, \tag{14}$$

and $c_2 = 1$. Once $\boldsymbol{\mu}_k'$ and $\mathbf{\Sigma}_k$ are obtained in step 4, the MD for every position vector can be calculated according to Equation (11), which can be rewritten as follows:

$$d_{k,j}^2(\boldsymbol{\mu}_k, \mathbf{\Sigma}_k) = c_1^{-1}(\mathbf{p}_{k,j} - \boldsymbol{\mu}_k)^{\mathrm{T}}(\mathbf{\Sigma}_k')^{-1}(\mathbf{p}_{k,j} - \boldsymbol{\mu}_k). \tag{15}$$

Thus, the outliers are identified by comparing the squared MDs with the defined threshold d_{k0}^2 obtained in step 5. The choice of the threshold is based on the fact that, when the position matrix $\mathbf{P}_k \sim N(\boldsymbol{\mu}_k, \mathbf{\Sigma}_k)$, the squared MD $d_{k,j}^2$ is distributed as a χ^2 random variance with 2 degrees of freedom [22].

To reduce the false-alarm probability, we set an unreliable probability to every node in step 6. If $\mathbf{p}_{k,j}$ is detected as outliers, then the unreliable probability of s_j is set to be $q_{k,j} = 1/m$; otherwise, $q_{k,j} = 0$. After m position matrices are evaluated, the unreliable probability for every node in S_{rep} can be obtained by $q_j = \sum_{k=1}^{m} q_{k,j}$. Thus, the unreliable probabilities for the nodes from S_{rep} have been determined. To identify the unreliable nodes in S_{ref}, the detection method is repeated with different reference nodes, which are selected from the set of S_{rep} that have been identified as reliable.

3.3. WLS Based on Reliable Probability and Distance

When the unreliable probabilities for all nodes are determined, the WLS method with reliable probability q_i and distance $\hat{r}_i$ from the initial position to node s_i, $i = 1, \ldots, N$ is applied to perform localization and can be formulated as follows:

$$\hat{\mathbf{p}} = (\mathbf{A}^{\mathrm{T}}\mathbf{W}\mathbf{A})^{-1}\mathbf{A}^{\mathrm{T}}\mathbf{W}\mathbf{B}, \tag{16}$$

where

$$\mathbf{W} = \mathrm{diag}(w_1/\hat{r}_i, \ldots, w_N/\hat{r}_N). \tag{17}$$

$w_i = 1 - q_i$. The procedure for the proposed localization method is given by Algorithm 4.

Algorithm 4. The procedure of the proposed method based on DIG and MD: DIG_MD.

(1) Perform the DIG method to determine $\Omega_{ref} = \{\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_m\}$ ($m \geq 3$) and $\Omega_{rep} = \{\mathbf{s}_{m+1}, \mathbf{s}_{m+2}, \ldots, \mathbf{s}_N\}$, and then calculate m position matrices $\mathbf{P}_k$, $k = 1, \ldots, m$.
(2) Identify all the outliers in the matrices $\mathbf{P}_k$ $k = 1, \ldots, m$ based on Algorithm 2.
(3) Calculate the unreliable probabilities for nodes in Ω_{rep}.
(4) Estimate the source position based on Equation (16).
(5) Reselect the nodes with high low unreliable probability $q_i < 0.5$ from Ω_{rep} with the new initial position obtained from step 5.
(6) Repeat steps 1–4 to estimate the source position as the final localization results.

4. Simulations

In this section, we compare the performance of the proposed method, DIG_WD, with that of the PLE, the WLS-based distance method denoted as WLS (i.e., the reliable probabilities for all nodes are 1), and the EM-based method [10] through a series of computer simulations.

We assume that N nodes are placed uniformly in an $L \times L$ m^2 test area with a resolution of Δ_x and Δ_y along the horizontal and vertical directions, respectively. Each node is equipped with a microphone

array to estimate the AOA of the target, and 1000 Monte Carlo simulations are conducted for every case based on the parameters $L = 250$, $\Delta_x = \Delta_y = 50$, and $\alpha = 0.95$. Next, u randomly selected nodes are assumed to be subject to large noise or interference, and their standard deviation of the estimated bearing error is set to be σ_2; moreover, those of the remaining "reliable" nodes are set to be σ_1, $\sigma_2 \succ \sigma_1$. The initial positions for WLS, EM, and DIG_MD are obtained by the PLE.

For comparison purposes, we also apply the detection method, Algorithm 3, to identify the outliers from all IPs calculated by every two bearing lines. Instead of calculating the mean of IPs as the source position, the WLS estimator based on reliable probabilities and distance is also used to determine the position. When t $(t \leq n-1)$ the IPs can be obtained on the bearing line extending from s_i, $i = 1, 2, \ldots, n$; the unreliable probabilities of s_i is q/t if q IPs included in the t points are identified as outliers from all IPs. Next, the WLS based on Equation (16) is used to find the source position; this method is defined as the IP_WLS method. Furthermore, the center of all IPs after excluding all detected outliers is defined as CIP.

4.1. The RMSEs for Different σ_1 and σ_2

The localization performances of various approaches are influenced by the standard deviation of the estimated bearings error. For the source located at $\mathbf{p} = [73.3, 62.3]^{\mathrm{T}}$, the RMSEs of different algorithms with different σ_1 are plotted as shown in Figure 4a when $\sigma_2 = 15^\circ$, $u = 6$ and $m = 4$.

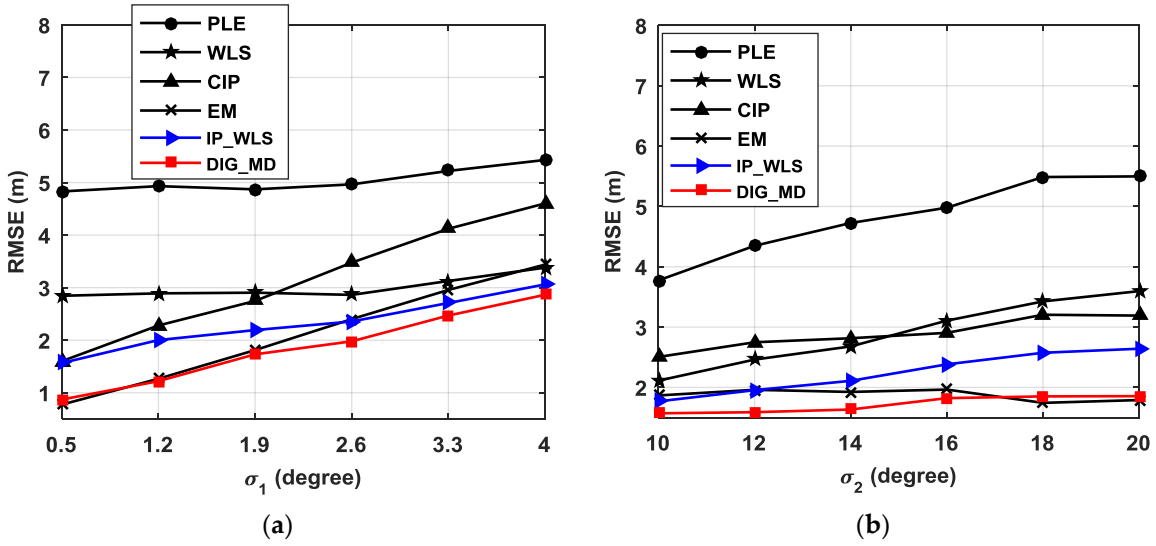

Figure 4. The RMSEs of the pseudolinear estimator (PLE), weighted least squares (WLS), center of all intersected points (CIP), expectation maximization (EM), IP_WLS, and division and greedy replacement–Mahalanobis distance (DIG_MD) methods with (**a**) different σ_1 when $\sigma_2 = 15^\circ$, (**b**) different σ_2 when $\sigma_1 = 2^\circ$. ($u = 6, m = 4$).

It can be seen that the existence of unreliable bearings can severely deteriorate the localization performance of the PLE, especially when σ_1 is small. When $\sigma_1 = 0.5^\circ$, the RMSE of the estimated errors for the PLE is as large as 4.83 m. Compared with WLS, the CIP method shows lower RMSEs only when $\sigma_1 < 2^\circ$, and IP_WLS always outperforms WLS for all values of σ_1, because it is easier to detect the outliers when the data are contaminated severely. This phenomenon also illustrates the importance of detecting outliers. From Figure 4a, we can also observe that IP_WLS shows better performance than CIP, illustrating the superiority of WLS over simple CIP. The EM exhibits a somewhat similar performance to that of DIG_WD when σ_1 is small; however, it shows a greater advantage as σ_1 increases.

The simulation is then conducted when $\sigma_1 = 2^\circ$ and σ_2 ranges from 10° to 20°, considering the fact that the background noise usually does not change greatly during a short period for certain applications. The results in Figure 4b indicate that the DIG_MD can significantly improve the localization performance compared with the conventional PLE and WLS. CIP can outperform WLS only when the difference between σ_1 and σ_2 is large, and it always has higher RMSEs than those of the IP_WLS method. EM performs slightly better than DIG_MD when σ_2 is significantly larger than σ_1. In contrast, the DIG_MD clearly outperforms EM when σ_2 is less than 16.

From Figure 4a,b, we can see that both IP_WLS and DIG_MD can improve the localization accuracy compared with the PLE and WLS. However, DIG_MD shows better performance than IP_WL. This is because IP_WLS is based on the outlier detection results of IP. These IPs are sensitive to the difference of two AOAs. When the source and two nodes are close to located at a line, the IP will be easily identified as outliers, and thus, false alarm probability will be increased. On the other hand, any of the two AOA errors will have an influence on the IP. When one IP is detected as an outlier, then two nodes will be allocated unreliable probabilities. As a result, the false alarm also exists if only one of them is reliable, especially when the node is close to the source. All these problems can be solved by the proposed DIG method.

4.2. The Influence of the Number of Reference Nodes

To discuss the effect of the number of reference nodes on the localization performance, the RMSEs for different scale of reference nodes are plotted in Figure 5 when $\sigma_1 = 2^\circ$ and $\sigma_2 = 15^\circ$.

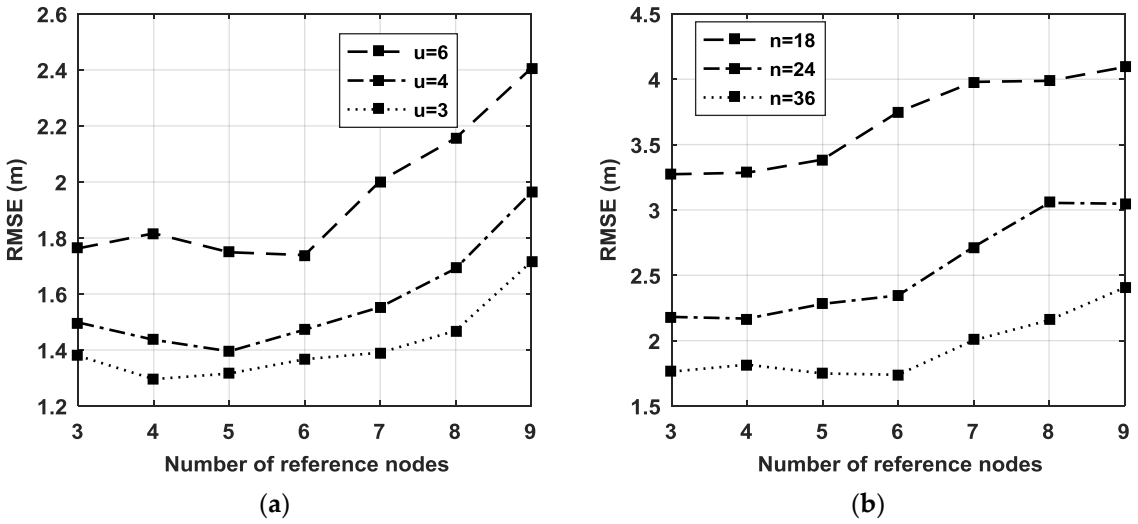

Figure 5. (**a**) The RMSEs of DIG_MD with the number of reference nodes when different unreliable nodes are present and $N = 36$; (**b**) the RMSEs of DIG_MD with the number of reference nodes when different nodes are used and $u = 6$.

We can see that more reference nodes should be used when the number of unreliable nodes increases, and the number of reference nodes should be no more than $[n/6]$ ($[\mathbf{x}]$ is the nearest integer to $\mathbf{x}$). Otherwise, the performance of DIG_WD will deteriorate seriously. As illustrated in Section 3, m position sets with $(N - m)$ elements in each set can be obtained using the DIG_MD method. When the number of reference nodes increases, the position sets also increase, whereas the number of estimated locations decreases. Only if there are enough positions to be evaluated should more position sets be used to increase the reliability of detection. To guarantee enough positions in each set to detect outliers, $(N - m)$ should be significantly greater than m. From the simulation results, it can be seen that the reference node number is preferred to be within the range from three to $[N/6]$.

4.3. The Localization Performance for Different Numbers of Unreliable Nodes

Figure 6 further shows the localization performance for different numbers of unreliable nodes. It can be seen that the RMSEs of all the methods increase as the number of unreliable nodes increases. Compared with the PLE, CIP can improve the localization performance when unreliable nodes are present; however, it exhibits slightly higher RMSE than PLE when there is no outlier. EM has the highest RMSE among EM, WLS, IP_WLS, and DIG_MD; however, it performs better than WLS when the number of unreliable nodes increases. The IP_WLS and DIG_MD methods can inhibit the effect of unreliable bearing measurements for all cases. The superiority of the DIG_MD method over other methods increases as the number of unreliable nodes increases.

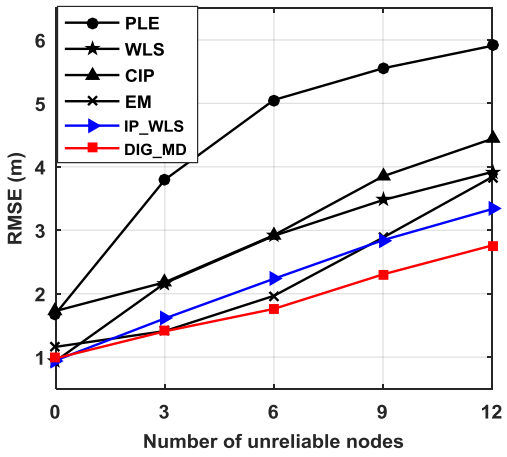

Figure 6. RMSE vs. the number of unreliable nodes when $\sigma_1 = 2^\circ$ and $\sigma_2 = 15^\circ$ and $m = 4$.

To investigate the robustness of the proposed method, the hit percentage of DIG_MD (when the errors of the evaluated methods are less than WLS or the PLE) is shown in Figure 7. The figure shows that the CIP method has the lowest hit percentages compared with both the PLE and WLS. EM has higher hit percentages than IP_WLS compared with the PLE, while the latter can improve localization accuracy with greater probability than EM compared with WLS. In contrast, the hit percentages of DIG_MD retain its superiority compared with both the PLE and WLS.

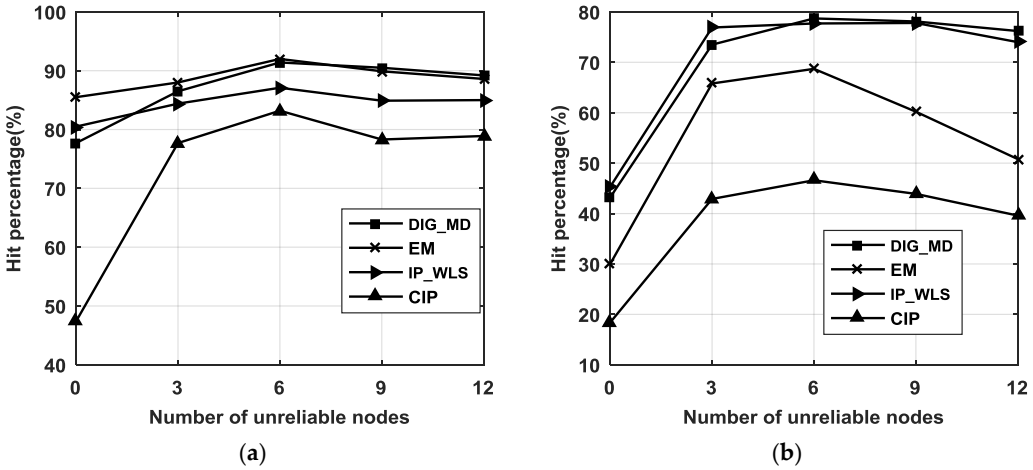

Figure 7. (**a**) Hit percentages of the DIG_MD, EM, IP_WLS, and CIP methods compared with the PLE when $\sigma_1 = 2^\circ$ and $\sigma_2 = 15^\circ$, and $m = 4$; (**b**) hit percentages of the DIG_MD, EM, IP_WLS, and CIP methods compared with WLS when $\sigma_1 = 2^\circ$ and $\sigma_2 = 15^\circ$, and $m = 4$.

4.4. The Localization Performance for Different Numbers of Nodes and for Different Source Positions

As the number of nodes usually has a great influence on the localization performance, we plot the relationship between RMSE and the number of nodes in Figure 8. For fairness, the number of unreliable nodes is $N/6$. The number of reference nodes is four. Figure 8 shows that EM has a higher RMSE than IP_WLS and DIG_MD methods when the number of nodes is 12. However, the IP_WLS method shows worse performance than EM as the number of nodes increases. The proposed method, DIG_MD, always has the best localization accuracy for the different cases.

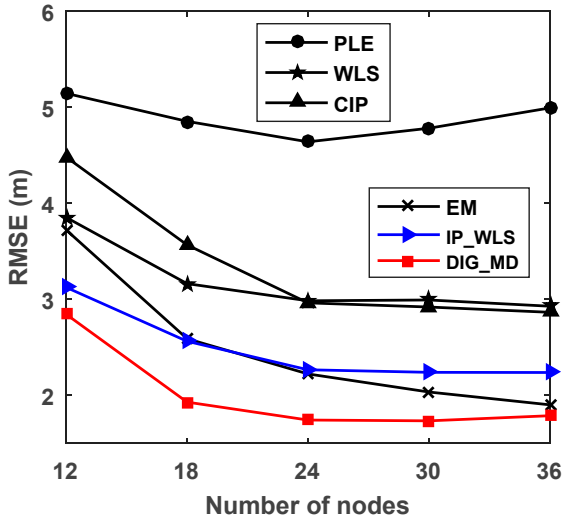

Figure 8. RMSE of PLE, WLS, CIP, EM, IP_WLS and DIG_MD vs. the number of nodes when $\sigma_1 = 2^\circ$, $\sigma_2 = 15^\circ$.

To study the efficiency of the proposed method for different source positions, Figure 9b shows the localization performance for five different source positions when six unreliable measurements are present. It is clear that the proposed method can improve the localization performance significantly for all source positions.

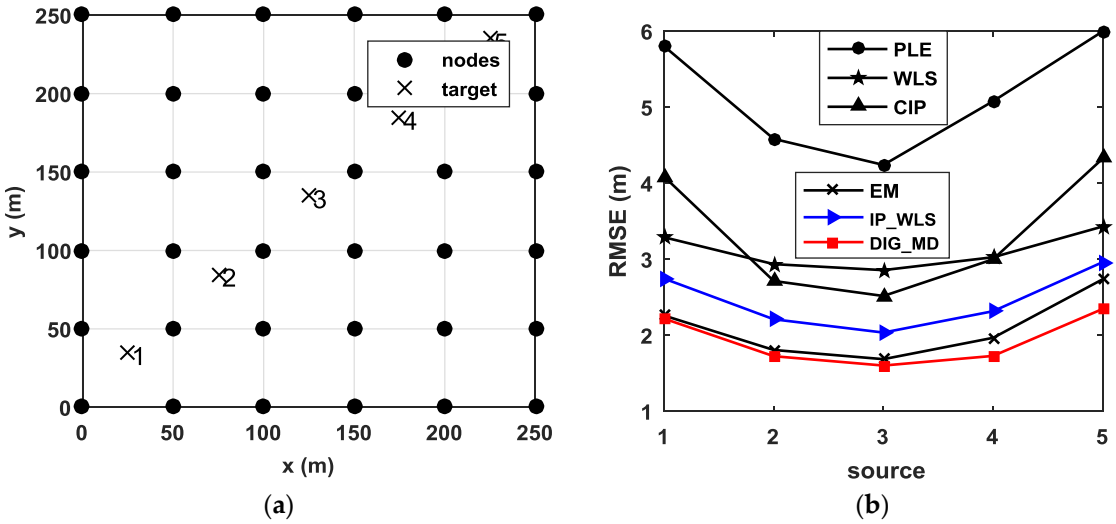

Figure 9. (**a**) Five target positions in a 36-node distributed network; (**b**) RMSE for different target positions when $m = 4$, $u = 5$, $\sigma_1 = 2^\circ$ and $\sigma_2 = 15^\circ$.

5. Outdoor Experiment Results and Analysis

In this section, we describe the verification of our proposed method using a 30-node network for acoustic source localization. All nodes were placed in an 11×11 m^2 square field, as shown in Figure 10. Each node is an autonomous vehicle equipped with a four-element cross microphone array, as shown in Figure 11. The microphone array is arranged into two orthogonal pairs 20 cm apart. Each pair of microphones estimates an AOA using the generalized cross correlation with phase transform (GCC-PHAT) [24] method. The final AOA is then obtained by the fusion of two AOAs obtained by the two pairs of microphones. The vertical distance from ground to microphone is also 20 cm. During the test, all nodes transmitted the estimated angles to the base station following a predefined collision-avoidance communication protocol. The localization tests were repeated 40 times. The acoustic source was a car engine noise generated by a loudspeaker orientated upward. Without loss of generality, we placed the speaker at the center of the test field (i.e., $\mathbf{x} = [5.5\text{m}, 5.5\text{m}]^{\mathrm{T}}$).

The experiment is conducted in an outdoor environment, with noise always present. However, the signal-noise-ratio (SNR) for each node is different as the distances from source to nodes are different. The range is from 5 to 15 dB. During the experiment, unreliable AOAs may be introduced by the following:

- Multipath signal: because the distance between the microphone array and the ground is only 20 cm, unreliable AOAs may be introduced by a multipath signal.
- Interferences: the movements of people and cars during the experiment are also causes of unreliable measurements.
- The low SNR: because of the possible nonstationary background, the SNR of the received signal of each node may vary in a large range, possibly resulting in unreliable measurements.

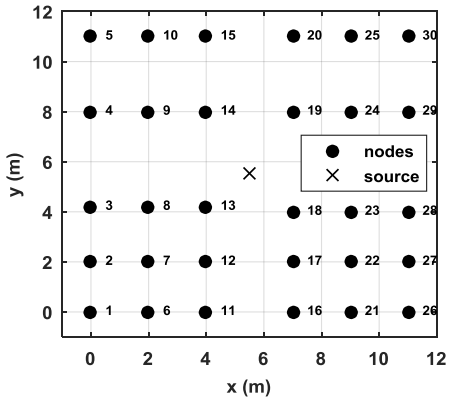

Figure 10. Node placement for the 35-node network.

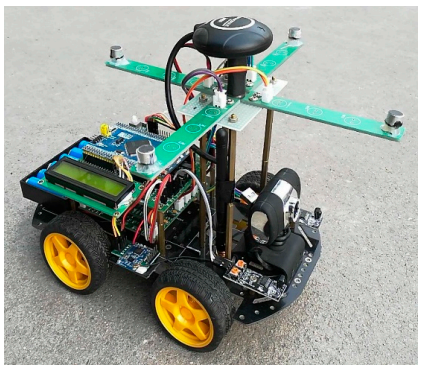

Figure 11. The four-element cross-microphone array.

To verify the influence of the number of reference nodes on the localization accuracy, we plotted the RMSEs of different numbers of reference nodes, as shown in Figure 12. We can see that the proposed method DIG_MD clearly outperforms other compared methods when the number of reference nodes is fewer than six. As the number exceeds six, the RMSEs of DIG_MD increase gradually. When more than nine reference nodes are used, the DIG_MD method yields similar localization performance to the IP_WLS. To show the localization results more clearly, we further compared the localization results of DIG_MD with the PLE for the 40 experiments with $m = 4$, as shown in Figure 13. The results show that while most of the large error peaks of PLE were substantially degraded, there were a few cases in which the DIG_MD method performed slightly better than the PLE (e.g., in the 11th and 33rd runs). To investigate the underlying reason, we plotted the unreliable sensor node detection results for the two cases, as shown in Figure 14a,b. For comparison purposes, the estimated AOA values and the detection results for the 10th and 32nd experimental runs for which the proposed method significantly improves the localization performance are plotted in Figure 14c,d, respectively.

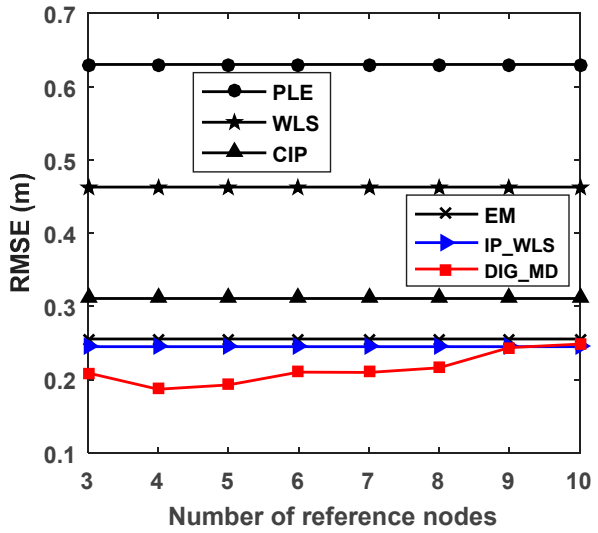

Figure 12. RMSEs for different numbers of reference nodes.

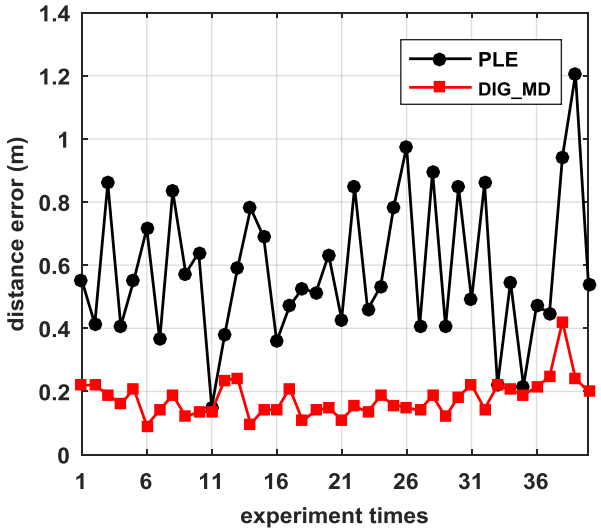

Figure 13. The localization error for 30 repetition estimation results.

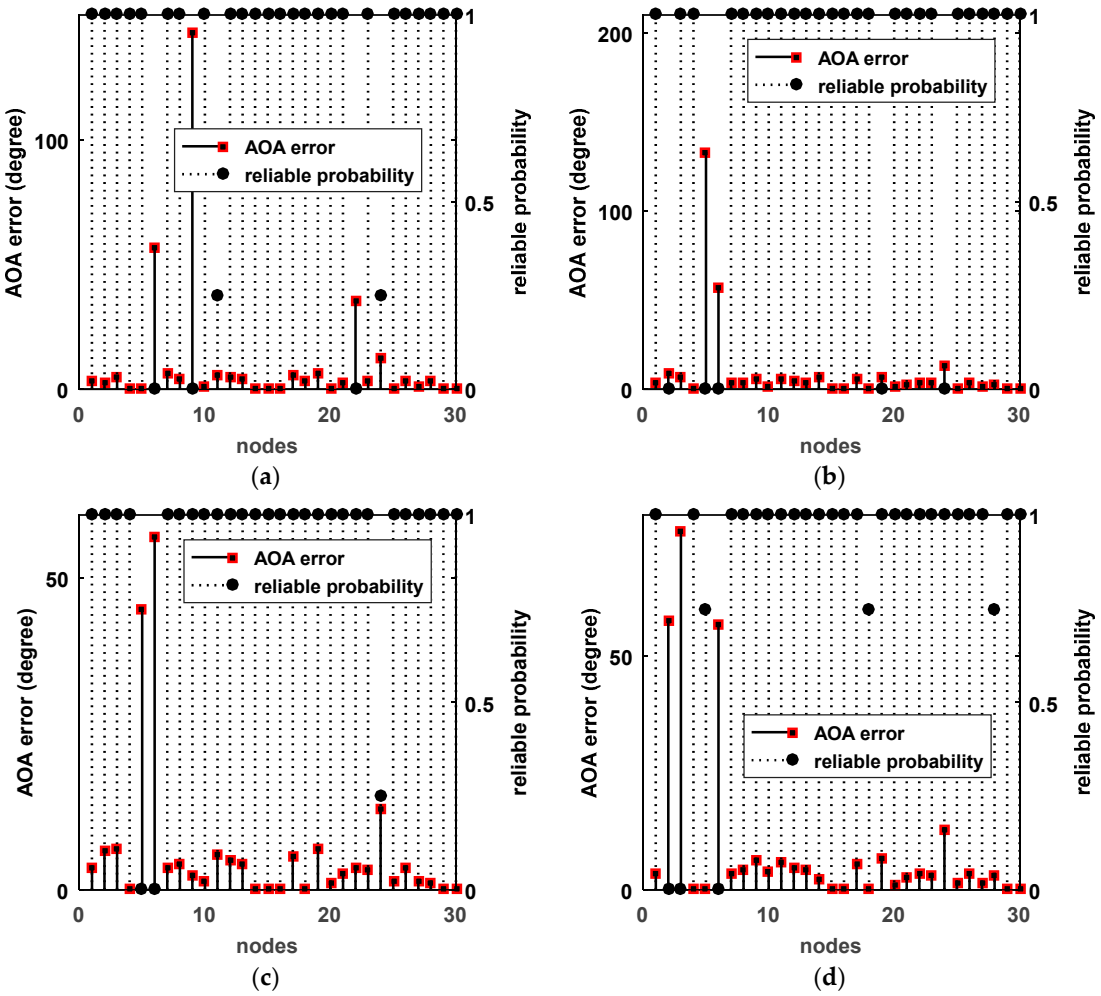

Figure 14. Estimated AOA errors and unreliable sensor detection results. (**a**) The 11th experiment; (**b**) the 33rd experiment; (**c**) the 10th experiment; (**d**) the 32nd experiment.

As nodes, s_{11}, s_{24}, are misjudged as unreliable for the 11th run, the localization error of DIG_MD is only slightly better than that of the PLE, even though the unreliable nodes s_6, s_9, s_{22} can be detected; a similar situation can also be found in the 33rd experiment. In contrast, the unreliable nodes in the 10th run can be detected correctly. For the 32nd run, the localization error can be significantly decreased while s_{18} and s_{28} are detected with a very low false-alarm probability.

To verify the localization performance under different numbers of nodes in a node network, we only use $s_1 \sim s_k$, $k = 20, 25, 30$ to perform localization when four reference nodes are used. Note that when different numbers of nodes are used, the source location is no longer located at the center of all nodes. The simulation results shown in Figure 15 reveal that DIG_MD has the best localization performance for all the cases considered.

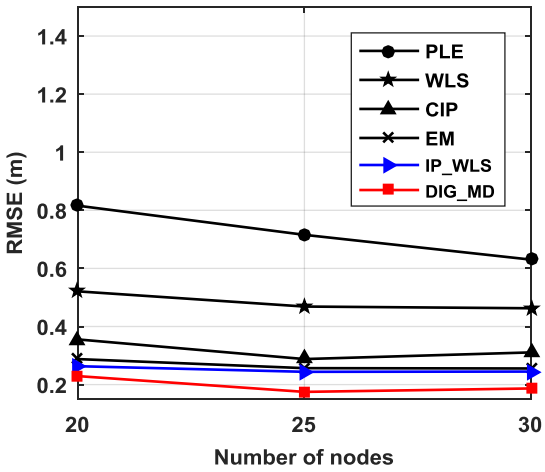

Figure 15. RMSEs for different numbers of nodes.

6. Conclusions

The localization performance of conventional AOA-based method, the PLE, is prone to be deteriorated when unreliable measurements are present. In this paper, we propose an unreliable node detection method based on the characteristics of the estimated positions of different node combinations. In the proposed approach, the DIG method is used to acquire different position sets, and the MD based on robust location and covariance matrix estimator is used to identify the outliers from the estimated target position sets. The proposed method does not require any prior information about the target and is easy to implement. Both simulation and outdoor experiment results show that DIG_MD is efficient and robust against the influence of unreliable measurements and can significantly improve the localization accuracy when the measurements are contaminated.

Author Contributions: Q.Y. proposed, programmed, and tested the proposed method. Q.Y. and J.C. designed the experiment and analyzed the experimental results, Q.Y. performed the experiment, Q.Y. wrote the paper, and J.C. and L.D.S. improved the paper.

Funding: This work is supported by the National Natural Science Foundation of China (No. 61501374) and NSFC-Zhejiang Joint Fund for the Integration of Industrialization and Information (U1609204).

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

References

1. Han, G.; Xu, H.; Duong, T.Q.; Jiang, J.; Hara, T. Localization algorithms of wireless sensor networks: A survey. *Telecommun. Syst.* **2013**, *52*, 2419–2436. [CrossRef]
2. Fresno, J.M.; Robles, G.; Martínez-Tarifa, J.M.; Stewart, B.G. Survey on the performance of source localization algorithms. *Sensors* **2017**, *17*, 2666. [CrossRef] [PubMed]
3. Cobos, M.; Antonacci, F.; Alexandridis, A.; Mouchtaris, A.; Lee, B. A survey of sound source localization methods in wireless acoustic sensor networks. *Wirel. Commun. Mob. Comput.* **2017**. [CrossRef]
4. Koks, D. *Passive Geolocation for Multiple Receivers with No Initial State Estimate*; No. DSTO RR-0222; Defence Science & Technology Organization: Edinburgh, SA, Australia, 2001.
5. Sayed, A.H.; Tarighat, A.; Khajehnouri, N. Network-Based wireless location: Challenges faced in developing techniques for accurate wireless location information. *IEEE Signal Process. Mag.* **2005**, *22*, 24–40. [CrossRef]
6. Lee, W.C. *Uncertainty in Wireless Sensor Networks*; Workshop on AFRL: Wright-Patterson Air Force Base, OH, USA, 2010.
7. Yan, Q.; Chen, J.; Ottoy, G.; Cox, B.; De Strycker, L. An accurate AOA localization method based on unreliable sensor detection. In Proceedings of the 2018 IEEE Sensors Applications Symposium (SAS), Seoul, Korea, 12–14 March 2018; pp. 1–6.

8. Yan, Q.; Chen, J.; Ottoy, G.; De Strycker, L. Robust AOA based acoustic source localization method with unreliable measurements. *Signal Process.* **2018**, *152*, 13–21. [CrossRef]
9. Yu, K.; Guo, Y.J. Statistical NLOS identification based on AOA, TOA, and signal strength. *IEEE Trans. Veh. Technol.* **2009**, *58*, 274–286. [CrossRef]
10. Giménez Febrer, P.J.; Silva Pereira, S.; López Valcarce, R. Distributed AOA-based source positioning in NLOS with sensor networks. In Proceedings of the 2015 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Brisbane, QLD, Australia, 19–24 April 2014; pp. 3197–3201.
11. Griffin, A.; Mouchtaris, A. Localizing multiple audio sources from DOA estimates in a wireless acoustic sensor network. In Proceedings of the 14th IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), New Paltz, NY, USA, 20–23 October 2013; pp. 1–4.
12. Do, H.; Silverman, H. A fast microphone array srp-Phat source location implementation using Coarse-To-Fine Region Contraction (CFRC). In Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), New Paltz, NY, USA, 21–24 October 2007; pp. 295–298.
13. Lee, D.D.; Kashyap, R.L. Robust maximum likelihood bearing estimation in contaminated Gaussian noise. *IEEE Trans. Signal Process.* **1992**, *40*, 1983–1986. [CrossRef]
14. Yardimci, Y.; Cetin, A.E.; Cadzow, J.A. Robust direction-of-arrival estimation in non-Gaussian noise. *IEEE Trans. Signal Process.* **1998**, *46*, 1443–1451. [CrossRef]
15. Besson, O.; Yuri, A.; Ben, J. Direction-of-arrival estimation in a mixture of K-Distributed and Gaussian noise. *Signal Process.* **2016**, *128*, 512–520. [CrossRef]
16. De Maesschalck, R.; Jouan-Rimbaud, D.; Massart, D.L. The mahalanobis distance. *Chemom. Intell. Lab. Syst.* **2000**, *50*, 1–18. [CrossRef]
17. Bishop, A.N.; Fidan, B.; Anderson, B.D.; Pathirana, P.N.; Dogancay, K. Optimality analysis of sensor node-Target geometries in passive localization: Part 1-Bearing-Only localization. In Proceedings of the 2007 3rd International Conference on Intelligent Sensors Sensor Networks and Information Processing (ISSNIP), Melbourne, VIC, Australia, 3–6 December 2007; pp. 7–12.
18. Gnanadesikan, R.; Kettenring, J.R. Robust estimates, residuals, and outlier detection with multiresponse data. *Biometrics* **1972**, *28*, 81–124. [CrossRef]
19. Maronna, R.A.; Ruben, H.Z. Robust estimates of location and dispersion for high-dimensional datasets. *Technometrics* **2002**, *44*, 307–317. [CrossRef]
20. Rousseeuw, P.J.; Molenberghs, G. Transformation of nonpositive semidefinite correlation matrices. *Commun. Stat. Theory Methods* **1993**, *22*, 965–984. [CrossRef]
21. Marcus, M.; Minc, H. *A Survey of Matrix Theory and Matrix Inequalities*; Courier Corporation: Chelmsford, MA, USA, 1992.
22. Rousseeuw, P.J.; Hubert, M. Robust statistics for outlier detection. In *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*; John Wiley & Sons: Hoboken, NJ, USA, 2011; pp. 73–79.
23. Rousseeuw, P.J.; Driessen, K.V. A fast algorithm for the minimum covariance determinant estimator. *Technometrics* **1999**, *41*, 212–223. [CrossRef]
24. Knapp, C.; Carter, G. The generalized correlation method for estimation of time delay. *IEEE Trans. Acoust. Speech Signal Process.* **1976**, *24*, 320–327. [CrossRef]