sensors

MDPI

Article

Sound Localization and Speech Enhancement Algorithm Based on Dual-Microphone

Tao Tao, Hong Zheng *, Jianfeng Yang, Zhongyuan Guo, Yiyang Zhang, Jiahui Ao, Yuao Chen, Weiting Lin and Xiao Tan

School of Electronic Information, Wuhan University, Wuhan 430072, China; tt1295@whu.edu.cn (T.T.); yjf@whu.edu.cn (J.Y.); guozhongyuan@whu.edu.cn (Z.G.); monster233@whu.edu.cn (Y.Z.); jiahui@whu.edu.cn (J.A.); chenyuao@whu.edu.cn (Y.C.); 2019262120001@whu.edu.cn (W.L.); haiantanxiao@whu.edu.cn (X.T.)

* Correspondence: zh@whu.edu.cn

Abstract: In order to simplify the complexity and reduce the cost of the microphone array, this paper proposes a dual-microphone based sound localization and speech enhancement algorithm. Based on the time delay estimation of the signal received by the dual microphones, this paper combines energy difference estimation and controllable beam response power to realize the 3D coordinate calculation of the acoustic source and dual-microphone sound localization. Based on the azimuth angle of the acoustic source and the analysis of the independent quantity of the speech signal, the separation of the speaker signal of the acoustic source is realized. On this basis, post-wiener filtering is used to amplify and suppress the voice signal of the speaker, which can help to achieve speech enhancement. Experimental results show that the dual-microphone sound localization algorithm proposed in this paper can accurately identify the sound location, and the speech enhancement algorithm is more robust and adaptable than the original algorithm.

Keywords: dual-microphone array; sound localization; speech enhancement; time delay estimation; post-filtering

check for updates

Citation: Tao, T.; Zheng, H.; Yang, J.; Guo, Z.; Zhang, Y.; Ao, J.; Chen, Y.; Lin, W.; Tan, X. Sound Localization and Speech Enhancement Algorithm Based on Dual-Microphone. *Sensors* **2022**, *22*, 715. https://doi.org/10.3390/s22030715

Received: 5 December 2021
Accepted: 12 January 2022
Published: 18 January 2022

1. Introduction

Microphone array is a key technology of human–computer interaction (HCI). It can enhance the efficiency of HCI and adapt intelligent speech device to more complex and changing environments [1–3]. Microphone array, which can acquire the voice signal by microphones, uses digital electronic technology to sample, process and analyze the acoustic field characteristics, so that the collected voice signal is easier to be processed. Due to factors such as cost control, performance optimization, and environmental adaptability, acoustic signal processing based on dual-microphones is a challenging task [4,5].

Acoustic signal processing technology based on microphone array includes multiple technologies such as sound localization, speech separation, and speech enhancement. As a simple acoustic signal receiving device, the microphone is widely used in various sound localization experiments [6,7]. Ganguly A. et al. [8] proposed a dictionary-based singular value decomposition algorithm to solve the sound localization problem with the help of the non-linear and non-uniform microphone array in the smart phone and proved the accuracy of the algorithm in an environment with extremely low signal-noise ratio (SNR) through experimental results. Nevertheless, this algorithm cannot obtain the spatial position of the acoustic source and the distance from the acoustic source to the center of the microphone array. Jiaze Li and Jie Liu [9] derived and compared the four-element cross microphone array and the five-element cross microphone array based on the generalized cross-correlation time delay estimation algorithm. The experiment showed that the four-element cross microphone array has a blind spot for sound localization, while the five-element microphone can better improve localization accuracy and reduce errors.

However, the structure of the microphone array is very complicated, which increases the hardware cost. Jelmer Tiete, Federico Domínguez, etc., [10] exploited the sensing capabilities of the Sound-Compass in a wireless sensor network to localize noise pollution, whose live tests produced a sound localization accuracy of a few centimeters in a 25 m^2 anechoic chamber, while simulation results accurately located up to five broadband acoustic sources in a 10,000 m^2 open field. The system requires 25 sensors, which makes it difficult to meet the requirements of miniaturization. Hongyan Xing, Xu Yang [11] made a theoretical model of a three-plane five-element microphone array is established, using time-delay values to judge the acoustic source's quadrant position, which derived a formula for the sound azimuth calculation of a single-plane five-element microphone array based on sound geometric localization. It is also necessary to detect the environmental adaptability of the system and the working accuracy in a high-noise environment.

As the main method of acoustic signal processing, speech enhancement is a key issue to improve the accuracy of acoustic information extraction. Shujau M. et al. [12] proposed a multi-channel speech enhancement algorithm based on independent component analysis (ICA) for co-located microphone recording. Experiments show that the algorithm significantly improves the quality and clarity of the acoustic signal. This algorithm is more suitable for linear directional microphone arrays. Xunyu Zhu [13] advanced a deep neural network combining beamforming and deep complex U-net network to denoise acoustic signals from small-scale microphone arrays, which has certain advantages in environments such as homes, conference rooms, and classrooms. The author did not solve the human voice interference, especially the human voice interference that is in the same direction as the target acoustic source. On the basis of dual microphone arrays, Hairong Jia et al. [14] proposed a speech enhancement algorithm based on dual-channel neural network time-frequency masking, which combines single-channel neural network, adaptive mask orientation and proper positioning, and convolutional beamforming. Compared with traditional single-channel and dual-channel algorithms, the algorithm can extract voice information more clearly. In the algorithm, the network model has a large amount of calculation, which leads to higher hardware performance requirements for system implementation.

Traditional microphone arrays require many microphones, resulting in high cost and high design requirements. At present, there are new design concepts for intelligent home speech modules, such as lightweight, high integration and cost control. Speech module design based on dual microphones or even single microphone is an increasingly popular direction in the field of intelligent acoustic signal processing [15,16].

With the current lightweight and highly integrated design concepts of intelligent homes and interactive robots, combined with the current status and development trend of voice signal processing, this paper proposes a dual-microphone-oriented sound localization and voice enhancement optimization algorithm. The algorithm can use two microphones to locate the speaker target, realize the enhancement of the acoustic signal, and output a high SNR corpus that is more convenient for back-end analysis.

The rest of the paper is arranged as follows. Section 2 introduces two acoustic signal models in detail. Section 3 describes the improved algorithm of sound localization based on dual-microphone. Section 4 presents the advanced speech enhancement algorithm based on sound localization. Section 5 gives the experimental results. Section 6 concludes the paper.

2. Acoustic Signal Model

On the one hand, analyzed from the propagation mode, the acoustic signal is a longitudinal wave. That is to say, it is a wave in which the particles in the medium move along the direction of propagation. On the other hand, the acoustic signal can also be seen as a spherical wave. After vibration occurs at the acoustic source to generate an acoustic signal, the medium near the acoustic source appears accompanied by vibration, and the voice signal spreads around along with the medium simultaneously [17].

According to the distance between the acoustic source and the microphone array, the acoustic field model can be divided into two types: Near-field model and far-field model. The near-field model regards the voice signal as a spherical wave, and it considers the amplitude difference of the voice signal received by the sensors on the microphone array.

Generally, the near-field model and the far-field model are defined according to the relationship between the distance between the acoustic source and the center point of the microphone array element and the acoustic wavelength [18]:

$$\begin{cases} \mathrm{L} > \frac{2\mathrm{d}^2}{\lambda_{\min}}, \text{ the far} - \text{field model} \\ \mathrm{L} < \frac{2\mathrm{d}^2}{\lambda_{\min}}, \text{ the near} - \text{field model} \end{cases} \tag{1}$$

In Equation (1), L is the distance between the acoustic source and the center of the microphone array, and d is the aperture of the array element, and $\lambda_{\min}$ is the minimum wavelength of the current voice.

3. Sound Localization Algorithm by Dual-Microphone

3.1. Time-Delay Estimation

We can calculate the azimuth angle of the acoustic source by processing multi-channel signals based on sound localization algorithms. When calculating the azimuth angle, the phase difference of the signals received by the microphones at different positions is used to estimate the position of the speaker. Generally, because the distances between the acoustic source and the two microphones are not same, the arrival time difference of the acoustic wave is reflected in the waveform diagram as the phase difference of the voice waveforms received by the two microphones. The distance difference between the speaker and the microphone array is equal to the product of the acoustic signal propagate speed in the air and the relative delay between the two microphones. As mentioned in Section 2, the acoustic signal can be seen as propagating outward in the form of waves.

As shown in Figure 1, referring to the far-field model, we can estimate the azimuth in a 2D plane by dual-microphone array. However, if it is expanded to a 3D space, the estimated value of the azimuth angle will be a sector, so that the acoustic wave reaching the microphone is a spherical wave. At this time, the arrival angle θ cannot be expressed as a function of time delay, which is the difficulty of the sound localization algorithm based on dual microphones. After the time delay is obtained, the distance difference between the two microphones and the speaker can be calculated [19].

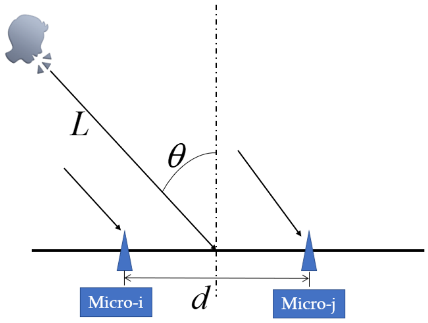


Figure 1. Principle of speaker positioning.

As shown in Figure 2, the dual-microphone acoustic field is similar to the hyperbolic model. The distance difference between the point on the hyperbola and the two focal points is a fixed quantity, so the acoustic source must be located on the hyperbola. If there is another distance difference at the same time, the corresponding hyperbolas can also be calculated [20,21]. The intersection of the two hyperbolas is the speaker position, as shown in Figure 3.

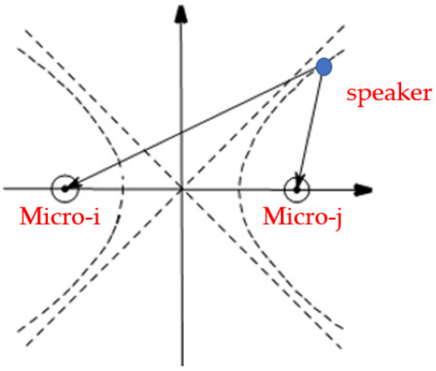


Figure 2. Dual-microphone positioning model.

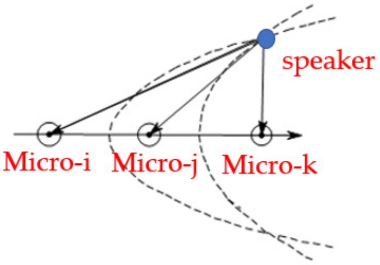


Figure 3. Multi-microphone positioning model.

Taking the midpoint of microphones i and j as the center of the coordinate system, the distance between microphones i and j is d. Let the coordinates of microphone i and j be (−d/2,0) and (d/2,0), respectively, and the acoustic source coordinates are (x,y). Then, the distance between the speaker and the two microphones is:

$$\begin{cases} L_{is} = \sqrt{\left(\frac{d}{2}+x\right)^2+y^2} \\ L_{js} = \sqrt{\left(\frac{d}{2}-x\right)^2+y^2} \end{cases} \tag{2}$$

In Equation (2), L_{is}, L_{js} are the distance between the speaker and dual-microphone. The distance difference $L\prime_{ij} = |L_{is} - L_{js}|$ between the speaker and the two microphones can be obtained. According to the estimated distance difference calculated by the time delay $\hat{L}\prime_{ij} = c \times \Delta t_{ij}$, the problem is transformed into the estimation of the position coordinate x of the speaker by minimizing the error between $L\prime_{ij}$ and $\hat{L}\prime_{ij}$ when the microphone coordinates and Δt_{ij} are known.

The sound localization method based on time delay requires at least two sets of data to construct two sets of hyperbolas and calculate their intersection points, because only the linear function relationship between x and y can be obtained by the information of time delay, which can only be embodied as the azimuth angle between the acoustic source and the center of the dual microphone array. If there is another distance difference exist, extra azimuth angle can be calculated by the new set of hyperbolas, and the intersection of the two hyperbolas is the acoustic source position. It can be seen that in a 2D space, three microphones can be used to estimate the position of the acoustic source. Therefore, the problem of sound localization is transformed to a problem of solving the intersection of two hyperbolas [22].

3.2. Energy Difference Estimation

In the process of acoustic wave propagation, there is energy attenuation except time delay exist. Considering the energy attenuation and time delay at the same time, the mathematical model of the signal received by the dual-microphones can be solved.

The acoustic signals received by microphones i and j are defined as follows:

$$\begin{cases} x_i = \frac{2L(n-t_i)}{d_i} \\ x_j = \frac{2L(n-t_j)}{d_j} \end{cases} \tag{3}$$

In Equation (3), x_i and x_j are the voice signals received by microphone i and microphone j, respectively. L(n) is the acoustic source, and t_i and t_j are the time when the two microphones receive signals. Respectively, d_i and d_j are the distances from the sound source to the two microphones. We define the sound intensity amplitude of the signal received by the microphone i as E_i, which can be actually measured. Finally, the sound intensity amplitude is derived as shown in Equation (4).

$$\frac{E_i}{E_j} = \frac{d_i^2}{d_j^2} \tag{4}$$

Combining Equations (3) and (4), we can get:

$$\begin{cases} \sqrt{(-\frac{d}{2} - x_s)^2 + (0 - y_s)^2} = \frac{d_{ij}\sqrt{E_j}}{\sqrt{E_j}-\sqrt{E_i}} \\ \sqrt{(\frac{d}{2} - x_s)^2 + (0 - y_s)^2} = \frac{d_{ij}\sqrt{E_i}}{\sqrt{E_j}-\sqrt{E_i}} \end{cases} \tag{5}$$

In Equation (5), d_{ij} is the distance difference between the sound source and the two microphones. It can be seen from Equation (5) that the two equations also construct a coordinate system similar to Figure 3 in the difference estimation of the energy field. In the energy field, the geometric model is two circles with the microphone i and j coordinates as the center and $d_{ij}\sqrt{E_j}/(\sqrt{E_j} - \sqrt{E_i})$, $d_{ij}\sqrt{E_i}/(\sqrt{E_j} - \sqrt{E_i})$ as the radius, respectively. The intersection of the two circles is the acoustic source position.

According to Euclidean geometry, when the distance between the centers of the two circles is greater than the difference between the radii of the two circles and less than the sum of the radii of the two circles, the two circles must intersect. Which is:

$$d_{ij} = \frac{d_{ij}(\sqrt{E_j} - \sqrt{E_i})}{\sqrt{E_j} - \sqrt{E_i}} \leq d \leq \frac{d_{ij}(\sqrt{E_j} + \sqrt{E_i})}{\sqrt{E_j} - \sqrt{E_i}} \tag{6}$$

Obviously, Equation (6) is always established, so Equation (5) must have two sets of real number solutions that are symmetrical about the microphone connection. Finally, combined with the actual scene, the optimal solution in the 2D space is selected.

3.3. Sound Source Localization

Based on the time delay estimation and the energy difference estimation, the sound source position and the sound source direction angle under 2D coordinates will be obtained. However, it is still impossible to obtain the acoustic source distance in the 3D space.

In order to solve this problem, this paper introduces the Steered Response Power-Phase Transform (SRP-PHAT) based on the weighted phase transformation to achieve the maximum autocorrelation estimation, thereby obtaining the most likely acoustic source position in the 3D space.

Before the sound localization, pre-emphasis, framing and other pre-processing are performed on the acoustic signal. Based on short-time Fourier transform (STFT), the spectrum analysis of two single-channel speech signals is carried out with acoustic equal-frame modeling technique.

The PHAT algorithm in this paper uses a steerable beam response power algorithm to sum all possible phase transforms. SRP-PHAT can directly transform and process multi-channel microphone signals and use multiple microphones to improve the accuracy of position estimation.

SRP can be implemented using a block processing scheme that uses a short-time digital Fourier transform as an estimate of the microphone signal spectrum. Divide the array signal into blocks in the time domain and calculate the steering response for each block. The digital Fourier transform of the signal block is denoted by $X_{k,b}[k]$. Where, m is the microphone index, b is the block index, and $G_{k,b}[k]$ is the Fourier transform of the discrete-time filter of microphone m, which is performed separately in each block. The steering response of block b can be defined as follows:

$$\widetilde{P}_b\Big[\Delta_1, \Delta_2] = \sum_{k=1}^{2} Y_{b'}[k, \Delta_1, \Delta_2]\widetilde{Y}_b[k, \Delta_1, \Delta_2] \tag{7}$$

$\widetilde{Y}_b[k, \Delta_1, \Delta_2]$ is a discrete frequency function and successive steering delays with index k. Where, Δ_1, Δ_2 represents all successive steering delays of the dual-microphone array in theory, it is necessary to process the data of all frequency bands in the signal. However, in actual, the data of one or more frequency bands are generally selected for processing. At the same time, although the k steering delays are continuous, in actual use, sampling is performed at a predefined set of spatial positions or directions, and the steering response power is obtained by summing k discrete frequencies.

$$\widetilde{Y}_b[k, \Delta_1, \Delta_2] = \sum_{k=1}^{2} G_{k,b}[k]X_{k,b}[k]e^{-jw\Delta_k} \tag{8}$$

The discrete filter G(t) is defined as Equation (9):

$$G_{m,b}(k) = \frac{1}{F_{m,b}(k)}, m = 1, 2 \tag{9}$$

where, b is the block index after framing, $F_{m,b}(k)$ is the Fourier transform of the signal block after framing, m is the microphone index.

Substituting Equation (9) into Equation (7), the controllable response weighted by the phase transformation is expressed as:

$$\widetilde{Y}_b^{PHAT}(\Delta_1, \Delta_2) = \sum_{k=1}^{2} \frac{F_{m,b}(k)}{|F_{m,b}(k)|} e^{-jw\Delta_m}, m = 1, 2 \tag{10}$$

Substituting Equation (10) into Equation (8), the controllable response power SRP-PHAT weighted by phase transformation can be obtained as:

$$\widetilde{P}_b^{PHAT}(\Delta_1, \Delta_2) = \sum_{k=1}^{2} \widetilde{Y}_b^{PHAT}(k, \Delta_1, \Delta_2)\widetilde{Y}_{b'}^{PHAT}(k, \Delta_1, \Delta_2) \tag{11}$$

In theory, it is necessary to analyze the data of all frequency bands in the acoustic signal. However, in the algorithm realization process, the acoustic signal processing method is somewhat different from the theory. Firstly, a predefined set of spatial positions or directions. Secondly, the voice signal is sampled, and the discrete frequencies are summed. Finally, the steering response power $\widetilde{P}_b^{PHAT}$ can be obtained.

The sound localization steps are as follows:

(1) Calculating the controllable time delay of the 2D azimuth direction in Section 3.2, which is according to the physical parameters of the microphone array;

(2) Using the STFT of the acoustic signal and the controllable time delay to calculate the SRP-PHAT for all frequencies in this direction;

(3) Repeating the above operations until SRP-PHAT in all directions is obtained;

(4) Selecting the direction corresponding to the maximum value as the azimuth angle of the sound source in 3D;

(5) Obtaining the 3D position of the sound source;
(6) The sound source localization algorithm flow is shown in Figure 4.

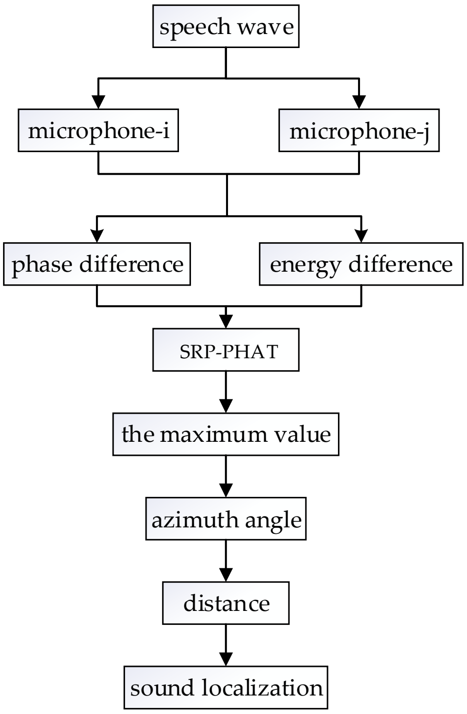


Figure 4. Sound localization algorithm based on dual-microphone.

4. Speech Enhancement Algorithm Based on Sound Localization

Traditional algorithms have insufficient speech enhancement effects in strong-noise or multi-noise environments. Correlation noise will be generated and there are higher requirements for the microphone array. With the development of signal processing technology, more and more speech enhancement algorithms have emerged, such as wavelet transformation, speech enhancement algorithms based on empirical mode decomposition and deep learning [23]. New speech enhancement algorithms pay more attention to noise feature analysis and statistics. According to the analysis results of the noise characteristics, the noise signal and the original speech signal are separated to further obtain the original speech signal, but the algorithm time efficiency and economic efficiency are low.

Combining the results of sound localization in Section 3, this paper proposes an optimization algorithm for indoor speech enhancement based on post-filtering. According to the azimuth information of the acoustic source, the enhancement algorithm only amplifies the acoustic signal from the speaker, while other signals are judged as background noise and will be effectively suppressed.

4.1. Speech Separation Algorithm Based on the Azimuth of the Target Sound Source

The ultimate goal of speech enhancement technology is to extract the source signal, but the source signal is often unclear in the living environment, resulting in the speaker signal entraining other interference signals or noise during the enhancement process [24]. The speech separation algorithm can not only remove the environmental noise and interference components, but also effectively separates the speech signals of different speakers. Independent Component Analysis (ICA) has better performance and higher stability, which is currently the most conventional and popular speech separation algorithm [25,26]. The principle of ICA is decomposing the aliased signal to obtain several independent signals. In this article, we define multiple independent source signals as S and the observation signal X after passing through the mixing matrix A, which is expressed as a matrix:

$$X(t) = AS(t) \tag{12}$$

where, the observation signal X(t) is the linear aliasing of n mutually independent unknown source signals S(t), and A is an m×n aliasing matrix whose aliasing weight coefficient of the matrix is unknown. When both S(t) and A are unknown, the core of the ICA algorithm is to solve the demixing matrix W so that the final output signal Y(t) optimally approximates the source signal S(t) according to certain criteria (such as independence criteria):

$$Y(t) = WX(t) \tag{13}$$

The process of solving the demixing matrix W is the process of feature extraction. This paper selects the azimuth angle information between acoustic source and dual-microphone as ICA analysis feature. Based on the definition of negative entropy, need to define was a column vector of matrix W. The objective function of the ICA algorithm is:

$$J(w) \propto \left\{E[G(w^T X)] - E[G(u)]\right\}^2 \tag{14}$$

where, u is a Gaussian variable with zero mean unit variance; G is a random non-negative quadratic function; X is a target sound source position vector, which is as the signal characteristic value. Taking the partial derivative of Equation (14) to get:

$$\frac{\partial J}{\partial w} = 2\left\{E[G(w^T X)] - E[G(u)]\right\}E[Xg(w^T X)] \tag{15}$$

In Equation (15), the g function is the derivative of the G function. Setting $\partial J/\partial w = 0$ directly will lead to poor convergence of the algorithm. Associating Equation (14) and Equation (15), it shows that the maximum value of the objective function J(w) can be obtained by the optimal solution of $E[G(w^T X)]$.

According to KKT constrained optimization, the optimal solution of $E[G(w^T X)]$ is an unconstrained optimization problem:

$$J'(w) = E[G(W^T X)] + \psi(\|w\| - 1) \tag{16}$$

where, ψ is a constant parameter. Based on Equation (16), the function H(w) is defined as follows:

$$H(w) = E[Xg(w^T X)] - \psi w \tag{17}$$

Derivation:

$$\frac{\partial H}{\partial w} = E[XX^T g'(w^T X)] - \psi \tag{18}$$

Finally, the matrix W is solved according to the Newton iteration method:

$$\begin{cases} w(j+1) = E\{Xg[w^T(j)X]\} - E\{g'[w^T(j)X]\}w(j) \\ w(j+1) = \frac{w(j+1)}{\|w(j+1)\|_2} \end{cases} \tag{19}$$

4.2. Speech Enhancement Algorithm Based on Post-Adaptive Filter

The idea of sub-frame block-index in Section 3.3 will also be applied to the adjustment of Wiener filter parameters in speech enhancement. The core of the adaptive algorithm is to modify the parameters of the filter based on the analysis of the first three voice framing blocks of the dual-channel voice signal collected by the front-end dual microphones, so as to achieve the optimal filtering.

Spectral subtraction is one of the effective technologies to enhance the quality of the voice signal, it has a good noise reduction effect at low SNR, the convergence rate and imbalance are affected by step size in LMS adaptive filtering algorithm. This paper introduces a method to enhance the quality of speech signal based on the combination of spectral subtraction and variable-step LMS adaptive filtering algorithm, to adjust the step size by changing the squared term of error, the step size follows the principle of change after the first fixed, achieves the purpose to improve the convergence rate and reduces the steady-state error.

As shown in Figure 5, $x(t)$ is the original signal input, $y(t)$ is the output signal of the system after the adaptive filter, $e(t)$ is the expected response, and $N(t)$ is the noise signal of the signal.

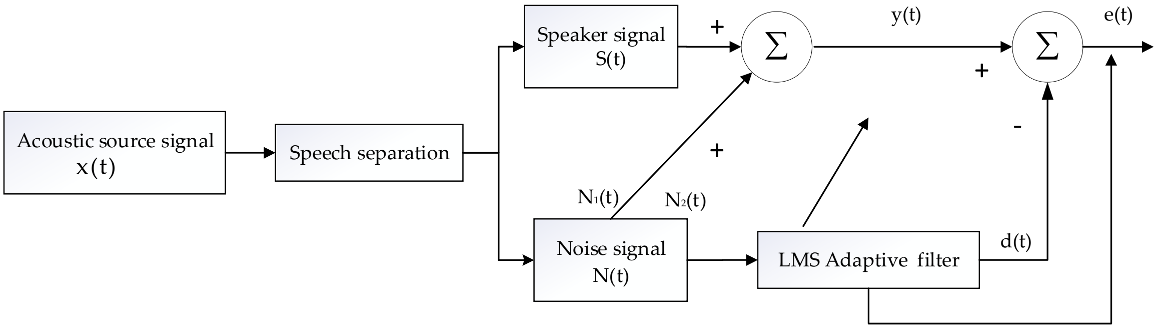


Figure 5. Adaptive filter flow.

The key to adaptive noise filtering is to obtain the best estimate of noise. The filter parameters obtained from the previous speech frame are used to adjust the control parameters of the latter speech frame, so as to obtain the error function of the system for improving the SNR. If the reference noise is related to the noise in the signal, the randomness of the noise can be better offset, and the noise can be completely eliminated. However, when the reference noise is not correlated with the noise in the signal or the correlation is weak, the noise cannot be completely cancelled out, and the filtering effect is not obvious. From Figure 5, we can get:

$$e(t) = y(t) - d(t) = S(t) + N_1(t) - d(t) \tag{20}$$

Then:

$$e^2(t) = S^2(t) + [N_1(t) - d(t)]^2 + 2[N_1(t) - d(t)]S(t) \tag{21}$$

Equation (21) takes the expectation on both sides of the equal sign to get:

$$E[e^2(t)] = E[s^2(t)] + E[N_1(t) - d(t)]^2 + 2E[(N_1(t) - d(t) \cdot S(t)] \tag{22}$$

Since $S(t)$ is not related to $N_1(t)$, and $S(t)$ is not related to $N_{2(}t)$, $2E[(N(t) - d(t) \cdot S(t)] = 0$:

$$E[e^2(t)] = E[s^2(t)] + E[N_1(t) - d(t)]^2 \tag{23}$$

The weight coefficient is adjusted by the LMS adaptive filter to obtain the minimum point of the nonlinear function $E[e^2(t)]$. When the value of $E[e^2(t)]$ in Equation (23) is minimum, the value of $E[N_1(t) - d(t)]^2$ in Equation (23) is also minimum. When the value of $E[s^2(m)]$ does not change, the output of the adaptive filter $d(t)$ is the best estimate of $N_1(t)$, and the system output is:

$$e(m) = s(t) + N(t) - d(t) \tag{24}$$

In this way, when the value of $d(t)$ is closest to the value of $N(t)$, the output of the adaptive LMS filter is $e(m) = s(m)$.

This paper proposes a speech enhancement algorithm based on post Wiener filtering. First, the algorithm uses spectral subtraction to perform speech enhancement on the speech signal of the current sound source, which will obtain acoustic signal containing autocorrelation noise. Then, the parameters of the post-wiener filter are used to suppress noise and amplify the target of the sound source signal. Finally, the algorithm fits the optimal filter. The principal flow chart of the optimization algorithm is shown in Figure 6.

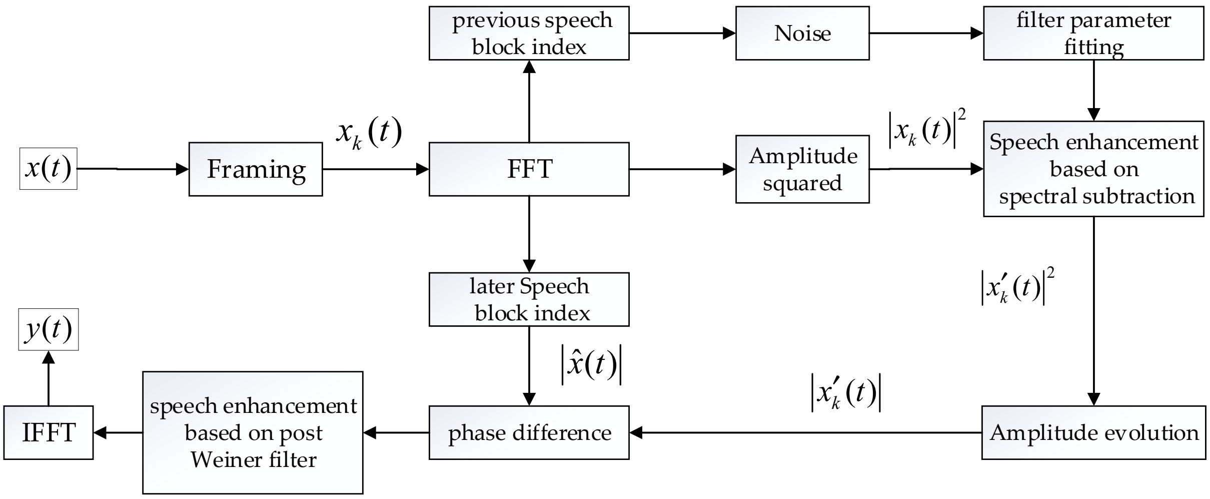


Figure 6. Adaptive speech enhancement flow.

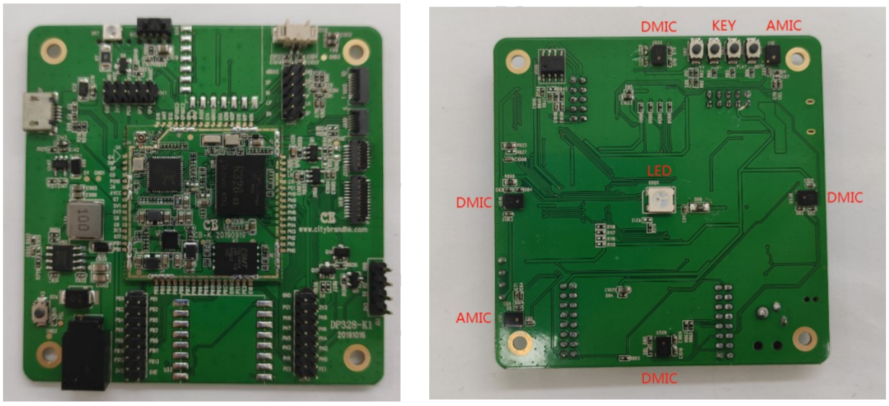
5. Experiment Design and Result Analysis

In order to verify the real performance and effectiveness of the algorithm in this paper, two experiments were designed. Experiment-I verifies the performance of this algorithm in real sound localization. Selecting some sound source points, we calculate the sound source position and compare it with the existing sound source localization method (based on the TDOA algorithm). Experiment-∏ is to verify the effectiveness of the proposed new method in speech enhancement. We process speech signals in different noise environments and compare them with other speech enhancement algorithms.

The experimental hardware uses the Allwinner R328 microphone array. Allwinner R328 relies on the computing ability of the cost-effective dual-core CortexTM-A7 CPU to provide the best computing ability at the lowest cost. The highly integrated CODEC can support key voice pick-and-place solutions without external DSP voice chip circuits. As shown in Figure 7, the Allwinner R328 microphone array has six microphones, including two digital microphones and four analog microphones. The back of Allwinner R328 microphone array also has four keys to adjust the recording volume and a LED to indicate that the device is working normally. In the experiment, only two digital microphones were used for recording. The distance between the two digital microphones is 15 cm, so the value of d in Equation (2) is 0.2.

When the voice signal is sampled, the two digital microphones on the array are used as recording devices. The distance between the two microphones is 20 cm, and the sampling rate is 16 KHz.

The experimental site was chosen as a hall of 10 m × 15 m × 4 m. The early reverberation time of the room is calculated to be 15 m through experiments. In living environment, there are many kinds of noises such as other people talking, air conditioners, and computer fans.



(**a**) (**b**)

Figure 7. Allwinner R328 microphone array physical picture: (**a**) the front of Allwinner R328 microphone array; (**b**) the back of Allwinner R328 microphone array.

5.1. Acoustic Localization Experiment by Dual-Microphone

Firstly, we build a test prototype for the collection of the circular microphone array, and the programming the development board. In the experiment, the USB interface is used to connect with the PC, which not only supplies power to the hardware circuit, but also transmits the processed voice signal to the computer.

To test the dual-microphone sound source localization function, there are other speakers speaking in the laboratory to interfere with the target speaker's voice signal, while one audio is also set to play different volume of interference noise (the noise level is divided into three levels according to the volume of the sound, and the three-level noise interference is the most serious).

Determining the accuracy of the acoustic azimuth angle measurement, the target speaker stands at different angle positions 4 m away from the center of the microphone. In the serial port tool, entering the relevant commands, the development board will record the target speaker and calculate its azimuth angle. In the actual positioning experiment, the measurement is repeated five times at each experimental point, and the average value is taken as the final positioning result of the point.

We then conduct experiments on the accuracy of sound source distance measurement. Under four noise environments, the target speaker stands at the same angular position from different distances to the center of the microphone. Then we use the previous method to perform recording and 3D distance calculation. Similarly, the measurement is repeated five times at each experimental point, and the average value is taken as the final positioning result of the point. The experimental results are shown in the Figure 8.

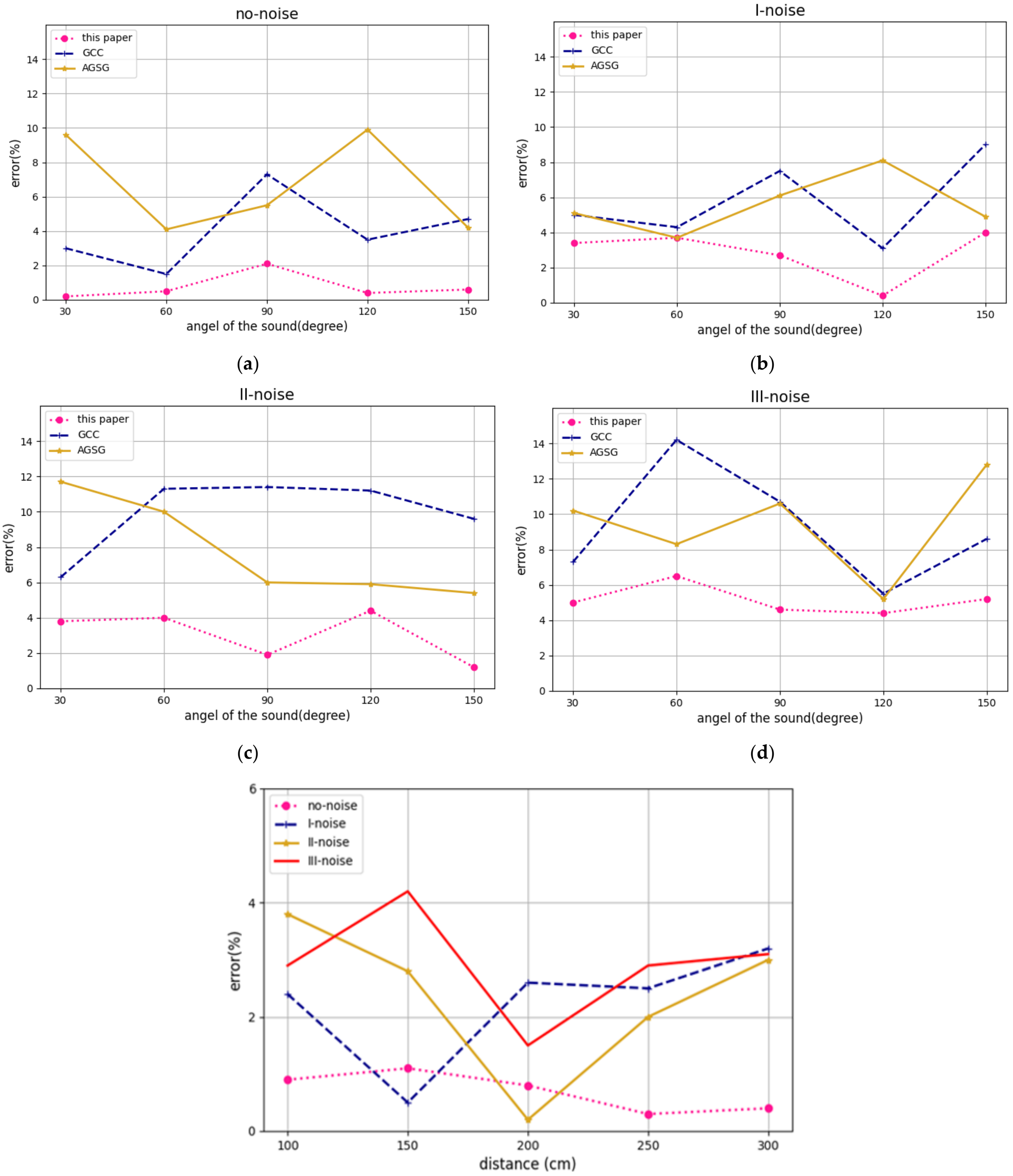
5.2. Speech Enhancement Experiment

5.2.1. Known Noise Simulation Experiment

In the simulation experiment, we will select 20 groups of speech files in a noise-free scene as clean speech signal. There are 4 kinds of noise in NOISE-92, which are babble, street, car and train. The SNR of added noise are -5 dB, 0 dB, and 5 dB. The sampling rate is 16 kHz. The quantization precision is 16 bits.

Perceptual evaluation of speech quality (PESQ) is an objective, full-reference voice quality assessment method. The PESQ algorithm requires a noisy attenuated signal and an original reference signal, which can provide an evaluation criterion for speech. The PESQ score is from -0.5 to 4.5. The higher the score, the better the voice quality.

Table 1 shows the quality value of noisy speech (not enhanced by the enhancement algorithm), the quality value enhanced by the GCC/AGSC algorithm, and the quality value after speech enhancement algorithm proposed in this paper. The processing standard of the two algorithms is controlled in the same way, and this quality value is the average of the 20 groups of speech files.



(**a**) (**b**)

(**c**) (**d**)

(**e**)

Figure 8. The experiment result of localization experiment: (**a**) the comparison of azimuth error around no-noise; (**b**) the comparison of azimuth error around I-noise; (**c**) the comparison of azimuth error around II-noise; (**d**) the comparison of azimuth error around III-noise; (**e**) distance error around noise for this paper.

Table 1. The comparison of the PESQ value.

The Type of Noise	SNR	The Speech with Noise	GCC-Enhanced Speech	AGSC-Enhanced Speech	The Speech Enhanced by the Algorithm in This Paper
babble	−5 dB	1.38	1.35	1.53	1.74
	0	1.69	1.55	1.59	1.85
	5 dB	2.12	1.97	1.98	2.26
street	−5 dB	1.23	1.15	1.28	1.55
	0	1.72	1.61	1.80	1.96
	5 dB	2.21	2.16	2.27	2.29
car	−5 dB	1.48	1.33	1.46	1.77
	0	1.89	1.67	1.92	1.91
	5 dB	2.45	2.44	2.43	2.63
train	−5 dB	1.27	1.30	1.28	1.46
	0	1.56	1.67	1.66	1.91
	5 dB	2.17	2.18	2.15	2.52

It can be seen from the table that the algorithm proposed in this paper has a higher quality value than the noisy speech and GCC/AGSC algorithm under all noise conditions, which proves that the algorithm proposed in this paper can greatly improve the enhanced speech quality. We compared the PESQ between the algorithm proposed in this paper and the GCC/AGSC algorithm to intuitively show the improvement of PESQ. From Table 1, it is clear that the PESQ are increased by the algorithm proposed in this paper is improved under all three kinds of SNR conditions. Except for babble, the lower the signal noise ratio, the higher the quality value. The algorithm proposed herein is more advantageous to improve the quality of the speech under low signal noise ratio.

5.2.2. Unknown Noise Reality Simulation Experiment

In the speech enhancement experiment, the acoustic files in the first-level noise and the third-level noise environment are selected to perform subsequent enhancement processing on the acoustic signal. According to the foregoing, the azimuth angle information of the sound source is used as a feature vector for acoustic signal separation. The voice system will only amplify the voice signal from this position and suppress other signals to achieve voice enhancement. Finally, the advantages of the algorithm in this paper are demonstrated through comparative experiments.

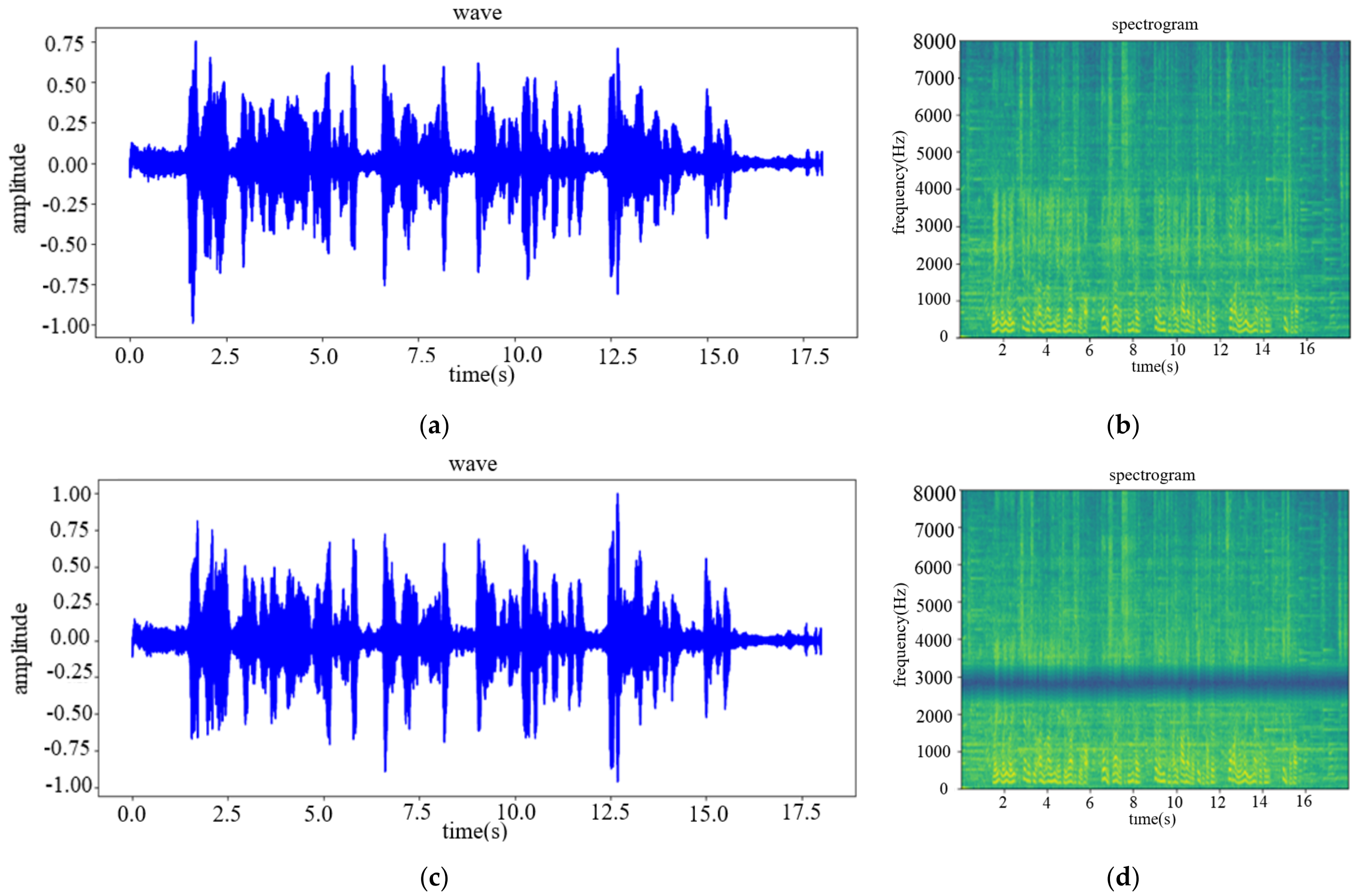
This paper uses the experimental data to test the technical solution in the laboratory and compares speech enhancement effect of the algorithm in this paper with GCC algorithm and AGSC algorithm. Figure 9 shows the high-noisy experimental speech and the output results of the two algorithms. Figure 10 shows the low-noisy experimental speech and the output results of the two algorithms.

From the comparison of the spectrogram, it can be found that when the GCC algorithm and the AGSC algorithm enhance the dual-channel speech signal, there will be auto-correlation noise and speech distortion; while the speech enhancement algorithm in this paper has neither obvious auto-correlation noise nor speech distortion. In addition, from the speech waveform information, the GCC algorithm and the AGSC algorithm have no accuracy of the sound source azimuth estimation with low SNR of the acoustic signal, which affects the speech enhancement performance. While the speech enhancement algorithm in this paper has better effect of background noise reduction and acoustic source target signal amplification.

Finally, in order to verify the comprehensibility of the corpus enhanced by the algorithm in this paper, eight speech files in the experiment were sequentially used for speech recognition by the speech transcribing module of iFLYTEK. In each speech file, the speaker said a total of 52 Chinese characters. The correct rate of speech recognition for each corpus is shown in Table 2.

Table 2. The correct rate of speech recognition.

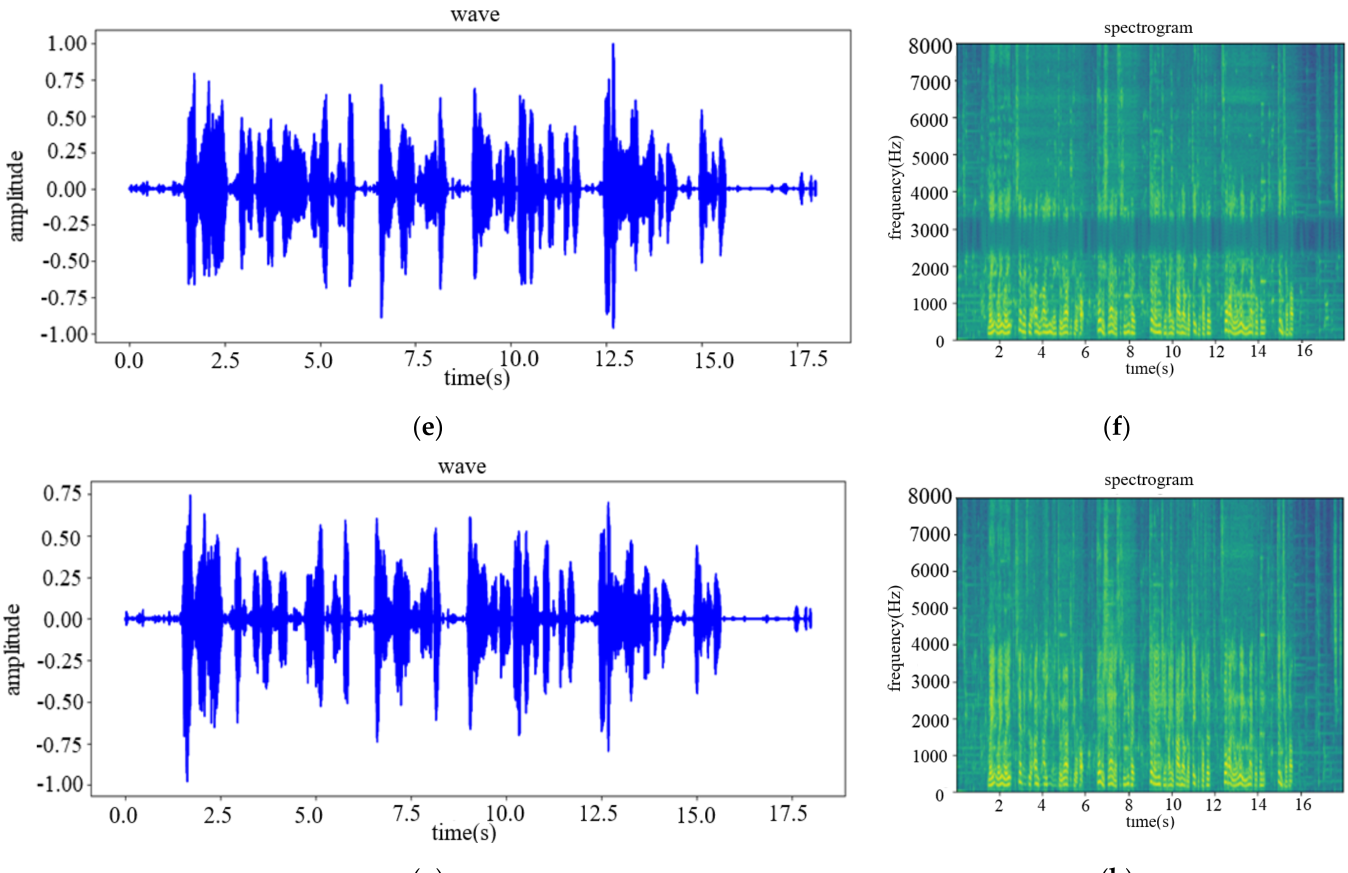
Type	Test Environment	Correct Rate (%)
Original speech file	low-noise high-noise	67.31 57.69
GCC-enhanced speech file	low-noise high-noise	80.77 73.77
AGSC-enhanced speech file	low-noise high-noise	90.38 78.85
the speech file enhanced by the algorithm in this paper	low-noise high-noise	100 98.77



(**a**) (**b**)

(**c**) (**d**)

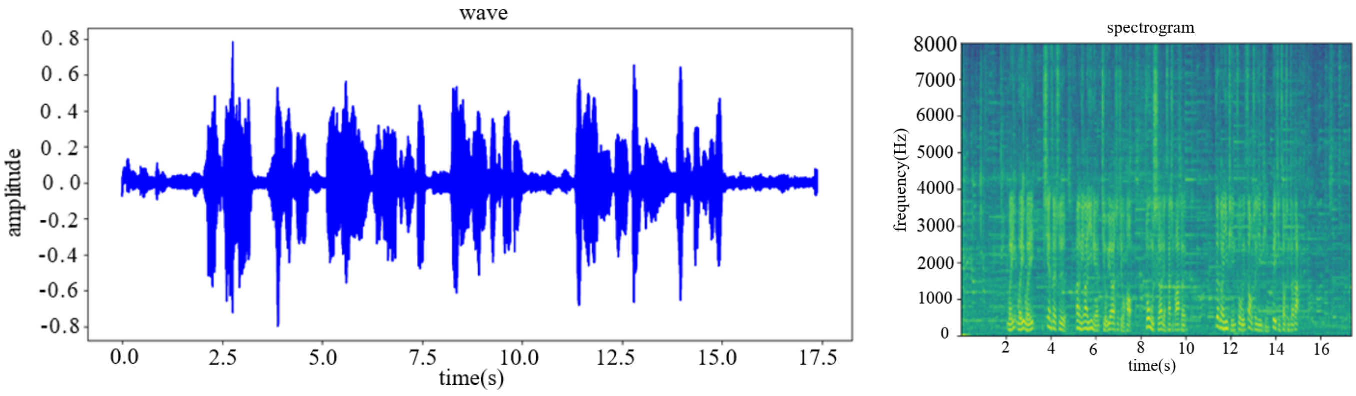
Figure 9. *Cont.*



(**e**) (**f**)

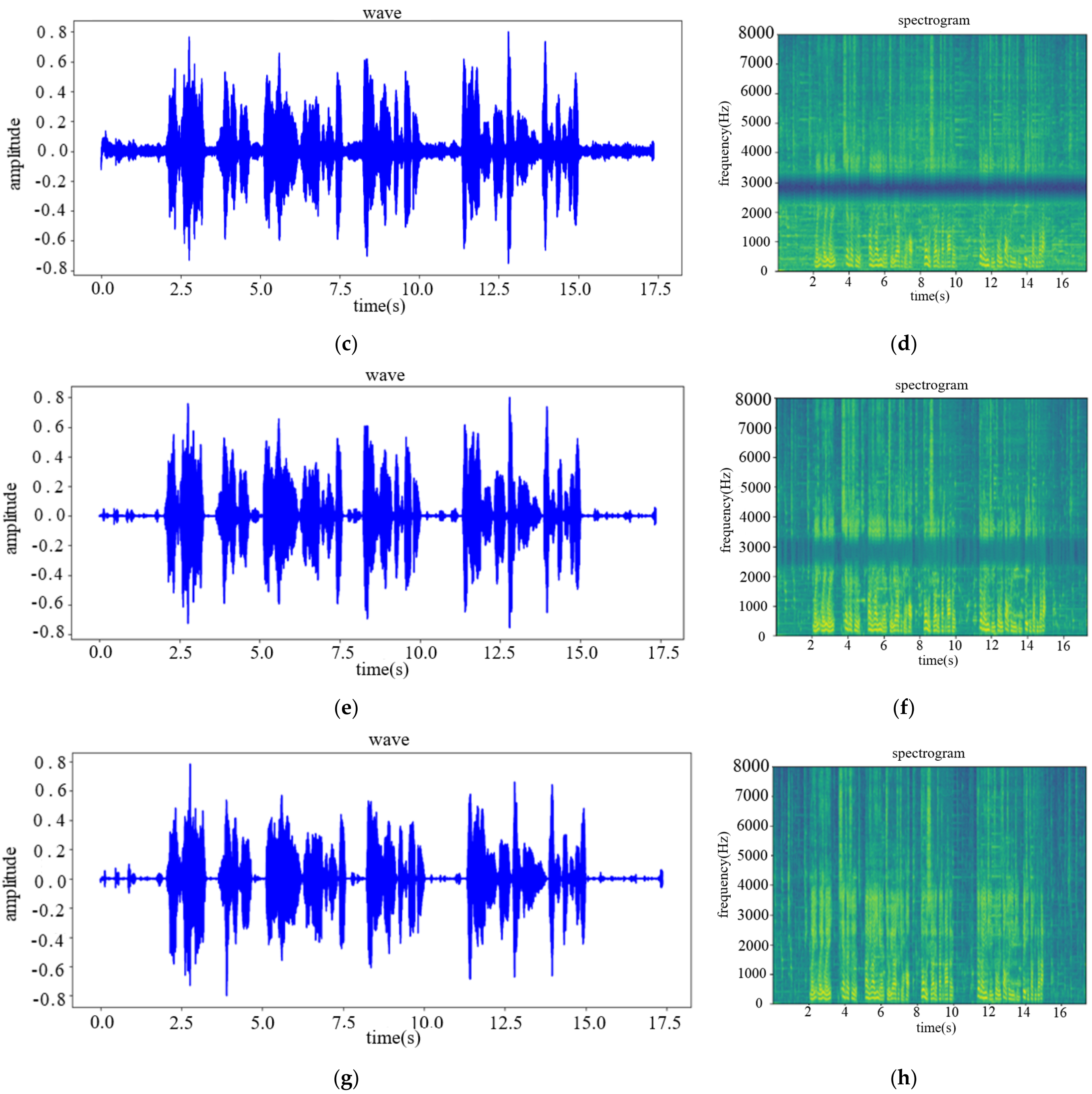
(**g**) (**h**)

Figure 9. The results of speech enhancement comparison experiment in high-noise: (**a**) Original speech wav; (**b**) Original speech spectrogram; (**c**) GCC-enhanced speech wave; (**d**) GCC-enhanced speech spectrogram; (**e**) AGSC-enhanced speech wave; (**f**) AGSC-enhanced speech spectrogram; (**g**) the speech wave enhanced by the algorithm in this paper; (**h**) the speech spectrogram enhanced by the algorithm in this paper.



(**a**) (**b**)

Figure 10. *Cont.*



(**c**) (**d**)

(**e**) (**f**)

(**g**) (**h**)

Figure 10. The results of speech enhancement comparison experiment in low-noise: (**a**) Original speech wave; (**b**) Original speech spectrogram; (**c**) GCC-enhanced speech wave; (**d**) GCC-enhanced speech spectrogram; (**e**) AGSC-enhanced speech wave; (**f**) AGSC-enhanced speech spectrogram; (**g**) the speech wave enhanced by the algorithm in this paper; (**h**) the speech spectrogram enhanced by the algorithm in this paper.

6. Conclusions

This paper proposes an improved sound source localization and speech enhancement algorithm. By introducing the maximum controllable response power, based on the traditional time delay estimation, combined with the energy attenuation estimation, only two microphones are needed to complete the position settlement of the sound source in the three-dimensional space, which simplifies the design complexity and reduce cost of the microphone array. It also improves the accuracy of the sound source localization algorithm.

Then the results of sound source localization are used to realize speech separation based on the azimuth of the target speaker, and complete speech enhancement based on adaptive filtering, and output a corpus with a higher SNR. Finally, related experiments are completed in combination with actual scenarios and hardware construction. The experimental results show that the dual-microphone-based sound source localization and speech enhancement algorithm proposed in this paper has extremely high accuracy and robustness. Compared with other speech enhancement algorithm, the corpus enhanced by the algorithm in this paper has a higher SNR.

Author Contributions: Conceptualization, T.T.; Data curation, T.T. and X.T.; Funding acquisition, J.Y.; Investigation, T.T. and J.A.; Methodology, W.L.; Resources, H.Z.; Software, T.T. and Y.Z.; Supervision, H.Z. and Z.G.; Visualization, Y.C.; Writing—original draft, T.T.; Writing—review & editing, T.T. All authors have read and agreed to the published version of the manuscript.

Funding: This work was funded by the National Key Research and Development Program of China (No. 2020YFF0304902) and the National Natural Science Foundation of China (No. 61771352).

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

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

References

1. Okuno, H.G.; Nakadai, K. Robot audition: Its rise and perspectives. In Proceedings of the 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2015), Brisbane, Australia, 19–24 April 2015; pp. 5610–5614.
2. Nakadai, K.; Takahashi, T.; Okuno, H.G.; Nakajima, H.; Hasegawa, Y.; Tsujino, H. Design and Implementation of Robot Audition System 'HARK'—Open Source Software for Listening to Three Simultaneous Speakers. *Adv. Robot.* **2010**, *24*, 739–761. [CrossRef]
3. Hoshiba, K.; Washizaki, K.; Wakabayashi, M.; Ishiki, T.; Kumon, M.; Bando, Y.; Gabriel, D.; Nakadai, K.; Okuno, H.G. Design of UAV-embedded microphone array system for sound source localization in outdoor environments. *Sensors* **2017**, *17*, 2535. [CrossRef] [PubMed]
4. Seltzer, M.L. Microphone Array Processing for Robust Speech Recognition. Ph.D. Thesis, Carnegie Mellon University, Pittsburgh, PA, USA, 2003.
5. DiBiase, J.H. A High-Accuracy, Low-Latency Technique for Talker Localization in Reverberant Environments Using Microphone Arrays. Ph.D. Thesis, Brown University, Providence, RI, USA, 2000.
6. Sallai, J.; Hedgecock, W.; Volgyesi, P.; Nadas, A.; Balogh, G.; Ledeczi, A. Weapon classification and shooter localization using distributed multichannel acoustic sensors. *J. Syst. Arch.* **2011**, *57*, 869–885. [CrossRef]
7. Blumstein, D.T.; Mennill, D.J.; Clemins, P.; Girod, L.; Yao, K.; Patricelli, G.; Deppe, J.L.; Krakauer, A.H.; Clark, C.; Cortopassi, K.A.; et al. Acoustic monitoring in terrestrial environments using microphone arrays: Applications, technological considerations and prospectus. *J. Appl. Ecol.* **2011**, *48*, 758–767. [CrossRef]
8. Ganguly, A.; Reddy, C.; Hao, Y.; Panahi, I. Improving Sound Localization for Hearing Aid Devices Using Smartphone Assisted Technology. In Proceedings of the 2016 IEEE International Workshop on Signal Processing Systems (SiPS), Dallas, TX, USA, 26–28 October 2016; pp. 165–170.
9. Li, J.Z.; Liu, H. Research on Sound Localization Algorithm of Five-element Cross Microphone Array. *Inf. Technol. Informatiz.* **2021**, *9*, 4246. [CrossRef]
10. Tiete, J.; Domínguez, F.; Silva, B.D.; Segers, L.; Steenhaut, K.; Touhafi, A. Sound-Compass: A Distributed MEMS Microphone Array-Based Sensor for Sound Source Localization. *Sensors* **2014**, *14*, 1918–1949. [CrossRef] [PubMed]
11. Xing, H.; Yang, X. Sound source localization fusion algorithm and performance analysis of a three-plane five-element microphone array. *Appl. Sci.* **2019**, *9*, 2417. [CrossRef]
12. Shujau, M.; Ritz, C.H.; Burnett, I.S. Speech enhancement via separation of sources from co-located microphone recordings. In Proceedings of the IEEE International Conference on Acoustics Speech & Signal Processing, Dallas, TX, USA, 14–19 March 2010; pp. 137–140.
13. Zhu, C.Y. Research on Speech Enhancement of Small-Scale Microphone Array Based on Deep Learning. Ph.D. Thesis, Zhejiang University, Hangzhou, China, 2021.
14. Jia, H.R.; Mei, S.L.; Zhang, M. Dual channel neural network speech enhancement algorithm based on time frequency masking. *J. Huazhong Univ. Sci. Technol. (Nat. Sci. Ed.)* **2021**, *49*, 43–49. [CrossRef]
15. Alonso-Martín, F.; Gorostiza, J.F.; Malfaz, M.; Salichs, M.A. User localization during human-robot interaction. *Sensors* **2012**, *12*, 9913–9935. [CrossRef] [PubMed]
16. Kilis, N.; Mitianoudis, N. A novel scheme for single-channel speech dereverberation. *Acoustics* **2019**, *1*, 42. [CrossRef]
17. Alam, F.; Usman, M.; Alkhammash, H.I.; Wajid, M. Improved Direction-of-Arrival Estimation of an Acoustic Source Using Support Vector Regression and Signal Correlation. *Sensors* **2021**, *21*, 2692. [CrossRef] [PubMed]

18. Min, X.Y.; Wang, Q.L.; Ran, Y.F. Speech enhancement algorithm based on microphone array. *Comput. Eng. Des.* **2020**, *41*, 1074–1079. [CrossRef]
19. Su, D.; Miro, J.V.; Vidal-Calleja, T. Real-time sound source localisation for target tracking applications using an asynchronous microphone array. In Proceedings of the 2015 IEEE 10th Conference on Industrial Electronics and Applications (ICIEA), Auckland, New Zealand, 15–17 June 2015; pp. 1261–1266.
20. Fallon, M.F.; Godsill, S.J. Acoustic source localization and tracking of a time-varying number of speakers. *IEEE Trans. Audio Speech Lang. Process.* **2011**, *20*, 1409–1415. [CrossRef]
21. Han, J.H. Tracking control of moving sound source using fuzzy-gain scheduling of PD control. *Electronics* **2020**, *9*, 14. [CrossRef]
22. Dong, Y.Y.; Dong, C.X.; Liu, W.; Cai, J. Generalized $\ell 2-\ell$p minimization based DOA estimation for sources with known waveforms in impulsive noise. *Signal Process.* **2022**, *1*, 108313. [CrossRef]
23. Cho, B.J.; Lee, J.M.; Park, H.M. A beamforming algorithm based on maximum likelihood of a complex Gaussian distribution with time-varying variances for robust speech recognition. *IEEE Signal Process. Lett.* **2019**, *26*, 1398–1402. [CrossRef]
24. Gannot, S.; Cohen, I. Speech enhancement based on the general transfer function GSC and postfiltering. *IEEE Trans. Speech Audio Process.* **2004**, *12*, 561–571. [CrossRef]
25. Deng, T.; Zheng, R.; Jun, J. Towards Robust Multiple Blind Source Localization Using Source Separation and Beamforming. *Sensors* **2021**, *21*, 205. [CrossRef]
26. Leng, Y.H.; Zheng, C.S.; Li, X.D. Fast independent vector analysis using power ratio correlation-based bands partition. *J. Signal Process.* **2019**, *35*, 1314–1323. [CrossRef]