Journal of Electronic Research and Application, 2025, Volume 9, Issue 4
http://ojs.bbwpublisher.com/index.php/JERA
ISSN Online: 2208-3510
ISSN Print: 2208-3502

Real-Time Sound Source Localization Method Based on Selective SRP-PHAT and Vision Fusion

Jinde Huang*

Guangxi Police College, Nanning 530028, Guangxi, China

**Author to whom correspondence should be addressed.*

Abstract: Aiming at the problem that the traditional SRP-PHAT sound source localization method performs intensive search in a 360-degree space, resulting in high computational complexity and difficulty in meeting real-time requirements, an innovative high-precision sound source localization method is proposed. This method combines the selective SRP-PHAT algorithm with real-time visual analysis. Its core innovations include using face detection to dynamically determine the scanning angle range to achieve visually guided selective scanning, distinguishing face sound sources from background noise through a sound source classification mechanism, and implementing intelligent background orientation selection to ensure comprehensive monitoring of environmental noise. Experimental results show that the method achieves a positioning accuracy of ±5 degrees and a processing speed of more than 10FPS in complex real environments, and its performance is significantly better than the traditional full-angle scanning method.

Keywords: Sound source localization; SRP-PHAT; Audio-visual fusion; Real-time processing; Microphone array

Online publication: August 7, 2025

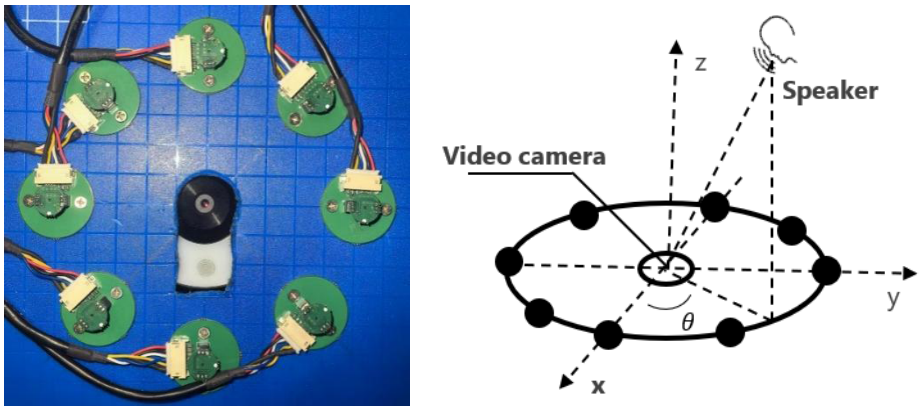
1. Introduction

Direction of Arrival (DOA) technology is widely used in video conferencing, intelligent monitoring, and robot interaction [1]. Traditional methods such as beamforming and time difference of arrival (TDOA) have significantly reduced performance in reverberant environments [2]. The SRP-PHAT algorithm is highly robust to reverberation and noise due to its phase transformation weighted characteristics, but its computational complexity is proportional to the size of the search space [3,4].

This method combines real-time visual analysis, proposes a visually guided selective SRP-PHAT scanning strategy, and designs a sound source classification mechanism and a dynamic background orientation selection algorithm. Experiments were conducted on equipment and complex environments, and the results showed that the algorithm speed was increased by 86%, accurately distinguishing the face sound source and background noise, and achieving a positioning accuracy of ±5 degrees.

2. Signal model

In order to facilitate sound source localization (DOA) with visual fusion, a uniform circular array of 8 isotropic microphones is designed, and the camera is placed at the center of the circle. The actual object is shown in **Figure 1 (a)**. This method considers the sound source to be the sound of human speech, with a frequency range of 500–2000Hz and a sound speed of c=343m/s. In order to avoid the problem of spatial sampling overlap [reference] and consider the convenience of calculation, the diameter of the designed uniform circular array is 10cm, and the array logic structure is shown in **Figure 1 (b)**.



(a) Physical structure of the array (b) Array logical structure

Figure 1. Microphone array structure

The coordinates of the array element are [5],

$$\begin{cases} x_i = 0.05 \cdot cos\left(\frac{\pi}{4}i\right) \\ y_i = 0.05 \cdot sin\left(\frac{\pi}{4}i\right) \\ z_i = 0 \end{cases}, i = 0,1,\ldots,7 \quad (1)$$

Because the array is a planar array, therefore z_i=0, it also means that the array can only estimate the azimuth of the sound source.

As shown in **Figure 1(b)**, the azimuth angle of the sound source projected onto the plane where the array is located is θ, then the signal received by the mth microphone $x_m(t)$ is as shown in equation (2),

$$x_m(t) = s(t - \tau_m(\theta)) + n_m(t) \quad (2)$$

where, $S(t)$ is the source signal, $\tau_m(\theta)$ is the theoretical delay of microphone m, and $n_m(t)$ is the additive noise.

3. Core algorithm

3.1. Basic SRP-PHAT algorithm

In **Figure 1(b)**, for the microphone pair consisting of the mth and nth array elements, the theoretical delay between them can be expressed by equation (3),

$$\tau_{mn}(\theta) = \frac{(\vec{r}_m - \vec{r}_n)\cdot\vec{u}(\theta)}{c} \quad (3)$$

among them, $\vec{r}_m$ and $\vec{r}_n$ position vectors of microphones m and n respectively, $\vec{u}(\theta)$ is the unit vector of the

sound source direction, C is the speed of sound, θ and is the angle of the sound source direction.

In fact, if the azimuth θ is unknown, $\tau_{mn}(\theta)$ it must be estimated in other ways. The S RP-PHAT method is $\tau_{mn}(\theta)$ a suitable method for estimation. The SRP-PHAT method delays the received signal by the array element for different time lengths, and calculates the cross-power spectrum for each time length delay. The delay time length that can maximize the cross-power spectrum is the actual delay corresponding to the actual sound source azimuth. The PHAT weighted cross power spectrum is defined by formula (4),

$$\mathrm{PHAT}_{mn}(f) = \frac{X_m(f)X_n^*(f)}{|X_m(f)X_n^*(f)|+\epsilon} \tag{4}$$

among them, $X_m(f)$ and $X_n(f)$ represent the complex spectra of microphones m and n at frequency f, respectively, $X^*_n(f)$ is the complex conjugate of $X_n(f)$, $|\cdot|$ is the modulus (amplitude) of the complex number, ϵ and is a small positive number (to prevent division by zero errors, usually 1e-10). By weighting the cross-power spectrum so that only phase information is retained and amplitude information is ignored, the algorithm is more robust to reverberation.

SRP (Steered Response Power) in a specific direction θ, as shown in formula (5),

$$P(\theta) = \sum_{f\in\mathcal{F}} \sum_{m=0}^{6} \sum_{n=m+1}^{7} \Re\{\mathrm{PHAT}_{mn}(f) \cdot e^{-j2\pi f \tau_{mn}(\theta)}\} \tag{5}$$

where, $\mathcal{F}$ is the effective frequency band (500-4000Hz), m,n is the microphone index (8 microphones), $\Re\{\}$ is the real part of the complex number, and j is the imaginary unit.

Using equation (5), we scan the 360° direction space (usually with a step size of 1°), calculate P(θ) for each direction θ, and find the direction that maximizes P(θ), which can be expressed by equation (6),

$$\hat{\theta} = \arg\max_{\theta} P(\theta) \tag{6}$$

This direction $\hat{\theta}$ is the estimated sound source direction.

3.2. Vision-guided scanning strategy

In the basic SRP-PHAT algorithm, intensive search is required in the 360° azimuth space (usually 1° step), resulting in a computational complexity of up to O ($N_\theta \times N_f \times N_{pairs}$), where N_θ =360, N f is about 350 (500-4000Hz frequency band), N_{pairs} =28 (the number of combinations of 8 microphones), and a total of about 3.5 million calculations, which is difficult to meet real-time processing requirements. To address this challenge, an innovative vision-guided scanning strategy is proposed, which reduces the scanning points by 86% by combining real-time face detection information while maintaining positioning accuracy.

3.2.1. Face angle calculation and key point extraction.

First, real-time face detection is performed using the OpenCV DNN module, using the pre-trained model of the SSD framework [6]. For each detected face, its bounding box and center point coordinates are calculated (x_{face},y_{face}). The center point of the image is used (x_c,y_c) as the origin of the reference coordinate system, and the face direction angle is calculated as shown in formula (7),

$$\theta_{face} = \arctan2\left(\frac{y_c - y_{face}}{x_{face} - x_c}\right) \times \frac{180}{\pi} \tag{7}$$

where arctan2 is the four-quadrant inverse tangent function, ensuring that the angle calculation is [0°,360°) within range.

To further determine the spatial range of the face, the system extracts five key points: the four corner points of the bounding box (x_1,y_1), (x_1,y_2), (x_2,y_1), (x_2,y_2). Each key point calculates its azimuth $\theta_{keypotint}$ to form a face angle set, as shown in formula (8),

$$\mathcal{A}_{face} = \{\theta_{keypoint,i} | i = 1, ..., 5 \times N_{faces}\} \tag{8}$$

3.2.2. Face angle range modeling and boundary processing

In order to accurately represent the orientation space occupied by the face, the minimum and maximum angles are calculated, as shown in formula (9),

$$\theta_{face}^{min} = min(\mathcal{A}_{face}), \theta_{face}^{max} = max(\mathcal{A}_{face}) \tag{9}$$

When the angle range crosses the 0°/360° boundary, (e.g., $\theta_{face}^{min} = 350°$, $\theta_{face}^{max} = 10°$), it is divided into two intervals, as shown in formula (10),

$$\mathcal{R}_{face} = \begin{cases} \{[\theta_{face}^{min}, \theta_{face}^{max}]\} & , \text{ if } \theta_{face}^{max} - \theta_{face}^{min} \leq 180° \\ \{[\theta_{face}^{min}, 360°], [0°, \theta_{face}^{max}]\} & , \text{ otherwise} \end{cases} \tag{10}$$

This processing ensures that the face area is correctly represented in the orientation space and avoids scanning blind areas caused by boundary effects.

The background orientation step is obtained through formula (10).

(1) Generate candidate angle sets:

$$\Theta_{candidate} = \{0°, 45°, 90°, 135°, 180°, 225°, 270°, 315°\}$$

(2) Filter the face area:

$$\Theta_{filtered} = \{\theta \in \Theta_{candidate} | \theta \notin \mathcal{R}_{face}\}$$

(3) Select background orientation:

(a) If $|\Theta_{filtered}| \geq 4$: 4 angles are selected evenly

(b) If $|\Theta_{filtered}| < 4$:

(i) Randomly generate new angles $\theta_{new} \sim U(0,360)$

(ii) If $\theta_{new} \notin \mathcal{R}_{face}$, add to $\theta_{new} \notin \mathcal{R}_{face}$

(iii) Repeat until $|\Theta_{filtered}| = 4$

(4) Supplement fixed angles: If it is still insufficient, add {0°,90°,180°,270°} available angles.

3.2.3. Background orientation intelligent selection algorithm

Comprehensive monitoring of ambient sound sources, the system selects 4 background positions $\mathcal{A}_{bg}$ to ensure:

(1) Uniform coverage of 360° space

(2) Avoid the face area

(3) Provide environmental noise reference

Therefore, the background orientation selection algorithm is defined as in formula (11),

$$\mathcal{A}_{bg} = \{\theta_i | \theta_i \notin \cup \mathcal{R}_{face}\}_{i=1}^{4} \tag{11}$$

3.2.4. Final scan set construction

The scan set $\mathcal{A}_{scan}$ is the union of the face key point angle set and the background orientation set, as shown in formula (12),

$$\mathcal{A}_{\text{scan}} = \mathcal{A}_{face} \cup \mathcal{A}_{bg} \tag{12}$$

The method deduplicates and sorts the scan set, and the average number of scan points is 12.4±3.2 (experimental statistical value), which reduces the computational effort by 96.6% compared to traditional 360-point scanning.

3.3. Sound source classification mechanism

achieves reliable distinction between the speaker's voice source and background noise by comparing the maximum SRP values of the face area and the background area. The specific decision rule is shown in formula (13),

$$\text{SourceType} = \begin{cases} \text{Speaker} & \text{,if } max(P(\mathcal{A}_{\text{face}})) > \alpha \cdot max(P(\mathcal{A}_{\text{bg}})) \\ \text{Background} & \text{,otherwise} \end{cases} \tag{13}$$

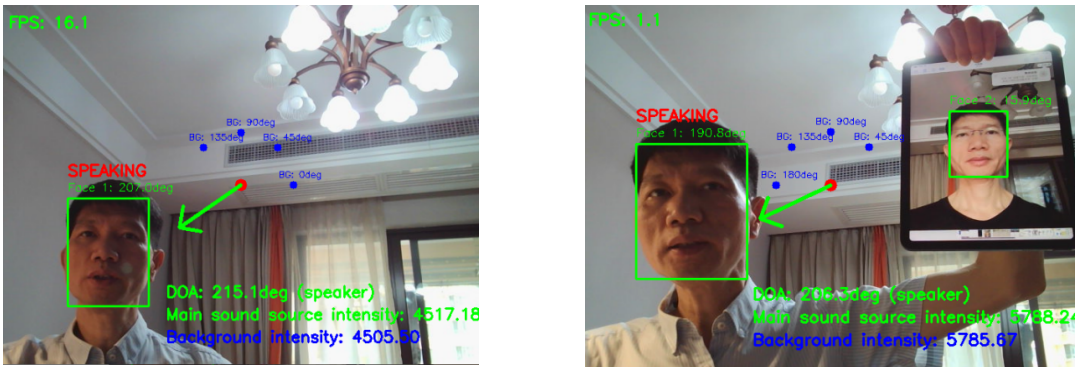
where α=1.2 is an adjustable threshold parameter.

4. Experimental results

Experiment was conducted in a standard office environment. The hardware conditions of the experiment were as follows: in order to simulate edge devices, the hardware devices selected had medium performance: 8-channel USB microphone array (sampling rate 48kHz), ordinary home camera (Logitech C270, resolution 720p), Intel(R) Core (TM) i5-8350U CPU @ 1.70GHz, 16GB RAM. The test scenarios were: (1) Single-person speech: the tester conducted 2 minutes of voice interaction in 8 directions from 0° to 315°; (2) Adding static avatar images to interfere with face detection (simulating a virtual meeting scene); (3) Background music interference: Adding pop music background (SNR=10dB).

By calibrating each position, the test is performed in real time. The result of a single experiment is shown in **Figure 3**. It contains: FPs - the number of frames that can be processed per second; Face - the angle between the face and the center of the corresponding face frame; DOA - the direction of the sound source. When the sound source is located in a face frame, the word SPEAKING in red font will be displayed above the face frame.

The quantitative evaluation was performed through three core indicators, and the results were compared with the traditional SRP-PHAT method, as shown in **Table 1**.



(a) Single speaker localization results (b) Localization results of a single person speaking with an avatar added

Figure 3. Results of a single speaker localization experiment

Table 1. Performance comparison

Performance indicators	Methods	Traditional SRP-PHAT	Performance improvements
Positioning error (°)	4.5 ± 2.1	4.8 ± 1.9	0.3°
Angular resolution (°)	1	1	-
Max FPS	15.5	1.6	+13.9FPS

The average FPS of the proposed method is 15.5FPS > 11FPS. Under the current experimental conditions, it can be displayed smoothly without any lag, which shows that the proposed algorithm has good real-time performance.

The performance of the sound source classification mechanism in different scenarios is shown in **Table 2.**

Table 2. Performance in different scenarios

Test scenario	Accuracy
Single-person, quiet environment	96.7%
Background music (SNR=10dB)	82.5%
Visual distraction scenes	85.3%

As can be seen from **Table 2**, the method proposed in this paper has a high accuracy in sound source localization, even in the presence of background noise and visual interference.

5. Conclusion

This paper proposes a real-time sound source localization method based on selective SRP-PHAT and visual fusion. The main contributions include the development of a visually guided selective scanning strategy that reduces the amount of computation by 86% while maintaining positioning accuracy; the design of a sound source classification mechanism that effectively distinguishes face sound sources from background noise; and the proposal of a dynamic background orientation selection algorithm to comprehensively monitor environmental sound sources. Experimental results show that the system has excellent positioning performance and real-time processing capabilities in complex acoustic environments.

Funding

This article is the research result of the 2024 Guangxi Higher Education Undergraduate Teaching Reform Project "OBE-Guided, Digitally Empowered 'Hadoop Big Data Development Technology' Course Ideological and Political Construction Innovation Exploration and Practice" (Project No.: 2024JGA396).

Disclosure statement

The author declares no conflict of interest.

References

[1] Schmidt RO, 1986, Multiple Emitter Location and Signal Parameter Estimation. IEEE Transactions on Antennas and Propagation, 34(3).

[2] Omologo M, Svaizer P, 1997, Use of the Crosspower-Spectrum Phase in Acoustic Event Location. IEEE Trans Speech Audio Process, 5(3): 288–292.

[3] Brumann K, Doclo S, 2024, Steered Response Power-Based Direction-of-Arrival Estimation Exploiting an Auxiliary Microphone. European Signal Processing Conference, 917–921.

[4] Li C, Hendriks RC, 2023, Alternating Least-Squares-Based Microphone Array Parameter Estimation for a Single-Source Reverberant and Noisy Acoustic Scenario, in IEEE/ACM Transactions on Audio, Speech, and Language Processing, 31, 3922–3934.

[5] Diaz-Guerra D, Miguel A, JR Beltran JR, 2021, Robust Sound Source Tracking Using SRP-PHAT and 3D Convolutional Neural Networks, in IEEE/ACM Transactions on Audio, Speech, and Language Processing, 29, 300–311.

[6] Alghareb FS, Hasan BT, 2025, Multitask Learning-Based Pipeline-Parallel Computation Offloading Architecture for Deep Face Analysis. Computers, 14: 29.