Efficient and robust approaches for three-dimensional sound source recognition and localization using humanoid robots sensor arrays

Hao Chen, Chengju Liu and Qijun Chen

Abstract
Efficient and robust sound source recognition and localization is one of the basic techniques for humanoid robots in terms of reaction to environments. Due to the fixed sensor arrays and limited computation resources in humanoid robots, there comes challenge for sound source recognition and localization. This article proposes a sound source recognition and localization framework to realize real-time and precise sound source recognition and localization system using humanoid robots’ sensor arrays. The type of the audio is recognized according to the cross-correlation function. And steered response power-phase transform function in discrete angle space is used to search the sound source direction. The sound source recognition and localization framework presents a new multi-robots collaboration system to get the precise three-dimensional sound source position and introduces a distance weighting revision way to optimize the localization performance. Additionally, the experiment results carried out on humanoid robot NAO demonstrate that the proposed approaches can recognize and localize the sound source efficiently and robustly.

Keywords
Auditory perception, sound source recognition and localization, cross-correlation, multi-robots collaboration system, humanoid robot

Introduction
Humanoid robots have been designed to interact with people and react to environments. As a representation of its intelligence, auditory perception is essential for its recognition of environmental changes. For example, humanoid robots in RoboCup competition need to recognize the whistle blown by referee as a signal to start the game. In this situation, the technology of sound source recognition and localization (SSRL) is employed to classify if the audio type is whistle and localize the accurate sound source position to avoid misidentification of whistle from other fields. Due to the fixed sensor arrays in humanoid robots and the fact that a sound event occurs at an arbitrary direction in three-dimensional (3D) space with noise and reverberation, there comes challenge for SSRL. In addition, a lot of techniques like visual recognition need to be carried out
simultaneously with the given computation resources limited humanoid robots. As a result, the computationally efficient and robust real-time 3D SSRL method is increasingly required.

In the last few years, many SSRL algorithms have been applied to intelligent robots. Confronted with sound recognition, frequency domain analysis is frequently adopted. Detected principal frequency components are classified to predefined frequency domain space to determine whether it is the certain audio type. Sensor arrays in humanoid robots, however, are usually fixed and equipped with low sample rate microphone. Thus, false recognition often occurs. To solve the problem, a new sound source recognition method which is not limited to frequency discrimination has been adopted in this article.

Despite the sensor arrays differ between various applications, commonly used sound source localization (SSL) cues are inter-channel time difference (ICTD) and inter-channel level difference. In Rascon and Meza, a survey on SSL in robotics is presented. The typical ways using ICTD such as generalized cross-correlation phase transformation (GCC-PHAT) and steered response power-phase transform (SRP-PHAT) estimate the sound source by time difference of arrival (TDOA) feature. Some multi signal classification (MUSIC)-based sound localization ways are presented in Takeda and Komatini, Birnie et al., and Hoshiba et al. The basic principle of MUSIC algorithm is to decompose the covariance matrix of the output data, obtaining signal subspace and noise subspace. And the incident direction of the signal is estimated by using the orthogonality of the two subspaces. However, these methods are often hard to meet strict performance requirements in humanoid robots, for example, limited computation resource, noisy environment, and the requirement of reliability and efficiency.

Some applications are implemented on robots equipped with fixed microphone array configurations in Valin et al., Lee et al., Grondin and Michaud, and Bustamante et al. In Valin et al., the mobile robot can localize the sound source over a range of 3 m and with a precision of 3°. In Athanasopoulos et al., an application on humanoid robot NAO is presented, which reaches a precision of 5° to locate the sound source at the same elevation as the microphone array’s horizontal plane. Furthermore, with more computation resources, a deep neural network for SSL is introduced in He et al., Zhang and Wang, Yalta et al., and Sun et al.

The methods mentioned above are still unable to meet the requirements in RoboCup soccer robot competition. In RoboCup SPL League, all teams must use NAO humanoid robots manufactured by Soft-Bank Robotics. The match is held on a total area of length 10.4 m and width 7.4 m. The whistle may be blown from same field or neighboring field as robots. So the strict requirements for distance range and positioning accuracy are challenging for SSRL algorithm. In 2019, a directional whistle challenge was held by organizing committee to investigate the possibilities of localizing the point where the referee whistle is blown. The details can be found on https://spl.robocup.org/technical-challenges-2019/. This article will take RoboCup SPL League competition and whistle challenge environment as one of the testbench of the methods proposed.

In this article, we propose an SSRL framework addressing the SSRL in humanoid robots. A cross-correlation-based recognition way is given to classify the audio type and a multi-robots collaboration system is used to locate the sound source. The cross-correlation-based recognition way takes prerecorded reference signal as basic clue rather than the frequency domain feature to distinguish the audio type. The single robot sound source direction estimator (SSDE) estimates the source direction by searching maximum SRP-PHAT function in discrete angle space while 2D multi-robots SSL algorithm and 3D SSL determine the source position. The SSRL aims at resolving the SSRL problem under indoor and outdoor circumstances. Figure 1 shows the structure of SSRL.

The contribution of this article is included as follows: (1) A robust and anti-noise sound recognition algorithm based on cross-correlation feature. (2) Improved SRP-PHAT function and simplified discrete angle space search with less computation cost. (3) Multi-robots collaboration localization system attaching distance weighting revision for more precise localization result. (4) Real-time application on humanoid robots NAO equipped with microphone arrays like human.

The rest of the article is structured as follows. In the second section, we discuss sound recognition algorithm
in SSRL, and the cross-correlation algorithm with noise filter is explained. We introduce SSL algorithm including 2D SSL and 3D SSL in the third section. Enhanced SRP-PHAT function and distance weighting revision is also presented in the third section. Experiments and result analysis are shown in the fourth section. Finally, in the fifth section, the article is concluded with a discussion and possible future research directions.

### Sound source recognition algorithm

Considering the existence of noise and reverberation mixed in sampled audio data, the misidentification may occur if we classify the audio type by only detecting the principal frequency components.

The main procedure of sound source recognition algorithm in this article is demonstrated in Algorithm 1. The audio data are acquired and sampled from microphone in robot’s head. After transferring the time-domain signal to frequency-domain signal using fast Fourier transform, a cross-correlation function between the signal and prerecorded reference signal is calculated to identify the certain audio type if its value is over the presetting threshold.

### Preprocessing: Noise filter

The noise in audio data origins from two aspects: the external noise from noisy environment and the internal noise caused by the processor’s ventilation fan, gyroscopes, and so on. It is essential to reduce the noise using noise filter before the recognition. A band-pass filter is carried out in our algorithm and the upper-lower limits are determined by the audio type. For instance, the frequency of whistle is 2.5–3.5 kHz in most cases, so we can set the limits of band-pass filter \( f_L - f_H \) as 1.5–4.5 kHz to filter out the noise. The noise filter procedure is executed before calculating the cross-correlation function to suppress the noise interference effectively.

### Cross-correlation function

The audio type identification cue in SSRL is cross-correlation function between the sampled audio data segments and prerecorded reference sound source.

Given two sound source signal \( x_1(n) \) and \( x_2(n) \), while \( x_1(n) \) is the audio segments to be analyzed and \( x_2(n) \) is the reference audio segments. The sampling frequency is both \( f \) and the number of samples is \( N \).

For two continuous signals, the cross-correlation function is defined as equation (1)

\[
R(t) = \mathbb{E}[x_1(m)x_2(m + t)]
\]

When we process the audio data in humanoid robots, the discrete sampling is carried out firstly. The cross-correlation function of two discrete signals is defined as equation (2), and the length of \( R(\tau_n) \) is \( 2N - 1 \)

\[
R(\tau_n) = \sum_{n=-N}^{N-1} x_1(n)x_2(n + \tau_n)
\]

However, the method below to calculate the cross-correlation is time-consuming in humanoid robots with limited computation resources. Hence, we need to do the calculation in frequency domain.

According to Wiener–Khinchin theorem, the auto-correlation function of a wide-sense-stationary random process has a spectral decomposition given by the power spectrum of that process

\[
P(\omega) = \int_{-\infty}^{+\infty} R(\tau)e^{-j\omega \tau} d\tau
\]

\[
R(\tau) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} P(\omega)e^{j\omega \tau} d\omega
\]

where \( P(\omega) \) is the cross power spectrum of \( x_1 \) and \( x_2 \)

\[
P(\omega) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} x_1(t)x_2(t + \tau)dt \cdot e^{-j\omega \tau} d\tau
\]
It can be simplified by the exchange integral property and the shift property of Fourier transform
\[ P(\omega) = F_1^*(\omega)F_2(\omega) \]  
(5)

Thus, the cross-correlation function can be presented as
\[ R(\tau) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} F_1^*(\omega)F_2(\omega)e^{i\omega \tau}d\omega \]  
(6)

And its discrete representation is
\[ R(\tau_n) = \frac{1}{N} \sum_{w=0}^{N-1} F_1^*(w)F_2(w)e^{j\frac{2\pi}{N}w} \]  
(7)

For any audio data segment from microphones, we calculate the cross-correlation function between it and the reference sound. The signal is identified as type T if the ratio of maximum value of cross-correlation function \( \max(R(\tau_n)) \) to maximum value of reference sound autocorrelation function \( \max(R'(\tau_n)) \) is larger than the presetting threshold of type T
\[ \frac{\max(R(\tau_n))}{\max(R'(\tau_n))} > \text{threshold}_T \]  
(8)

The empirical value of threshold is usually set to be 0.4 in practical use. As a result, we can deduce the audio type accurately based on the cross-correlation feature. The cross-correlation-based way can be more robust compared with principal frequency components detection-based way.

**SSL algorithm**

In this section, the proposed SSL algorithm is described. It consists of a single robot SSDE and multi-robots collaboration SSL algorithm. In “Single robot SSDE” section, a single robot SSDE will be introduced. “Multi-robot 2D SSL using distance weighting revision” and “Three-dimensional sound source localization” sections give more detailed introduction about 2D SSL algorithm and 3D SSL algorithm.

**Single robot SSDE**

As is depicted in Figure 2, the humanoid robot NAO is equipped with four microphones. The array configuration is analogous to human-ears as they are distributed on left and right sides. An important cue for SSL is the TDOA,\(^2,28,29\) but it is not enough to locate the sound source based on the TDOA model between two microphones because the cone of confusion will lead to a mirror-image position. The classical GCC-PHAT estimates the angle by TDOA under the assumption of far-field.\(^30\) As Figure 3 shows, the angle of sound arriving at mic1 and mic2 is \( \alpha \), and the difference of arrival distance is \( c\Delta T \), so we can infer the direction of sound source according to the geometric relationship based on TDOA and distance between microphones. In this section, we propose a single robot SSDE using all the four microphones to distinguish the source direction.

The theoretical basis of SSL is the TDOA model of each microphone pairs. When the sound source \( S_i \) is close to the microphone pairs, the sound wave received by microphone \( M_i \) can be considered as a polar wave. Under the polar wave propagation assumption, time of arrival of each microphone can be estimated as the ratio of sound-to-microphone distance to sound speed \( c \) in air
\[ \Delta T = \frac{||S_i - M_i||}{c} \]  
(9)

The coordinates that describe the source direction are shown in Figure 4.

Let the height of source be \( h \), SSDE describes the source direction using azimuth \( \alpha \) and elevation \( \beta \) with relative to robot’s coordinates. The SRP-PHAT feature is unique for each direction in discrete angle space, so the direction can be deduced by a set of microphone signals.

Given \( X_M(n) \) as the signal segments received by the \( M \)th microphone and \( q(x, y, z) \) as the assumption sound source position, the SRP-PHAT feature function\(^31\) is defined as equation (10)
Figure 4. The coordinates used in SSDE. 2D position and 3D position in the figure are hypothesis points as the height is assumed to be the height of a referee in a robot soccer game. SSDE: sound source direction estimator.

\[
\hat{P}(q) = \sum_{l=1}^{M} \sum_{m=l+1}^{M} \hat{R}_{lm}[\tau_{lm}(q(x,y,z))]
\]

where \( \hat{R}_{lm}[\tau_{lm}(q(x,y,z))] \) is the generalized cross-correlation function between the \( l \)th and \( m \)th microphone signal as

\[
\hat{R}_{lm}(\tau_n) = \frac{1}{N} \sum_{w=0}^{N-1} F_{l}^{*}(w) F_{m}(w) e^{i2\pi n \tau_n}
\]

where \( F_{m}(w) \) is the fast Fourier transform of \( X_{m}(n) \). \( \tau_{lm}(q(x,y,z)) \) is the TDOA between the \( l \)th and \( m \)th microphone signal. Let \( r_{l} \) and \( r_{m} \) represent the position of microphone, \( c \) is the sound propagation speed in air (usually 343 m/s), then \( \tau_{lm}(q(x,y,z)) \) is given by

\[
\tau_{lm}(q(x,y,z)) = \frac{||q - r_{m}|| - ||q - r_{l}||}{c}
\]

Therefore, the estimation of real sound source position can be presented as equation (13)

\[
\hat{q}_{s} = \arg \max_{q \in Q} \hat{P}(q)
\]

As is mentioned before, it is time-consuming if we search in the 3D space coordinates point by point. Then, we propose SSDE to deal with the problem by discretizing the space \( Q \) to discrete azimuth-elevation angle space \( Q^{d}(\alpha, \beta) \).

The sound source height is assumed to be \( h \), we can get \( \tau(q^{d}(\alpha, \beta)) \) by using \( \hat{q}^{d}(\alpha, \beta) \). Given \( r_{o} \) as the head chain origin position in world coordinates, the 3D coordinates \( q^{d}(\alpha, \beta) \) can be approximated as equation (14)

\[
q_{x} = r_{o}(x) + \frac{h \times \sin \alpha}{\tan \beta}
\]

\[
q_{y} = r_{o}(y) + \frac{h \times \cos \alpha}{\tan \beta}
\]

\[
q_{z} = r_{o}(z) + h
\]

Now we are able to write down the SSDE in discrete angle space as equation (15)

\[
\hat{q}_{s}(\alpha, \beta_{s}) = \arg \max_{q^{d} \in Q^{d}} \hat{P}(q^{d}(\alpha, \beta))
\]

\[
\hat{P}(q^{d}(\alpha, \beta)) = \sum_{l=1}^{M} \sum_{m=l+1}^{M} \hat{R}_{lm}[\tau_{lm}(q^{d}(\alpha, \beta))]
\]

\[
\tau_{lm}(q^{d}(\alpha, \beta)) = \frac{(||q^{d} - r_{m}|| - ||q^{d} - r_{l}||)}{c}
\]

It has to be noted that the height assumption may bring some estimation error but the searching method in discrete space reduces the computation load so we can realize the real-time processing in humanoid robot NAO. In “Multi-robot 2D SSL using distance weighting revision” and “Three-dimensional sound source localization” sections, we will revise the 3D position to improve the result precision. The SSDE algorithm is concluded as Algorithm 2.

Besides, the localization resolution in discrete angle space is limited by the humanoid robot’s microphone array and sampling frequency since we use TDOA as our localization cue. Suppose the maximum frequency as \( f_{\text{max}} \), and the distance between the \( l \)th and \( m \)th microphone is \( d_{lm} \). The maximum time delay which can be measured is \( d_{lm}/c \), so \( d_{lm} \times f_{\text{max}} \) kinds of angle can be distinguished under the maximum frequency. Given \( M \) microphones, the possible maximum resolution can be

\[
\Delta \theta = \frac{360^\circ}{\prod_{l=1}^{M} \frac{d_{lm}}{c} \times f_{\text{max}}}, \quad m = l + 1
\]

The distance between microphone arrays and the sampling frequency are the main factors that limit the SSDE resolution. As a result, the discretization interval should be set larger than \( \Delta \theta \) in angle space.

**Multi-robot 2D SSL using distance weighting revision**

SSDE provides the estimated sound source direction relative to single robot. SSRL introduces a multi-robot collaboration SSL algorithm using SSDE. As is shown in Figure 5, a 2D position can be estimated by crossing the azimuth angles and distance weighting revision. After
combining with the elevation angle, the height can be revised and we can get a more precise 3D position. The 2D SSL algorithm based on distance weighting revision will be presented in this section.

It needs to be stated that the robots rely on Wi-Fi communication. After each robot finish processing the sound signal, the angle information will be sent to the master robot who is responsible for executing 2D SSL and 3D SSL algorithm in real time.

To illustrate 2D SSL, let’s assume there are \( R \) humanoid robots NAO (\( R = 3 \) in this article) with initial pose \( L_r = [x_r, y_r, \theta_r] \), \( r = 1 \ldots R \). For each robot, the relative sound source direction \( q_r(\alpha_r, \beta_r) \) is calculated using SSDE, where \( \alpha_r \) is the azimuth and \( \beta_r \) is the elevation.

Intersect \( R \) robots’ azimuth angle rays, and we can get \( C_R^2 \) intersections \( P_{ij} \), \( i = 1 \ldots C_R \). After taking average, an estimated 2D position is achieved as equation (17)

\[
P_{\text{uncorrected}} = \frac{1}{C_R} \sum_i P_i
\]  

(17)

When the sound source distance is far greater than the microphone interval, we can analyze the sound propagation by assuming microphones are configured at the head chain origin point. For angle resolution sector region, the larger the radius, the longer the arc will be (\( L = r \times \alpha \)). All points in the arc will be regarded as the same angle in current angle resolution. That is to say, in the same angle resolution zone, the closer the microphone is to the sound source, the smaller the arc length \( L \) and the smaller the area it represents. Correspondingly, the small shifting of sound source \( \Delta q \) will be reflected in small deviation of angle \( \Delta \alpha \). Thus, closer distance to microphone means higher credibility of the identified sound source direction.

Using this criterion, 2D SSL revises the estimated 2D position by distance weighting revision. For each robot, the distance between it and uncorrected sound source position is calculated as \( d_r \), \( r = 1 \ldots R \). Then, we select the closest one to revise the position. The corrected position can be acquired by rotating the uncorrected pose to the azimuth ray of the closest one

\[
P_{2D} = P_{\text{corrected}} = L_r + d_r \cdot \begin{bmatrix} \cos(\alpha_r + \theta_r) \\ \sin(\alpha_r + \theta_r) \end{bmatrix}
\]

(18)

where \( r_i \) is the index of the robot with closest distance to sound source.

In 2D SSL, not only all robots’ sound direction information is combined into uncorrected position, but also the closest robot distance with corresponding angle is used to correct the position. The corrected 2D position combines the information of distance, and it is more reliable than the uncorrected one. 2D SSL in this section provides the 2D estimated position of sound source, and we will infer the 3D estimated position in the next section.
Three-dimensional sound source localization

This section aims at deducing the 3D estimated position of sound source. Once we get the 2D position from 2D SSL, we can combine it with the elevation and work out the height of sound source to revise the height assumed before. Figure 6 shows the 3D SSL coordinates.

The main procedure is described as follows:

Given two vectors with fixed starting points \( \vec{n} \) and \( \vec{b} \), where \( \vec{n} \) is the normal vector of horizontal plane with starting point as \( \vec{P}_{\text{corrected}} \) and direction as any normal vector. And \( \vec{b} \) is the elevation ray vector with starting point as \( \vec{L}_\alpha \) and direction as elevation direction \( \vec{\beta}_\alpha \).

Intersect \( \vec{n} \) and \( \vec{b} \), and we can get \( \vec{P}_{\text{cross}} \), which can be assumed as the 3D estimated sound source position \( \vec{P}_{3D} \).

The whole 3D SSL algorithm is concluded as shown in Figure 7. SSDE works out azimuth and elevation using acquired microphone signal firstly. Combining angle information and robots’ position, 2D position is deduced by 2D SSL. Next, the elevation angle is used to revise the height and get 3D position by 3D SSL.

The main procedure is described in Algorithm 3.

**Figure 6.** 3D SSL coordinates. (The dark green point is the ground truth 3D sound source location, and the dark red point is the estimated 3D position. The green point, red point, and blue point have the same meaning as described in Figure 4.) SSL: sound source localization.

**Figure 7.** 3D SSL algorithm structure. SSL: sound source localization.

**Algorithm 3.** Sound source localization.

### 2D Sound source localization

1: For robots located at \( L_r = [x_r, y_r, \theta_r], r = 1 ... R \), get their corresponding direction \( q_r(\alpha_r, \beta_r) \), \( r = 1 ... R \) using Sound Source Direction Estimator;

2: Intersect all robot-to-source direction line \( L_r q_r, r = 1 ... R \) and generate intersections \( P_i, i = 1 ... C_2 \);

3: Estimate uncorrected 2D sound source position as \( \vec{P}_{\text{uncorrected}} = \frac{1}{C_2} \sum_i P_i \);

4: Calculate distance \( d_i = ||L_r - \vec{P}_{\text{uncorrected}}|| \) and select \( r \) s.t. \( d_r = \min_i d_i \);

5: Get distance weighting revised 2D location as \( \vec{P}_{2D} = \vec{P}_{\text{corrected}} = L_r + d_r \begin{bmatrix} \cos(\alpha_r + \theta_r) \\ \sin(\alpha_r + \theta_r) \end{bmatrix} \);

### 3D Sound source localization

6: Intersect \( \vec{n} = (P_{\text{corrected}}, \vec{n}) \) and \( \vec{b} = (L_r, \vec{\beta}_r) \), thus get the crossed 3D sound source position as \( \vec{P}_{3D} \);
Experiments and analysis

We perform two experiments and one application to evaluate the proposed SSRL framework. In the first experiment, we compare the proposed sound recognition algorithm with another classical algorithm based on short-time Fourier transform (STFT).\(^3\)\(^3\)\(^4\) We perform a numerical analysis to show the accuracy and anti-interference ability of our system. Then, we test our framework in the indoor and outdoor environment to evaluate the performance of SSDE. We compare SSDE with GCC-PHAT-based algorithm and classical SRP-PHAT-based algorithm to illustrate the efficiency and robustness of our framework. Additionally, we apply the proposed system for one application in humanoid robots NAO and test it on RoboCup SPL Technical Challenge.

**Sound source recognition test**

We evaluate our sound recognition algorithm on humanoid robot NAO(V5) which is equipped with four microphones on its head. The configuration parameter of microphone arrays is presented in Table 1.

We put the robot in indoor environment and compare our algorithm with the STFT-based way. The whistle, speech, and clap sound are collected to test the whistle recognition accuracy and anti-interference ability. To test the anti-interference ability of our algorithm, we collect the same kind of whistle with noise. The noise is simulated by sound player which is located at 0.5 m away from the robot and test audio is set to be played at 1 m away from the robot. The level and frequency of audio is presented in Table 2. The frequency of whistle and clap sound is very close to each other.

The audio sampling frequency is set to be 48 kHz in this experiment. The sampled microphone signal including whistle, speech, and clap is normalized as shown in Figure 8. The signal segments are different in frequency as we can clearly see from the figure and the data are consistent with what we provided in Table 2. The frequency of whistle and clap sound is very close to each other.

The STFT-based method detects the principal frequency components using STFT. For instance, the STFT frequency spectrum of a section of whistle is shown in Figure 9. The range of whistle identification is 2.5–3.5 kHz, so the signal

---

**Table 1.** NAO(V5) robot body and head with microphone arrays configuration parameters.\(^*\)

| Part | Name   | X (m)  | Y (m)  | Z (m)  |
|------|--------|--------|--------|--------|
| A    | MicroFR| 0.0206 | -0.0309| 0.0986 |
| B    | MicroFL| 0.0206 | 0.0309 | 0.0986 |
| C    | MicroRL| -0.0215| 0.0558 | 0.0774 |
| D    | MicroRR| -0.0215| -0.0558| 0.0774 |

\(^*\)The position is relative to end transform of head chain.

**Table 2.** The level and frequency of audio.

| Audio type | Level (dB) | Frequency (kHz) |
|------------|------------|-----------------|
| Whistle    | 120        | 2.5–3.5         |
| Clap sound | 60         | 2–3             |
| Speech     | 40–60      | 0.3–3.4         |
| Noise      | 60         | 0.02–20         |

**Figure 8.** Three kinds of microphone signal segments wave including whistle, speech, and clap.
can be convinced to be whistle if the principal frequency is within the range.

The SSRL proposed in this article uses the cross-correlation function between current audio segment and reference whistle as the standard to identify the whistle. The ratio of cross-correlation function between two signals to reference signal auto-correlation function with a whistle blowing event is shown in Figure 10.

Two hundred samples are collected to compare the two algorithms. The identification results are shown in Table 3 and Figure 11.

The criterion used to judge the accuracy in this article is

\[ r\% = \frac{TP}{TP + FP} \]  \hspace{1cm} (19)  

And the criterion of anti-interference ability is

\[ e\% = \frac{TN}{FN + TN} \]  \hspace{1cm} (20)  

As we can see from the result, the two algorithms have different performance on the test set. The STFT-based recognition method gets an accuracy of 0.52 and anti-interference ability of 0.89 while SSRL proposed in this article can reach an accuracy of 0.75 and anti-interference ability of 0.98. Obviously, SSRL outperforms STFT in accuracy and anti-interference ability.

SSDE test

In this section, we will evaluate the efficiency and robustness of three source direction estimator including SSDE, GCC-PHAT, and SRP-PHAT in indoor and outdoor environments.
environment. The angle estimation accuracy and time cost are measured to distinguish which algorithm is better. To test the robustness of the algorithm, the corresponding audio is collected in the indoor and outdoor environment. The indoor environment is set in a totally enclosed room of length 13.0 m and width 8.0 m, while the outdoor environment is set in a very large stadium. There is reverberation caused by wall reflection in the indoor environment, while there are noisy voices in the outdoor environment.

Both SRP-PHAT and SSDE use the same kind of SRP feature. Classical SRP-PHAT algorithm estimates the source direction by searching in the 3D space using 3D coordinates while SSDE searches in angle space. The SRP-PHAT searching step is set to be 0.05 m in this experiment, and the searching domain is $9 \times 6 \text{ m}^2$ with height of 1.3–1.9 m.

| Ground truth (°) | SRP-PHAT (°) | SSDE (°) |
|------------------|--------------|----------|
| 14               | 13.565676    | 15.007578|
| 17               | 17.933968    | 17.300603|
| 21               | 20.612855    | 20.010115|
| 30               | 28.538149    | 30.015187|
| 50               | 49.919726    | 50.025328|
| 58               | 56.6964      | 60.030399|
| 14               | 13.565676    | 15.007578|

SRP-PHAT: steered response power-phase transform; SSDE: sound source direction estimator.

Table 5. Elevation angle test result in the indoor environment.

In indoor environment with reverberation, sound source in different azimuth-elevation position is acquired to test the angle estimation effect of three algorithms. Only some selected angle ranges are tested and results are shown in Tables 4 and 5. The error distribution and time cost are shown in Figure 12 and Table 6.

Similarly, we set up the same experiments in outdoor environments and test the three algorithms. The results are shown in Tables 7 to 9 and Figure 13.

The different error rates for different angles may be caused by the random noise, but it is clear to see from the results that angle estimation error of SSDE is less than GCC-PHAT and average time cost of SSDE is less than SRP-PHAT. In summary, GCC-PHAT-based way is faster than other two methods but performs not well in angle estimation in complicated environment. The SRP-PHAT-based way performs well in angle estimation but it has low efficiency and costs too much time in searching. For our proposed SSDE, it shows higher accuracy and efficiency and it is robust to both indoor and outdoor environment.

Application

We apply SSRL for 3D SSL on humanoid robots NAO. To test the recognition and localization simultaneously, we design a whistle recognition and localization experiments in robot soccer competition field.

![Figure 12. Angle estimation error distribution of three algorithms in indoor environments with reverberation. (The figure only shows the test angle data. In the error distribution map, the red line represents SSDE, green represents GCC-PHAT, and blue represents SRP-PHAT. And the scale of error coordinates is degree. What’s more, GCC-PHAT cannot estimate the elevation because of the arrival assumption.)](image)
Table 6. The average time cost of three algorithms when processing the same audio data segments indoors.

| Method      | Average time cost (ms) |
|-------------|------------------------|
| SSDE (ours) | 36.958                 |
| GCC-PHAT    | 32.585                 |
| SRP-PHAT    | 59.340                 |

SSDE: sound source direction estimator; GCC-PHAT: generalized cross-correlation phase transformation; SRP-PHAT: steered response power-phase transform.

Table 7. Azimuth angle test result in an outdoor environment.

| Ground truth (°) | GCC-PHAT (°) | SRP-PHAT (°) | SSDE (°) |
|------------------|--------------|--------------|----------|
| 0                | -1.429738    | -10.3108     | 2.292994 |
| 45               | 90           | 41.2240      | 50.0253  |
| 90               | 180          | 78.7283      | 85.043   |
| 135              | 136.72488    | 121.6938     | 135.068  |
| 180              | 155.896228   | 173.5138     | 165.083  |
| -135             | -173.0507    | -134.977     | -135.06848|
| -90              | -17.809810   | -89.3009     | -90.0456 |
| -45              | -9.138095    | -39.19307    | -40.0203 |

GCC-PHAT: generalized cross-correlation phase transformation; SRP-PHAT: steered response power-phase transform; SSDE: sound source direction estimator.

Table 8. Elevation angle test result in an outdoor environment.

| Ground truth (°) | SRP-PHAT (°) | SSDE (°) |
|------------------|--------------|----------|
| 21               | 21.011967    | 22.303140|
| 30               | 28.538149    | 30.015187|
| 50               | 48.103226    | 50.025328|

SRP-PHAT: steered response power-phase transform; SSDE: sound source direction estimator.

Table 9. The average time cost of three algorithms when processing the same audio data segments outdoors.

| Method      | Average time cost (ms) |
|-------------|------------------------|
| SSDE (ours) | 38.7301                 |
| GCC-PHAT    | 30.2985                 |
| SRP-PHAT    | 53.2898                 |

SSDE: sound source direction estimator; GCC-PHAT: generalized cross-correlation phase transformation; SRP-PHAT: steered response power-phase transform.

The experiment environment configuration is shown in Figure 14, robots are put in the predefined initial position and stay still. Robot 1 is placed at (2.5 m, 1.7 m) facing at −62.1°, robot 2 is placed at (−1.5 m, −1.3 m) facing at 40.9°, and robot 3 is placed at (−3.3 m, 0.6 m) facing at 0°. Once the referee blows the whistle in any place, the robot will detect the signal and recognize the whistle. Meanwhile, the audio data segments from all microphones will be used to locate the whistle. The robots will contact with each other via Wi-Fi and share their SSDE result. Finally, the robot will calculate the 3D sound source position using 3D SSL, and the console will present the result.

Three indexes are used to measure the localization result in this article: The first one is the 3D absolute position error, the second one is the relative distance error to the robot with closest distance to whistle, and the third one is the relative azimuth rotation error.

Suppose the actual position of whistle is \( q_{GT} \), and the SSRL estimated position is \( q_S \), then three indexes will be defined as equation (21)

\[
\begin{align*}
    e_1 &= ||q_{GT} - q_S|| \\
    e_2 &= ||q_{GT} - L_{min}|| - ||q_S - L_{min}|| \\
    e_3 &= |\alpha_{GT} - \alpha_S|
\end{align*}
\]

where \( L_{min} \) is the position of robot with closest distance to whistle position, \( \alpha_{GT} \) is the angle to whistle position in closest robot’s coordinates, and \( \alpha_S \) is the angle to estimated position in closest robot’s coordinates. All errors are normalized to their absolute value.

We take robot soccer competition field as the test field and place the robot at set place in Figure 15. The whistle is blown by a standing adult at every 0.5 m × 0.5 m² area. The yielding error distributed throughout the whole field is shown in Figures 15 to 17. In field with size of 9 × 6 m², the error is very small in most areas but there still exists some points with large error. The possible reason may be the misidentification caused by too close or too far distance and the accumulated angle estimation error of robots.

The average error of three indexes is shown in Table 10.

In addition, we applied the SSRL in humanoid robots NAO and took part in RoboCup2019 Standard Platform Directional Whistle Technical Challenge. Real competition field is set at a big exhibition hall and the whole area (18 m × 30 m²) contains four SPL standard field. There are walking spectators and noises in the hall so it brings a lot of challenges to our system. Some of the localization data are shown in Table 11.

As we can see from the result, using humanoid robots microphone arrays, the SSRL framework proposed in this article can locate the sound source in both indoor and outdoor environments. The 3D distance absolute error is approximately 0.8636 m, the relative distance error is 0.6760 m, and the relative azimuth angle error is 62.1°. In directional whistle technical challenge, the minimum distance error can reach to 0.049 m and the minimum angle error can achieve 1.352°.

The SSRL algorithm proposed in this article is proved to be able to locate the sound source in indoor and outdoor environments on humanoid robots NAO with limited computation resources and critical real-time requirement after application testing. Compared with other methods, the SSRL achieves higher efficiency and robustness.
Figure 13. Angle estimation error distribution of three algorithms in outdoor environments. (In error distribution map, the representation is as Figure 12.)

Figure 14. SSRL test environment. The robot is standing on the robot soccer standard platform competition field and the whistle will be blown by the referee. SSRL: sound source recognition and localization.

Figure 15. Error distribution map of $e_1$

Figure 16. Error distribution map of $e_2$
Conclusions

This article presents an SSRL framework approach implemented for 3D SSRL on humanoid robots. The humanoid robots microphone arrays are fixed and configured like human-ears, and limited computation resources make humanoid robots SSRL a challenging problem. This article addresses the problem by following ways.

The sound recognition based on cross-correlation feature classifies the sound type using sampled microphone signals. And the sound source direction is determined by a single robot SSDE which employs a new discrete angle space SRP-PHAT function. Aiming at getting the 3D position of the sound source, a 2D multi-robots SSL is firstly presented with distance weighting revision. The following 3D multi-robots SSL calculation works out the certain 3D sound source position.

Compared with other methods, SSRL is proved to be more efficient and robust after experiments test. However, extensive research is still necessary to further improve the system accuracy and robustness. Deep neural networks-based methods in the literature inspire us to extract deep feature of audio data and use it as a new cue to locate the sound source. We believe that more progress on SSRL will be made to transcend the field and improve robot audition in general.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported in part by the National Natural Science Foundation of China (grant nos 61673300, U1713211 and 61733013).

References

1. Magassouba A, Bertin N, and Chaumette F. Aural servo: sensor-based control from robot audition. IEEE Trans Robot 2018; 34(3): 572–585.
2. Basiri M, Schill F, Lima PU, et al. Localization of emergency acoustic sources by micro aerial vehicles. J Field Robot 2018; 35(2): 187–201.
3. Su D, Vidal-Calleja T, and Miro JV. Asynchronous microphone arrays calibration and sound source tracking. Auton Robots 2020; 44(2): 183–204.
4. Cowling M and Sitte R. Comparison of techniques for environmental sound recognition. Pattern Recog Lett 2003; 24(15): 2895–2907.
5. Rascon C and Meza I. Localization of sound sources in robotics: a review. Robot Auton Syst 2017; 96: 184–210.
6. Argentieri S, Danes P, and Souères P. A survey on sound source localization in robotics: from binaural to array processing methods. Comput Speech Lang 2015; 34(1): 87–112.
7. Kang HG, Graczyk M, and Skoglund J. On pre-filtering strategies for the GCC-PHAT algorithm. In: 2016 IEEE international workshop on acoustic signal enhancement (IWAENC), Xi’an, China, 13–16 September 2016, pp. 1–5. IEEE.
8. Velasco J, Taghizadeh MJ, Asaei A, et al. Novel GCC-PHAT model in diffuse sound field for microphone array pairwise distance based calibration. In: 2015 IEEE international conference on acoustics, speech and signal processing (ICASSP), Brisbane, QLD, Australia, 19–24 April 2015, pp. 2669–2673. IEEE.
9. Van Den Broeck B, Bertrand A, Karsmakers P, et al. Time-domain generalized cross correlation phase transform sound source localization for small microphone arrays. In: 2012 5th European DSP education and research conference (EDERC), Amsterdam, Netherlands, 13–14 September 2012, pp. 76–80. IEEE.
10. Do H, Silverman HF, and Yu Y. A real-time SRP-PHAT source location implementation using stochastic region contraction (SRC) on a large-aperture microphone array. In:
21. Zhang X and Wang D. Deep learning based binaural speech separation in reverbrant environments. IEEE/ACM Trans Audio Speech Lang Process 2017; 25(5): 1075–1084.

22. Yalta N, Nakadai K, and Ogata T. Sound source localization using deep learning models. J Robot Mechatron 2017; 29(1): 37–48.

23. Sun Y, Chen J, Yuen C, et al. Indoor sound source localization with probabilistic neural network. IEEE Trans Ind Electron 2017; 65(8): 6403–6413.

24. Brindza J, Thomas A, Lee S, et al. Active sound localization in a symmetric environment. Int J Adv Rob Sys 2013; 10(7): 301.

25. Adavanne S, Pertilä P, and Virtanen T. Sound event detection using spatial features and convolutional recurrent neural network. In: 2017 IEEE international conference on acoustics, speech and signal processing (ICASSP), New Orleans, LA, 5–9 March 2017, pp. 771–775. IEEE.

26. Lei Z, Yang K, and Ma Y. Passive localization in the deep ocean based on cross-correlation function matching. J Acoust Soc Am 2016; 139(6): EL196–EL201.

27. Freeman SE, Buckingham MJ, Freeman LA, et al. Cross-correlation, triangulation, and curved-wavefront focusing of coral reef sound using a bi-linear hydrophone array. J Acoust Soc Am 2015; 137(1): 30–41.

28. Qu X, Xie L, and Tan W. Iterative constrained weighted least squares source localization using TDOA and FDOA measurements. IEEE Trans Signal Process 2017; 65(15): 3990–4003.

29. Li X, Shen M, Wang W, et al. Real-time sound source localization for a mobile robot based on the guided spectral-temporal position method. Int J Adv Robot Syst 2012; 9(3): 78.

30. Imran M, Hussain A, Qazi NM, et al. A methodology for sound source localization and tracking: development of 3D microphone array for near-field and far-field applications. In: 2016 13th International Bhurban conference on applied sciences and technology (IBCAST), Islamabad, Pakistan, 12–16 January 2016, pp. 586–591. IEEE.

31. Velasco J, Martín-Arguedas CJ, Macias-Guarasa J, et al. Proposal and validation of an analytical generative model of SRP-PHAT power maps in reverberant scenarios. Signal Proc 2016; 119: 209–228.

32. Hao B, Zhao Y, Li Z, et al. A sensor selection method for TDOA and AOA localization in the presence of sensor errors. In: 2017 IEEE/CIC international conference on communications in China (ICCC), Qingdao Shi, Shandong Sheng, China, 22–24 October 2017, pp. 1–6. IEEE.

33. Parchami M, Zhu WP, Champagne B, et al. Recent developments in speech enhancement in the short-time Fourier transform domain. IEEE Circ Soc Am Mag 2016; 16(3): 45–77.

34. Nakagawa Y and Wada S. Diagnosis system for abnormal respiratory sound using divided pulmonary sound waveform. J Signal Proc 2016; 20(4): 221–224.