EURASIP Journal on Audio, Speech, and Music Processing

RESEARCH Open Access

Multi-source localization by using offset residual weight

Maoshen Jia*, Shang Gao and Changchun Bao

Abstract

Multiple sound source localization is a hot issue of concern in recent years. The Single Source Zone (SSZ) based localization methods achieve good performance due to the detection and utilization of the Time-Frequency (T-F) zone where only one source is dominant. However, some T-F points consisting of components from multiple sources are also included in the detected SSZ sometimes. Once a T-F point in SSZ is contributed by multiple components, this point is defined as an outlier. The existence of outliers within the detected SSZ is usually an unavoidable problem for SSZ-based methods. To solve this problem, a multi-source localization by using offset residual weight is proposed in this paper. In this method, an assumption is developed: the direction estimated by all the T-F points within the detected SSZ has a difference along with the actual direction of sources. But this difference is much smaller than the difference between the directions estimated by the outliers along with the actual source localization. After verifying this assumption experimentally, Point Offset Residual Weight (PORW) and Source Offset Residual Weight (SORW) are proposed to reduce the influence of outliers on the localization results. Then, a composite weight is formed by combining PORW and SORW, which can effectively distinguish the outliers and desired points. After that, the outliers are removed by composite weight. Finally, a statistical histogram of DOA estimation with outliers removed is used for multi-source localization. The objective evaluation of the proposed method is conducted in various simulated environments. The results show that the proposed method achieves a better performance compared with the reference methods in sources localization.

Keywords: Multiple sound sources localization, Direction of arrival estimation, Reverberation, Soundfield microphone

1 Introduction

Multiple sound source localization is a hot subject in audio signal processing and has gained extensive attention over decades for its vital role in various audio applications. An accurate estimation of sound source location can be applied in robotics [1], sound source separation [2, 3], hearing aids [4], human-machine interaction [5], and so on. The main task of multiple sound source localization is to obtain the position of sources in the acoustic scene by using observed signals from several sensors without knowing the information of sound sources and the generation process of the received signals. Generally speaking, the existing localization methods can be roughly divided into four categories. The first ones are based on the time difference of arrival (TDOA) and its extensions [6–8]. Various TDOA-based methods, including CC (cross-correlation), GCC (generalized cross-correlation), and MCCC (multichannel cross-correlation), were developed to solve the localization problem. However, these methods still suffer from the accuracy decline caused by ambient noise and reverberation. The robustness of the TDOA-based methods can be improved from three aspects [9]: incorporate priori knowledge to ameliorate the performance, increase the number of sensors to take advantage of the redundancy, and take the reverberation into account in signal model to improve the accuracy of TDOA. While limited by the practical application scenarios, priori knowledge is not always available [10] and the number of microphones set up in the space is usually restricted. Lots of the methods

* Correspondence: jiamaoshen@bjut.edu.cn
School of Information and Communication Engineering, Faculty of Information Technology, Beijing University of Technology, Beijing 100124, China

introduce reverberation in modeling while their performance still degrades inevitably when reverberation time increases.

The spectral estimation techniques are also used in multi-source localization, some of the representative methods include multiple signal classification (MUSIC) [11–13] and estimation of signal parameters via rotational invariance (ESPRIT) [14] algorithms. MUSIC is one of the famous subspace based methods for multi-source direction of arrival (DOA) estimation under super-deterministic conditions (i.e., the number of microphones is greater than the number of sound sources) [15, 16]. The MUSIC algorithm exhibits high resolution and is applicable for an array of arbitrary geometric shapes. A direct path dominance (DPD) test [17] based method is adopted to perform DOA estimation by cross-spectrum matrices. In this method, a focusing process is added to localize multiple sources using array with arbitrary configuration. Nevertheless, since the entire array manifold needs to be searched to find the steering vectors which are orthogonal to the noise subspace, most of spectral estimation techniques has poor computational efficiency [18]. Besides, the MUSIC-based methods need to know the number of sources in advance. Due to the sparseness of speech signals, the number of simultaneously active sources is not consistent in broadband, which make them not suitable to deal with broadband sources.

The third one is based on the independent component analysis (ICA), which is usually used to deal with the linear instantaneous mixtures. The convolutive blind source separation problem can be solved by transforming the problem into the frequency domain [19–22]. Recently, ICA-based methods have also been adopted to conduct multi-source localization [23–25]. As mentioned in [26], the ICA-based methods can be used in the Time-Frequency (T-F) domain to carry out the multi-source localization task so long as the number of dominant sources does not exceed the number of microphones in each T-F zone, which means that the requirement for the number of microphones becomes more relaxed.

For the last category, sparse components analysis (SCA) is applied to locate multiple sound sources. Most of the SCA-based methods rely on the w-disjoint orthogonal assumption [27]. In this assumption, there are usually some T-F zones, where only one source is active or dominant, exist, even if multiple sources sound simultaneously. These specific T-F zones are called single source zone (SSZ). According to [26], it has been proved that using T-F points within the detected SSZs to conduct multi-source localization can obtain high accuracy. Moreover, [28] proposed a DOA estimation method using a sound field microphone to realize highly accurate positioning. The SCA-based methods include not only the detection of "zone" level sparse components but also "point" level sparse components. These "point" level sparse components are the T-F points, which are called the single source point (SSP), where only one source is active or dominant. The methods that use SSP to perform multiple source localization are called the SSP-based methods [29–33]. The SSP detection methods include but are not limited to the coherent test based method [29], the energy threshold based method [31], and the energy decomposition based method in [32]. Recently, methods base on the phase feature of real and imaginary parts of mixture TF vectors have also been proposed. The method proposed in [33] falls into the last category which detects the low-reverberant single-source (LRSS) points to perform multi-source localization and achieves localization results with high accuracy. Different from these methods which focus on the extraction of sparse components from recorded signals, other methods aim to find the model that fitting the observed distribution [34]. All the methods mentioned above can achieve a good localization performance in simple acoustic scenarios (i.e., an acoustic scenario with low reverberation time and a low number of sound sources). However, as the number of sources or the reverberation time increases, the localization performance of both SSZ-based methods and SSP-based methods declines. For SSP-based methods, there could be many outliers, which contain the wrong localization information, mixed in with the detected results. For example, the outliers can be T-F points consist of multiple components with the same phase which cannot be discriminated by the phase-based SSP detection method. And the outliers can also be composed of a single reflection component that cannot be identified by both energy-based and phase-based criteria. As for the SSZ-based methods, in addition to the same problem as SSP-based methods, there are always some sources with relatively less DOA estimates in the histogram which are hard to be detected through peak searching. This kind of sources is called Statistically Weak Sources (SWS) and the other sources are called statistically dominant sources (SDS) [35]. A Statistically dominant source component removal (SDSCR) algorithm is proposed in the same paper to solve this problem. It has been proved that the SDSC R method can always obtain better localization results. However, the increasing number of outliers caused by multiple sources and high reverberation time is still unsolved and could lead to a significant decrease in localization accuracy.

In this paper, the problem of SSZ-based methods mentioned above has been proved by verification experiments. This problem can be summarized as: there is an inevitable presence of outliers in detected SSZs which leads to the decreasing accuracy of localization. To solve to this problem, a new assumption has been proposed in this paper: since the distribution of outlier within SSZ is sparse and the DOA estimates of outliers are randomly distributed, the average direction calculated by the T-F points inside SSZ does not far away from the location of sound source. Referring to this assumption, a residual based weighting idea has been proposed and applied in the SSZ-based method to realize a robust multi-source localization method for the reverberation environment. In the proposed method, two residual measurements are introduced to weaken the adverse effect brought by outliers. Among them, the residual measurement used in the outlier detection part aims to evaluate the angular difference, which is also called the offset, between the direction estimated by each T-F point and the average direction estimated by their corresponding SSZ. This measurement is named as "point offset residual measurement." The second proposed residual measurement, which is called "source offset residual measurement," is mainly used to measure the offset between the DOA estimated by T-F point and the coarse localization of their corresponding sound sources. Each of the residual measurement corresponds to a weight that is applied in the localization process. According to the residual measurement, they derivate from, they are named as "Point Offset Residual Weight" (PORW) and "Source Offset Residual Weight" (SORW), respectively. Both two weights have the characteristic that they give low values to outliers while high values for desired T-F points. Based on this characteristic, these false detections of outliers would be weighted and suppressed. Different from the traditional methods which give both outliers and desired T-F points a weight of 1 in the statistical histogram of DOA estimations, the proposed method combines two weights mentioned above to produce the histogram refer to the contribution of each T-F point in the direction of their corresponding actual source. It should be mentioned that the proposed method can be applied to a variety of array setups. The PORW can be applied to all the SSZ-based methods. The SORW can be applied to the methods where histograms are plotted to perform DOA estimation. At last, the advantages of the proposed method over traditional methods and its robustness in various environments are verified by several sets of objective and subjective experiments.

2 Modeling and angular calculation

In this section, the basic model of signals received by sound field microphone is introduced. Then, the SSZ detection criterion for T-F zones and B-format transformation, angular calculation for T-F points is reviewed.

2.1 The SSZ detection using sound field microphone

In this paper, the sound field microphone [28], which is the array of directional microphones, is used to record the sound signals. A sound field microphone consists of four closely placed cardioid microphone capsules. Since the distance between source and microphone is much larger than that between different microphone capsules, these microphone capsules can be regarded spatially coincident with respect to a sound source. Under this condition, the recorded signals from different channels also have no delay in time domain. The recorded signals of four channels are represented as $\{s_1, s_2, s_3, s_4\}$ which are corresponding to the microphone capsules pointing at front left up (FLU), front right down (FRD), back left down (BLD), and back right up (BRU), respectively. Assume an acoustic environment contains Q sources with reverberation and noise, the signals recorded by sound field microphone after short time Fourier transform (STFT) are modeled as the formula below:

$$s_p(n,k) = \sum\nolimits_{i=1}^{Q} h_{i,p}(k)\cdot x_i(n,k) + r_{\mathrm{ev}}(n,k) + v_{\mathrm{n}}(n,k) \quad (1)$$

where $p \in \{1, 2, 3, 4\}$ is the index of the microphone capsule. The signals from ith source in the frame n and frequency number k are represented as $x_i(n,k)$. $h_{i,\,p}(k)$ is the transfer function between ith source to the pth microphone capsule. The signals received by pth microphone capsule from sound field microphone are represented as $s_p(n,k)$. Since the reverberation components consist of signals from different sound sources in various of T-F points, this part of received signals is simplified as $r_{\mathrm{ev}}(n,k)$. The noise components are represented as $v_{\mathrm{n}}(n,k)$. It should be mentioned that because of the spatially coincident characteristic of the different microphone capsules, the recorded signals from four capsules of the sound field microphone have close-to-equal phase while different amplitude. Following equation should also be satisfied:

$$\frac{h_{i,p}(k)}{|h_{i,p}(k)|} = \frac{h_{i,q}(k)}{|h_{i,q}(k)|} \quad (2)$$

Meanwhile, if T-F point (n,k) only consists of the direct component from a single source, the following equation should also be satisfied:

$$\frac{s_p(n,k)}{|s_p(n,k)|} = \frac{s_q(n,k)}{|s_q(n,k)|} \quad (3)$$

Where $p,\ q \in \{1, 2, 3, 4\}$ $(p \neq q)$ represent the index of the sound field microphone channel. Since the signals from different channels of sound field microphone should have the same phase character, which means the signal waveform within SSZ in different channels should have high similarity. A SSZ detection criterion [26] is proposed based on this characteristic, which uses the normalized cross-correlation (NCC) coefficient between signal between channels to detect SSZ. The NCC coefficient between channel p and q in the T-F zone Z is defined as follow:

$$r_{pq}(Z) = \frac{R_{pq}(Z)}{\sqrt{R_{pp}(Z)\cdot R_{qq}(Z)}} \tag{4}$$

where $p, q \in \{1, 2, 3, 4\}$ $(p \neq q)$ denotes the index of the channel and Z denotes the T-F zone whose size is customized by user. $R_{pq}(Z)$ is cross-correlation coefficient given as follow:

$$R_{pq}(Z) = \sum\nolimits_{(n,k)\in Z} \left| s_p(n,k)\cdot s_q(n,k) \right| \tag{5}$$

When there are no reverberation components or noise components in the T-F zone used for analysis, all the T-F points within the T-F zone consist of the signals from the same source through the direct path, the result of (4) should satisfy $r_{pq}(Z) = 1$, while it is hard to realize in the actual experimental environment so the criterion is relaxed as below:

$$r_{pq}(Z) > 1 - \varepsilon \tag{6}$$

where ε is an empirical threshold set by user according to the practical scenario. This threshold should guarantee that enough T-F zones are detected to perform the localization; meanwhile, most of the T-F zones contaminated by reverberation and interfere sources should be removed.

2.2 Angular calculation for T-F points

The signals received by the sound field microphone directly are called the A-format signals. In previous work [35], a simple and intuitive angular calculation method has been proposed where A-format signals need to be changed into B-format signals. The B-format signals consist of four channels which are represented as $\{s_w, s_x, s_y, s_z\}$, and the transformation operation can be expressed as the formula below:

$$\begin{cases} s_w(n,k) = s_1(n,k) + s_2(n,k) + s_3(n,k) + s_4(n,k) \\ s_x(n,k) = s_1(n,k) + s_2(n,k) - s_3(n,k) - s_4(n,k) \\ s_y(n,k) = s_1(n,k) - s_2(n,k) + s_3(n,k) - s_4(n,k) \\ s_z(n,k) = s_1(n,k) - s_2(n,k) - s_3(n,k) + s_4(n,k) \end{cases} \tag{7}$$

where s_w is the signal received by the omnidirectional channel, and $\{s_x, s_y, s_z\}$ are the signal received by three channels correspond to the Cartesian coordinate.

As mentioned in [28], if a T-F point only consists of the direct component from a single source i and no reverberation components are involved, the model of received B-format signals from sound source i can be represented as:

$$\begin{cases} s_w(n,k) = \frac{\sqrt{2}}{2} x_i(n,k) \\ s_x(n,k) = cos\mu_i \cdot\ cos\gamma_i \cdot x_i(n,k) \\ s_y(n,k) = sin\mu_i \cdot\ cos\gamma_i \cdot x_i(n,k) \\ s_z(n,k) = sin\gamma_i \cdot x_i(n,k) \end{cases} \tag{8}$$

Where $x_i(n,k)$ represents the signals from source i in the frame n and frequency number k, μ_i, and γ_i are the azimuth and elevation of ith sound source, respectively. And thus, the localization of sound source i can be calculated. For simplicity, the formula below only calculates the azimuth of source i:

$$\hat{\mu}(n,k) = \tan^{-1}\left(\frac{Re\{s_w{}^*(n,k)\cdot s_y(n,k)\}}{Re\{s_w{}^*(n,k)\cdot s_x(n,k)\}} \right) \tag{9}$$

where $Re\{\cdot\}$ represents the operation of taking the real part, and $*$ denotes conjugation. Since the azimuth calculation process depends on the ratio of B-format signals from x, y channels, which means that for the T-F points with a similar direction of the vector $[s_x, s_y]$, the DOA estimations should have similar results. This particular trait is applied afterward to calculate the weight.

It should be noted that both (8) and (9) assume only one source is active in the T-F point (n, k) for simplicity. However, there are many T-F points contain reverberation components and\or multi-source components. That means there are more parameters and the components from multiple source in (8) which cannot be eliminated by the division in (9), which causes the decline of the localization accuracy.

3 Problem statement and the proposed point offset residual weight

As mentioned in the last section, the degradation of localization performance is mainly caused by the reverberation components and\or multi-source components. Once the estimated direction $\hat{\mu}(n,k)$ for T-F point (n, k) has a large difference to the directions of actual sources, this specific T-F point is defined as an outlier. In this section, the problem caused by the outlier is verified by experiments. After analyzing the reason causes this problem, a relaxed assumption is proposed to solve the problem. In this assumption, the direction estimated by all the T-F points in a detected SSZ is assumed to be much closer to the actual sources' direction than that for the outliers. Refer to this assumption, the point offset residual measurement is introduced to measure the angular difference between the direction estimated by all the T-F points within the detected SSZ and the direction estimated by each of them. Then, the point offset residual weight is calculated for the actual application.

3.1 Problem statement

In general, the SSZ-based methods use the relevance feature of the T-F zone to detect SSZ. Then, the T-F points within detected SSZ are used to perform multi-source localization. The SSZ-based methods could usually obtain a good performance. However, with the increase of reverberation time and\or the increase of the source number, the localization accuracy of SSZ-based methods is declining. This is mainly caused by the increasing ratio of outliers mixed in the detected SSZ. As this ratio increases, the wrongly counted number of sources could lead to the significant decrease in localization accuracy. To confirm the presence of outliers in the detected SSZ, a group of experiments is conducted.

In these experiments, the statistical analysis of DOA estimates for each T-F point within detected SSZ is performed using 100 sound source segments with a length of 8s from the NTT [36] database. The angle between sources is settled as 60°. The experiments are divided into two cases: changing the number of sound sources and reverberation time. As for the experiments of changing the number of sources, the source number is selected as $\{2, 3, 4, 5\}$ in an anechoic room. For the experiments of the changing of reverberation time, two sources are active simultaneously in the rooms with a reverberation time of 0 ms, 150 ms, 300 ms, and 450 ms, respectively. In these experiments, the SSZ is detected under four different thresholds of NCC, which means that the parameter from formula 6 mentioned above is chosen from {0.1, 0.25, 0.4, 0.55}. Among all the T-F points in the selected SSZs, the outliers are defined as the T-F points whose DOA estimations' deviation from the direction of the actual source by more than 20°. On the contrary, the rest of the T-F points within the detected SSZ are defined as the desired T-F points. The threshold of 20° is an empirical threshold chosen according to the informal experiments. The outlier ratio η is calculated as follow:

$$\eta = \frac{\sum_{\tau=1}^{\kappa} N_o(Z_\tau)}{\sum_{\tau=1}^{\kappa} N_a(Z_\tau)} \tag{10}$$

where κ is the number of the detected SSZ, Z_τ is the τth detected SSZ. N_o is the outliers' number in Z_τ and N_a is the size of Z_τ. The ratio of outliers in the different condition is given in Fig. 1a and b.

From Fig. 1, a visible trend can be found that the proportion of outlier is increasing with the number of source and the reverberation time. Besides, the increase in the outlier ratio caused by reverberation is much greater than that caused by the increase in source number. More specifically, we take the blue labeled (i.e., $\varepsilon = 0.1$) data in Fig. 1a as an example, when the room is anechoic with two sources, only 14.3% of outliers are mixed in the detected SSZ. When the number of sources rises to five, the ratio of the outlier is 28.6%, which means that despite the increase of the outlier ratio, most of the detected T-F zones are still dominant by only one source. However, the situation is quite different in the presence of reverberation. We still take the blue labeled

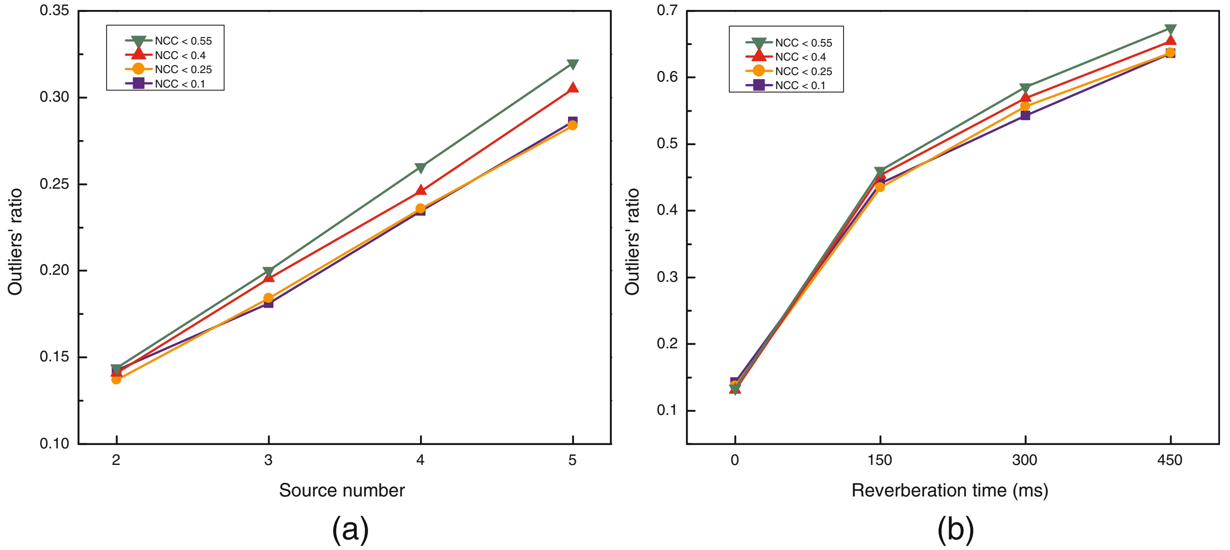

Fig. 1 Statistical analysis for outliers' ratio in different conditions. **a** Different source number in anechoic room (**b**) two sources with different reverberation time

data for analysis, in Fig. 1b, compared with the outlier ratio of 14.3% in the anechoic environment, the outlier ratio increases to 44.1% with 150 ms of reverberation. When the reverberation is 450 ms, the outlier is an account for 63.6% of the total T-F points within detected T-F zone which means the components carry the directional information of actual sources are overwhelmed by other components with a high possibility.

On the other hand, it can be found that the change of threshold can hardly decrease the ratio of outliers significantly. From both Fig. 1a and b, the red and green labeled data (i.e., $\varepsilon = 0.4$, $\varepsilon = 0.55$, respectively) always obtained a higher ratio of outliers than blue and orange labeled data (i.e., $\varepsilon = 0.1$, $\varepsilon = 0.25$, respectively), which means that the increase of the threshold could increase the ratio of outliers in the detect T-F zone. However, even lower the thresholds means fewer outliers are included in the detected T-F zone, the strict threshold also makes fewer T-F zones can be detected, which could lead to insufficient data for analysis. Above all, the change of the SSZ detection threshold could not reduce the number of outliers effectively, and the outliers are always existing in the detected SSZ.

3.2 The analysis for the cause of these outliers

From the description above, the outliers always exist in the detected SSZs, and the reasons for the presence of these outliers are analyzed as follows:

- The SSZ-based methods apply the zone-level characteristics to discriminate T-F zones dominated by only one sound source, and this leads to negligence in the performance of the individual T-F points. Even though the outliers may carry multiple sound components, these redundant components could hardly pull down the NCC of the whole T-F zone and that specific T-F zone can still be recognized as SSZ.
- In order to guarantee that enough SSZs are detected for the localization, the SSZ detection threshold ε should be relaxed. However, the relaxed threshold could also mean a large ratio of outliers mixed in with the detection result.

In conclusion, the existence of the outlier in the detected T-F zone is an unavoidable problem for SSZ-based methods and this problem gets worse as the reverberation time and source number increase.

3.3 The proposed point offset residual measurement

From the former section, we found that there are always outliers exist in the detected SSZs and the change of the SSZ detection threshold can hardly change the proportion of outliers. To solve this problem, a new assumption is proposed in this section: the DOA estimations using outliers are far away from the direction of the actual source. While the DOA estimations using all the T-F points within detected SSZ are much closer to the direction of the actual source. A set of experiments are conducted to verify this assumption. In this experiment, the root mean squared error (RMSE) is used to measure the error between estimate direction and true direction, which is defined as:

$$RMSE = \sqrt{\frac{1}{m}\sum\nolimits_{i=1}^{m}\left(\hat{\mu}(i) - \mu_j\right)^2} \tag{11}$$

where m represents the number of estimated directions. $\hat{\mu}(i)$ is the ith estimated direction and the direction of its corresponding source is μ_j, $j \in \{1, 2, \ldots, Q\}$, Q is an integer and presents the number of actual sources.

Three sets of directional estimation are used to calculate RMSE; the first set consists of the directions estimated by taking the average of all the T-F points within each SSZ while the second set is the average of directions estimated by outliers. The third one is the directions estimated by each outlier. The same experimental data and simulation environments as the experiments performed above are used. The results are shown in Fig. 2a and b.

In Fig. 2, the lines with blue and square marks are the RMSE of directions estimated by each outlier, and the line with green circle representing the RMSE corresponding to the average direction of the outliers within the detected T-F zone. The line with a label of orange and triangle is the RMSE of the direction estimated by all the T-F points from each SSZ. In both figures, the direction estimated by each SSZ is the closest to the actual direction. While the direction estimated by each outlier obtain the biggest RMSE. The same result can be obtained in a variety of environments which proves the proposed assumption.

Besides, the reason causes this phenomenon could also be found in this figure. The reason for the low RMSE of SSZ is not only because of the introduction of desired points. From Fig. 2, it can be found that in the same situation, the green line is lower than the blue line which proves that taking average of outliers also contributes to the decrease of RMSE.

Refer to the assumption mentioned above, a conclusion can be obtained: once the offset between the direction estimated by a T-F point and the direction of its corresponding SSZ is large; this specific T-F point has little contribution to the localization. Based on this conclusion, a measurement called point offset residual measurement is introduced to evaluate the distance

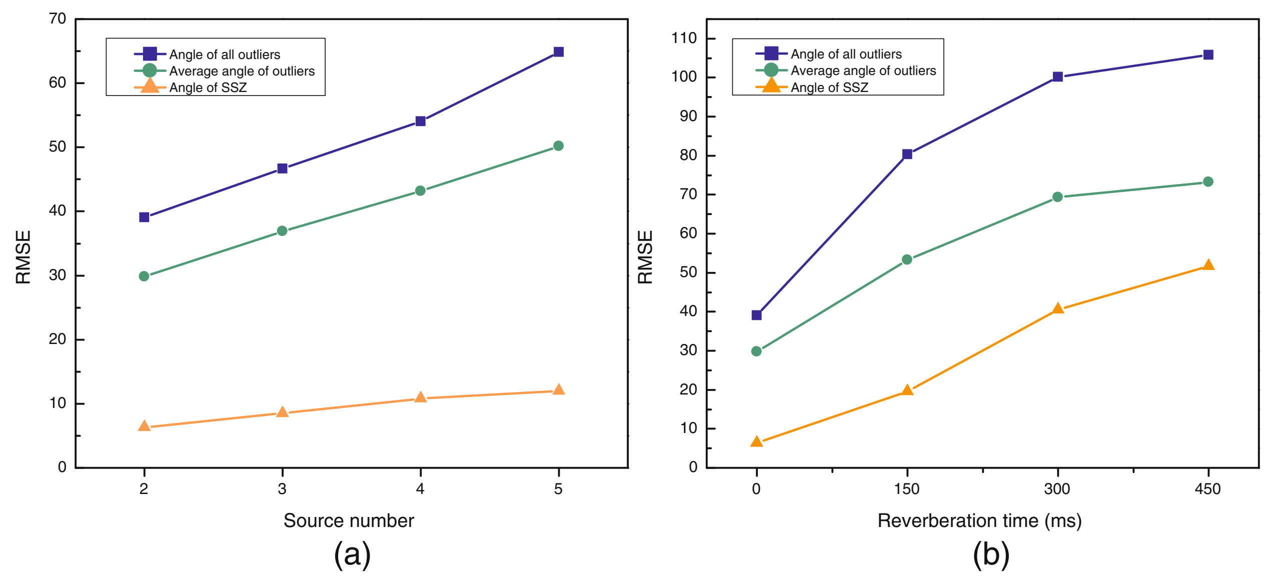

Fig. 2 Different methods to obtain azimuth and their corresponding RMSE. **a** Different source number in anechoic room (**b**) two sources with different reverberation time

between the direction estimated by SSZ and every T-F point. The point offset residual measurement is defined as:

$$M_p(n,k) = \frac{\left\langle \overline{\boldsymbol{V}}(\mathrm{Z})\cdot \boldsymbol{V}(n,k)\right\rangle}{\left\|\overline{\boldsymbol{V}}(\mathrm{Z})\right\|\cdot\left\|\boldsymbol{V}(n,k)\right\|} \quad (12)$$

where $\langle\cdot\rangle$ is the symbol of the inner product operator, and $\|\cdot\|$ denotes Euclidean norm. $\boldsymbol{V}(n,k)$ is a vector formed by B-format signal in (n,k), $\boldsymbol{V}(n,k) = [s_y(n,k), s_x(n,k)]'$, $[\cdot]'$ represents taking the transposition of the vector. $\overline{\boldsymbol{V}}(\mathrm{Z})$ is a vector formed by B-format signals in Z, which is given as:

$$\overline{\boldsymbol{V}}(\mathrm{Z}) = \sum\nolimits_{(n,k)\in Z}\frac{\boldsymbol{V}(n,k)}{K\cdot\left\|\boldsymbol{V}(n,k)\right\|} \quad (13)$$

where K represents the number of T-F points in the detected SSZ. According to the description below (9), the vector consists of B-format signals is corresponding to the estimated direction. That is why these vectors are introduced to measure the angular difference between SSZ and the T-F points.

The reason that the vector projection is used as the measurement is described as follows: In complex acoustic environments, most of the T-F points contain multiple components which can be seen as the superposition of plane waves. From the perspective of spatial geometry, if the microphone is recognized as a point in the space, then each of the plane waves corresponds to a vector. Thus, the signals received in a T-F point can be recognized as the composite vector, and formula 12 can be regarded as the projection of the vector. More specifically, the vector of each T-F points is projected onto the direction estimated by their corresponding SSZ to measure the contribution of each point in the direction of their corresponding SSZ. On the other hand, outliers could also contain the contribution of components from the source directly which is overwhelmed by reverberation components, and the proposed PORW is used to measure this contribution.

3.4 The proposed point offset residual weight

This proposed measurement calculates the orthogonal projection of each T-F point signal vector on SSZ average signal vector which aims to describe the offset between the direction of SSZ and the direction estimated by each T-F point. However, this measurement only describes the offset between the direction of SSZ and the directions of the points within it, which can be used to measure the contribution but not represent it. Therefore, a normalization step is necessary to change the angular difference into weight, which could represent the contribution of a T-F point and be directly assigned to a point. The *point offset residual weight* (PORW) is defined as:

$$W_p(n,k) = 1-\frac{\cos^{-1}\left(M_p(n,k)\right)}{\pi} \quad (14)$$

It should be noted that the higher PORW is given to the T-F points who have a higher contribution to the direction estimated by SSZ.

Since the direction estimated by SSZ is much closer to the direction of the actual source than the direction estimated by outliers. The proposed PORW should have a characteristic that a higher value for desired points, while a lower value for outliers. A set of experiments are conducted to verify this characteristic. The experimental data and simulation environment are the same as the experiment mentioned above. In this experiment, the average weights of four kinds of T-F points are calculated. These four categories are named as: Desired points in the desired SSZ (DPDS), desired points in falsely detected SSZ (DPFS), outliers in falsely detected SSZ (OFDS), and outliers in desired SSZ (ODS). The specific name of points and SSZs in the categories is explained as follows, the desired point is the same as the desired T-F point mentioned above. The desired SSZ represents a detected SSZ whose direction error is less than 20°, while the falsely detected SSZ does not meet this condition. More specifically, we take DPDS as an example: the DPDS collects all the desired T-F points in the SSZ with a direction error less than 20° and then, the average of their given PORW is calculated. The results are given in Fig. 3.

In Fig. 3, it can be found that desired points obtain higher weights than the outliers in all the situations. As the reverberation time increase, the difference of PORW for outliers and desired points is decreased. More specifically, in the environment with a reverberation time of 150 ms, the average PORWs for DPDS and DPFS are 0.96 and 0.87, respectively. While for the ODS and OFDS, the average PORWs in the same situation are just 0.81 and 0.80, respectively. As the reverberation time increase, it can be found that the PORW of DPDS holds steady and the PORW of DPFS has declined while the average PORW of the desired TF point is still higher than that of outliers, which shows the robustness of the proposed PORW.

It should be mentioned that the DPFS represents the desired T-F point in falsely detected SSZ. The reason DPFS gets a higher weight than outlier is that the falsely detected SSZ contains outliers with different directions which could also far away from their corresponding SSZ's direction. On the other perspective, the SSZ's direction has a feature that it does not far away from the actual source's azimuth. Even the falsely detected SSZ has a direction whose offset larger than the threshold, it still relatively closes to the true direction which brings a relatively high weight to the desired T-F points.

In summary, the proposed PORW can be used to attenuate the undesirable effects of outliers in SSZ-based methods. However, even the outliers' effect can be weakened; the outliers themselves are hard to be removed by using a single weight. To remove the outliers, the second kind of weight is proposed in the following process and more details are described in the next section.

4 Proposed method

Based on the proposed PORW, a multi-source localization method is proposed and the flowchart is presented in Fig. 4. The whole method is described below:

Firstly, the SSZ detection is operated for the received A-format signals. Secondly, all T-F points in detected SSZ are converted into B-format to form the signal vector and calculate PORW for every T-F points. Thirdly,

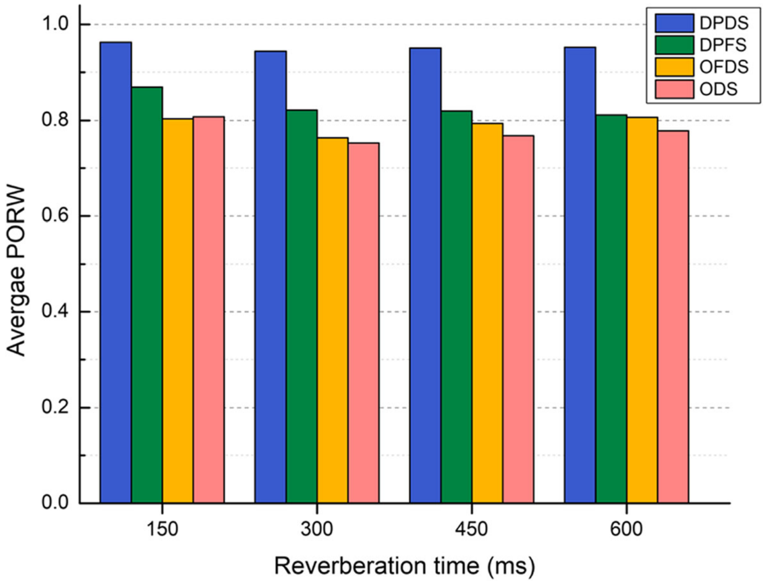

Fig. 3 PORW of four kinds of points with different reverberation time

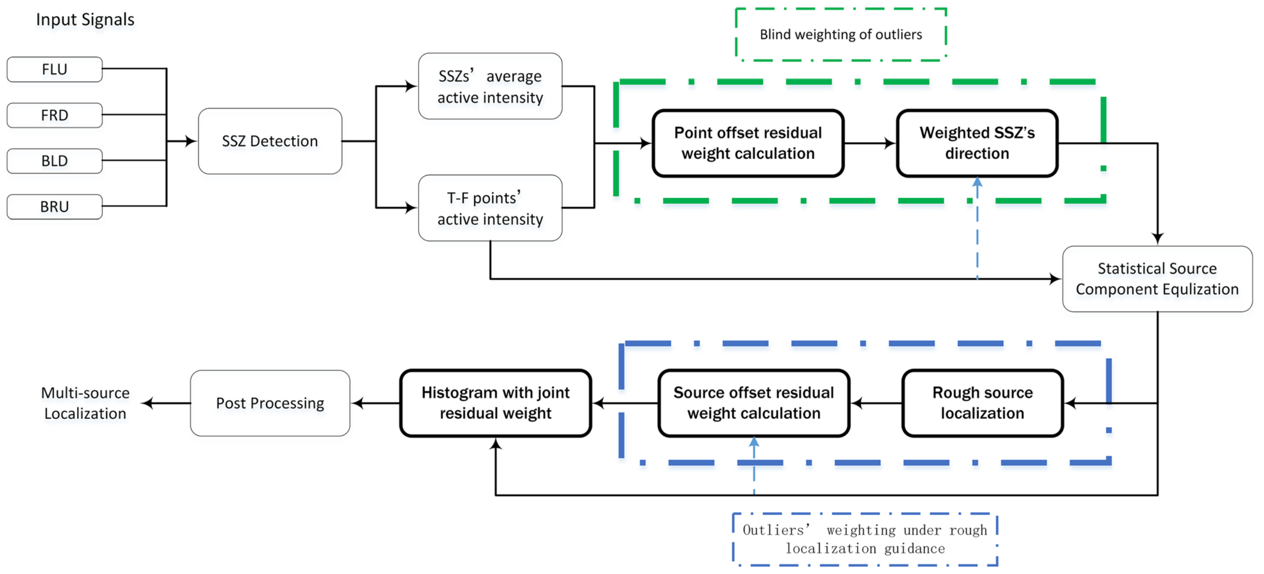
Input Signals
FLU
FRD
BLD
BRU
SSZ Detection
SSZs' average active intensity
T-F points' active intensity
Blind weighting of outliers
Point offset residual weight calculation
Weighted SSZ's direction
Statistical Source Component Equlization
Multi-source Localization
Post Processing
Histogram with joint residual weight
Source offset residual weight calculation
Rough source localization
Outliers' weighting under rough localization guidance

Fig. 4 The block diagram for the proposed multi-source localization framework

the PORWs are used to weight the B-format signal vectors to weaken the effect of outliers. A more accurate DOA estimation of SSZ after weighing can be obtained. Then, based on these DOA estimations, the statistical source component equalization (SSCE) [37] is conducted to obtain the rough localization of sources. Later, the SORW is proposed for further reducing the effect of outliers. Follow by that, the SORW and PORW are combined as a composite weight which is used to remove the outliers and draw the histogram of DOA estimation. Finally, post-processing applied in [35] including KDE and peak searching is conducted to complete the multi-source localization.

Since the directional analysis of B-format signals can be performed relatively simply based on an energetic analysis of the sound field, the activity intensity [38] which shows the direction of the flow of energy is introduced to replace the vector formed by B-format signals. The corresponding formula is shown as follow:

$$\begin{cases} I_x(n,k) = \dfrac{\sqrt{2}}{\rho c}\left[\, Re\left\{ s_w^*(n,k)\cdot s_x(n,k)\right\}\right] \\ I_y(n,k) = \dfrac{\sqrt{2}}{\rho c}\left[\, Re\left\{ s_w^*(n,k)\cdot s_y(n,k)\right\}\right] \\ I_z(n,k) = \dfrac{\sqrt{2}}{\rho c}\left[\, Re\left\{ s_w^*(n,k)\cdot s_z(n,k)\right\}\right] \end{cases} \tag{15}$$

where ρ and c are represented as the density of the medium and the speed of sound respectively and they are both constant. The representation of active intensity vector in the horizontal plane is $\boldsymbol{I}(n,k) = [I_x(n,k), I_y(n,k)]$. More details of the proposed method are described below:

4.1 The blind weighting of T-F points

Based on the description of the active intensity mentioned above, the proposed assumption can be changed to:

- There are always outliers in the detected SSZ whose directions far away from the actual source.
- The SSZ's direction is close to the direction calculated by the active intensity vector which contains the actual sources' azimuth information while the directions calculated by using the outliers' active intensity vectors have a big difference with the actual source's direction.

Therefore, the direction estimated by SSZ and the activity intensity vector of all the T-F points within SSZ is used to obtain the PORW. Then, the obtained PORWs are combined with the active intensity vector to weaken the impact of outliers. A more accurate direction estimated by taking the average of the weighted active intensity vector in each SSZ can be obtained by the following formula:

$$\widehat{\mu}(Z) = \tan^{-1}\left(\frac{Re\left\{ \sum_{(n,k)\in Z} W_p(n,k)\cdot I_y(n,k)\right\}}{Re\left\{ \sum_{(n,k)\in Z} W_p(n,k)\cdot I_x(n,k)\right\}} \right) \tag{16}$$

The estimated direction of the SSZ can be used as guidance for statistical source component equalization.

4.2 Statistical sound source component equalization

The statistical sound source component equalization [37] aims to deal with the masking phenomenon caused by the excessive disparity in the number of T-F points each sound source belongs to. The sound sources are divided into two categories, the statistically weak sources (SWS) and the statistically dominant sources (SDS). The former is likely to be masked by the latter in the DOA statistical histogram because of their few corresponding T-F points. Four thresholds are set to distinguish the T-F points belong to the SDS and remove a part of them to make the SWS sufficiently obvious to be retrieved by post-processing. Although this step solves the problem of the masked SWS, the effect of outliers on the localization is instead amplified after this step because the removal of components from sound sources is equal to the enhancement of the outliers' components.

4.3 The weighting of outliers under rough localization guidance

To solve the problem caused by the removal of SDS components, the source offset residual measurement is proposed. The peaks formed by outliers, which are called the pseudo-peaks, have some characteristics different from the peaks formed by actual sources' components, which are called the true peaks. These characteristics can be summarized in two aspects: Firstly, the pseudo-peaks usually lower than the true peaks. Secondly, the pseudo-peaks locate differently in the histogram before and after smoothing due to the randomly distributed outliers. While the true peaks always have the same location in the histogram before and after smoothing. Based on these characteristics, the SORW is introduced to deal with the problem caused by pseudo-peaks, the formula is shown below:

$$M_s(n,k) = \min\left\{\frac{\langle \tilde{\boldsymbol{I}}_i \cdot \boldsymbol{I}(n,k)\rangle}{\|\tilde{\boldsymbol{I}}_i\| \cdot \|\boldsymbol{I}(n,k)\|}\right\} \tag{17}$$

where $i \in [1, \tilde{Q}]$ is the index of the estimated source, $\tilde{Q}$ is the estimated number of sources (i.e., the pseudo-peaks are included). $\tilde{\boldsymbol{I}}_i$ is the active intensity vector of the T-F point that carries the azimuth information closest to the ith estimated azimuth.

Like the PORW, a normalization step is also necessary to the source offset residual measurement. Therefore, the *source offset residual weight* (SORW) is proposed to lower the pseudo-peaks in the statistical histogram which is defined as:

$$W_s(n,k) = 1 - \frac{\cos^{-1}(M_s(n,k))}{\pi} \tag{18}$$

Since the pseudo-peaks in smoothed histogram hardly correspond to local maximums formed by outliers in the DOA statistical histogram, the outliers who form the local maximum in DOA statistical histogram can hardly be given the highest SORW. While the T-F points with the accurate azimuth information could obtain a higher SORW due to the peaks formed by them have the same localization before and after the smoothing. Like the previous description, an experiment is conducted to verify the proposed SORW in different reverberation time using the same group of data, the average of SORW calculated by outliers and desired points is shown in Fig. 5:

It can be found that the SORWs which are given to the desired points are higher than that given to the outliers in different reverberation time. With these experimental results, SORW can be proven effective to distinguish the outliers and desired T-F points. Although the SORW can be easily integrated into other localization frameworks, it should be noted that the SORW is based on the rough localization using PORW. That means the calculation of PORW is a necessary step for the obtaining of SORW. A detailed explanation would be given in the next section combining with the experiments.

4.4 The composite weight based post-processing

In this section, the proposed PORW and SORW are combined into one composite weight, the formula is shown below:

$$W_c(n,k) = W_p(n,k) \cdot W_s(n,k) \tag{19}$$

The experiment is also conducted to evaluate the composite weight and Fig. 6 shows the results:

Experimental results show that both outliers and desired points have lower composite weights compared with the results of SORW but the outliers still get weights much lower than desired points. According to the analysis above, both PORW and SORW can distinguish the desired points and outliers by giving a lower weight to outliers which means that the composite weight should have a better performance. To verify this conclusion, the difference between the weights given to outliers and desired points are calculated by using PORW, SORW, and composite weight. The results are shown in Fig. 7.

It can be found that the composite weight has the highest difference between outliers and desired points in all the simulation environments and the conclusion mentioned above is proved. Due to its high efficiency to discriminate outliers, the composite weight is used to remove the outliers and draw the histogram for the final localization. The desired points should satisfy the equation given as follow:

$$W_c(n,k) \geq 1 - \delta \tag{20}$$

where δ is an empirical threshold set by uses according to the application scenario and the T-F points that do

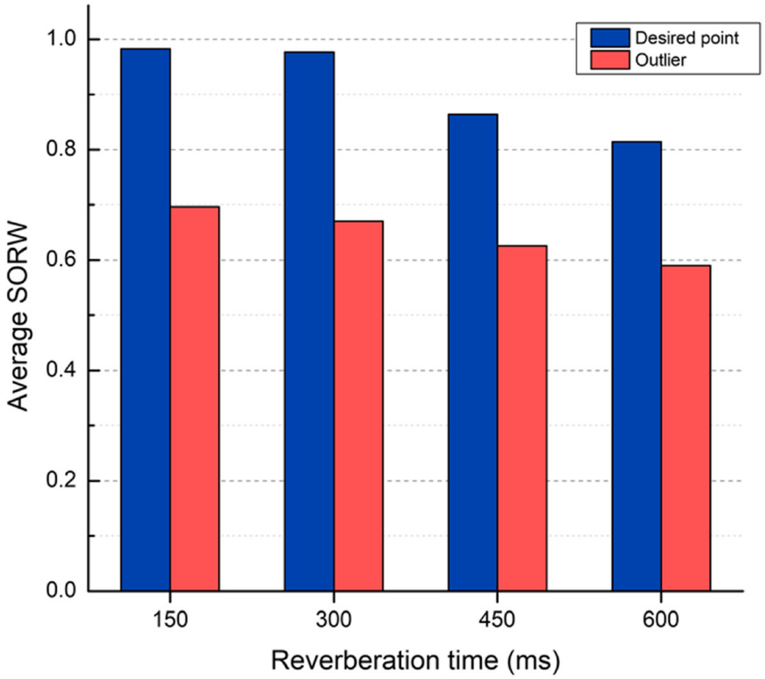

Fig. 5 SORW of desired points and outliers with different reverberation time

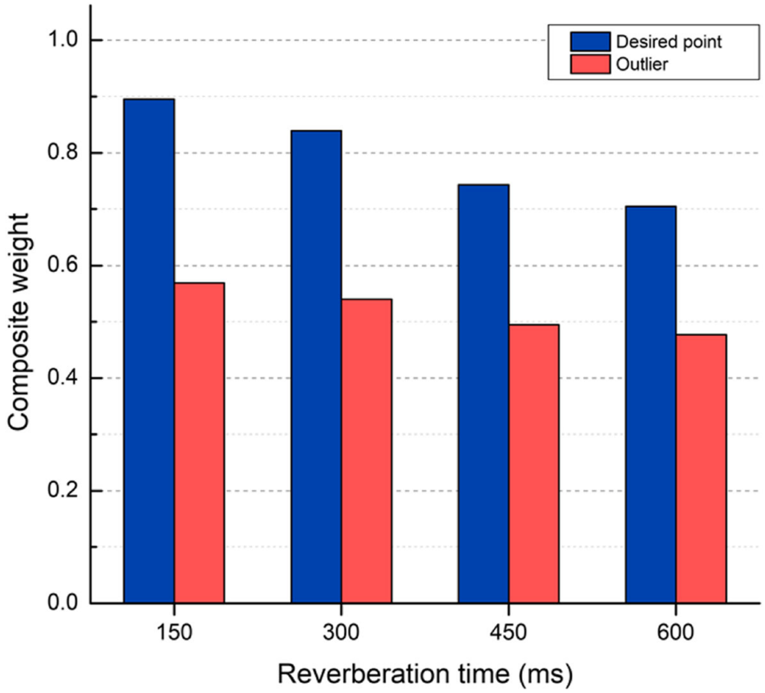

Fig. 6 Composite weight of desired points and outliers with different reverberation time

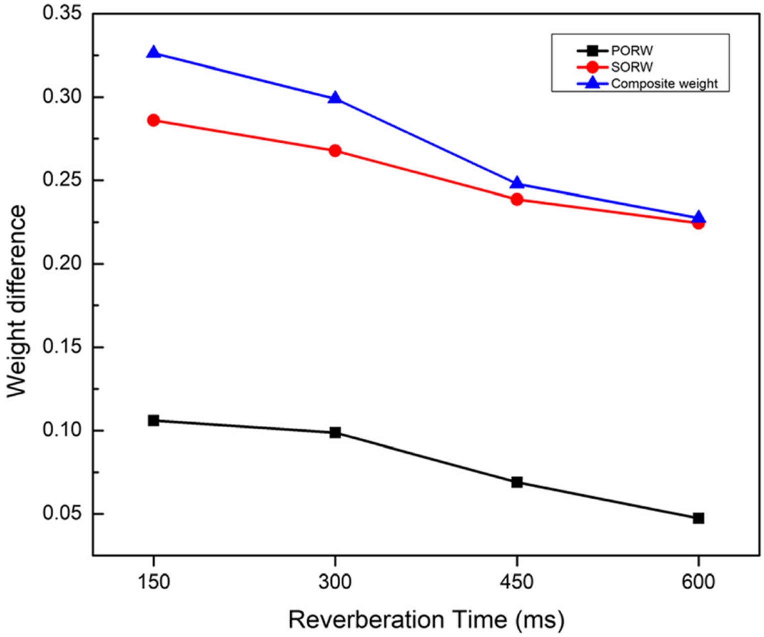

Fig. 7 The difference between weight given to the outliers and desired points

not satisfy this equation are recognized as outliers and their weights are assigned a value of 0. The histogram is drawn using the following formula:

$$Y(\mu) = \left\{ [\hat{\mu}(n,k)] = \mu | \sum W_c(n,k) \right\} \tag{21}$$

where $\mu \in [1{,}360]$, $(n,k) \in \mathbb{Z}$ is the angle in the histogram and $\mathbb{Z}$ represents the set of all the detected SSZ after equalization. The composite weight can be integrated into the histograms of SSZ-based methods. After that, the post-processing applied in [35] is used to perform the localization by KDE and peak searching.

5 Results and discussion

In this section, experiments are conducted to evaluate the performance of the proposed method. The experimental validation includes two aspects: the comparison between the localization results using different proposed weights and the comparison between different methods. The experiments are conducted in both simulation and actual environments, the experimental environments and reference methods is introduced as follows:

5.1 Experimental environments and reference methods

The simulation experiments were realized by using ROOMSIM package [39]. The simulation room used for experiments is uniformly set as a cube with a length of 6 m, a width of 3 m, and a height of 2 m. The sound field microphone is set in the center of the simulated room and the sources are set around the microphone with a distance of 1 m. It is noted that the elevation angle and the azimuth have the same calculation process except for the different channels of B format signals are selected. For the convenience of explanation and calculation, the elevation is set as 0° and not involved in the localization. It also means that the height of the sources is the same as that of the microphone. The original speech signals are chosen from the Chinese subdatabase from the NTT database [36]. The other parameters include but are not limited to the thresholds for STFT or SDSCR algorithm is shown in Table 1.

It should be noted that ε_p is the threshold for the peak searching on post-processing. The different values for the minimum difference threshold ε_a are set based on the prediction of the minimum possible distance between sound sources, in the experiments of changing separation of sources, the minimum difference threshold is changing with the separation of sources. As for the

Table 1 Experimental parameters

Parameter	Value
Sampling frequency of speech source	44.1kHz
Overlapping in frequency	50%
T-F zone width	32
SSZ detection threshold	0.25
STFT length	2048
Minimum difference threshold (ε_a)	50\40\30\20
Peak searching threshold (ε_p)	0.001
Outliers' removal threshold (δ)	0.6

outliers' removal threshold (δ), which is introduced in this paper, is set according to the results of statistical experiments. Combing with the results in Fig. 6, it can be found that the threshold of 0.6 can distinguish the outliers and the desired points in most of the experiment settings.

As for the experiments using actual recorded signals, the signals are recorded in an acoustic chamber with a dimension of 4.5 m × 3.5 m × 2.8 m. The reverberation time and background noise are estimated to be 400 ms and 28 dB, respectively. A Sennheiser Ambeo VR Mic was selected to record the signals. During the recording process, the microphone was set at the center of the room with a height of 1.8 m. Two male speakers were located 1.6 m away from the microphone.

The reference methods are selected from the representative methods of SSZ-based method [28], SDSCR algorithm [35], DPD [17] test based method and SSP-based method [33].

5.2 The statistical analysis for outliers' ratio with different source number

Before calculating the accuracy of the localization under different experimental environments, a group of experiments is performed to verify the effectiveness of the proposed method in terms of outlier removal. In these experiments, the reverberation time is set as 150 ms and the separation between sources is set as 60°, and the minimum difference threshold is set as 50°. The number of sources is chosen from $\{2, 3, 4, 5\}$. The results are shown in Fig. 8:

It should be noted that the way to select outliers and calculate outlier's ratio are the same as experiment shown by Fig. 1, which means that the DOA estimation is conducted for the points selected by the methods, once the DOA of a selected point has a deviation from the direction of the actual source by more than 20°, this point is recognized as an outlier. The outlier's ratio is obtained by calculating the proportion of outlier within selected points. From Fig. 8, it can be found that as the number of source increases, the outlier's ratio of all the methods increases. Among the points selected by all the methods, the points selected by proposed methods have the lowest outliers' ratio, which proves the effectiveness of the proposed method in terms of outlier removal.

5.3 The evaluation of the proposed method in different reverberation time

In this section, the proposed method is evaluated in the environment with different reverberation times. Four kinds of room with different reverberation times are set by the adjustment of the absorption of the walls inside the room and the experiments are conducted in these room with the reverberation time of $\{150\ \text{ms}, 300\ \text{ms}, 450\ \text{ms}, 600\ \text{ms}\}$. Three sound sources with a separation of 60° are active simultaneously and the RMSE of the estimated localization of sources are calculated. The minimum difference threshold is set as 50° in this group of experiments. The results are shown in Fig. 9:

In addition to the obvious trends of increasing RMSE with increasing reverberation time, the proposed method using composite weight can usually obtain the lowest

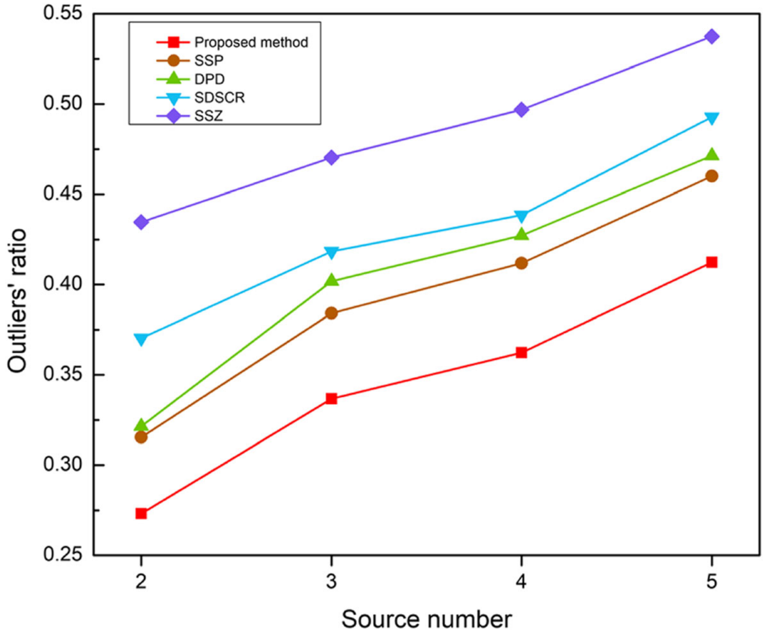

Fig. 8 Statistical analysis for outliers' ratio with different number of sources, RT60 = 150 ms, separation between adjacent sources is 60°

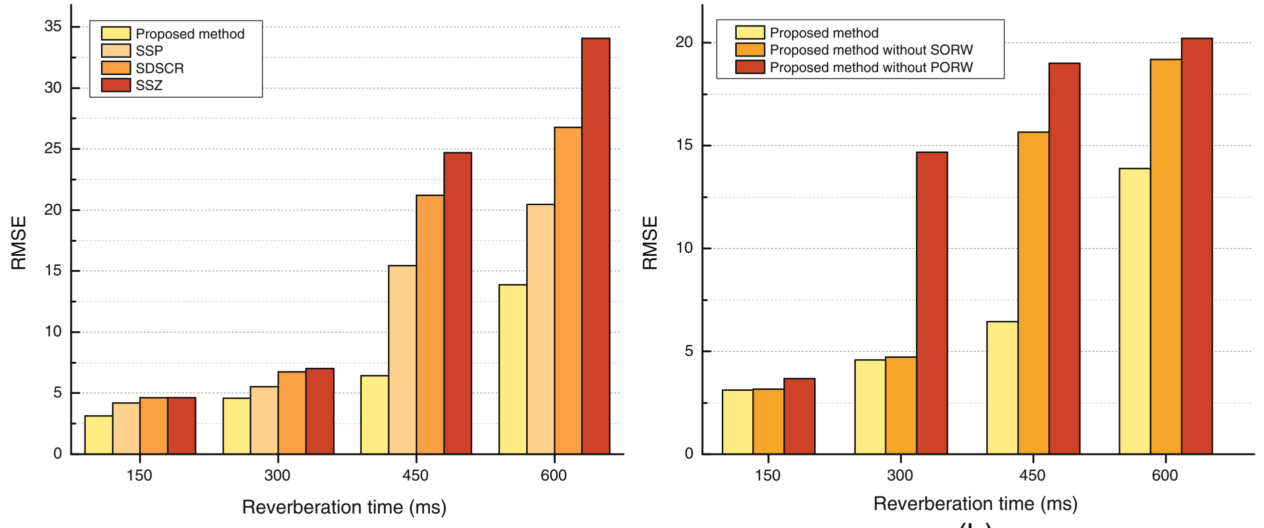

Fig. 9 RMSE of localization results in different reverberation conditions, three sources, separation between adjacent sources is 60°. **a** The RMSE of localization results using different methods. **b** The RMSE of localization results using different proposed weights

RMSE. As Fig. 9a shows, when the reverberation time is set as 150 ms, the difference of RMSE between the proposed method and SSP is only 1.1 while this difference widens to 6.58 when the reverberation time increases to 600 ms which demonstrates the advantage of the proposed method against reverberation environments. Comparing to that, the difference in localization accuracy between the proposed method and SSZ is even larger, especially in the environment with high reverberation time. In the room with a reverberation time of 600 ms, the RMSE of SSZ is up to 34.1, which is 20.2 higher than the RMSE of the proposed method. This severe decline in the localization accuracy is mainly due to the incorrect counting of source number. It should be noted that the data used to evaluate the localization performance is selected only if at least one method has its sound source counting correct. More specifically, the correct counting of sources' number means that the number of detected sources is equal to the number of actual sources. But in the complex acoustic environments, this criterion is too hard to realize for some methods, so this criterion has been relaxed. As long as all of the actual sources in the scenario have a corresponding localization estimation, and the error is not higher than the minimum difference threshold set above, the source counting is considered as a correct source counting result.

Since the traditional SSZ-based methods have poor robustness to the increase of reverberation time compared with SSP-based methods and the proposed method. In the scenario with severe reverberation interference, the SSZ-based method usually has the source counting problem more often than the proposed method which brings a huge difference in the localization accuracy.

Figure 9b aims to indicate that each step of the proposed method is necessary. It can be found that in this figure, the data with red mark always has a higher RMSE than the data with orange mark which means that the removal of PORW from the proposed method leads to a greater negative influence than removing SORW from the proposed method. As previously mentioned, the SORW is used to remove the pseudo-peaks consist of outliers based on the guidance of PORW. Once the PORW is not involved in the rough localization, the difference between pseudo-peaks and peaks consist of desired T-F points is shortened and it is possible that the peaks are wrongly removed. While the PORW is used in the blind weighing process which means that it can be used without the SORW. When the reverberation time is set as 150 ms and 300 ms, it only has little difference between the data marked yellow and the data marked orange because of the rare occurrence of pseudo-peaks in these scenarios. However, the absence of PORW does cause the localization accuracy decline in high reverberation scenarios. In the scenario with a reverberation time of 450 ms, the removal of SORW in the proposed method leads to the RMSE increase by 9.22, which means that the false detection of pseudo-peaks can cause a sharp decline of localization performance. It should be noted that when the reverberation time is 600 ms, this difference caused by removing SORW is 5.32, which is lower than that in the reverberation time of 450 ms. This phenomenon is caused by the increase in the error of the estimated directions when the reverberation time is 600 ms. The offset increase in the results with wrongly counted source number is not significant compare to the offset increase in the results with the source number counting correctly. In conclusion, the SORW is based on the PORW which means that without PORW, the proposed method is hard to be established. Even the PORW can exist only by itself, the problem of pseudo-peaks still leads to a dramatic increase of error in high reverberation situations. Both SORW and PORW play necessary roles in the proposed method.

5.4 The evaluation of proposed method with different separation of sources

In order to validate the proposed method under multiple conditions, experiments are conducted in the scenario of changing sources' separation. The reverberation time is set as 150 ms and the number of sources is three. The separations between sources are selected as {30°, 40°, 50°, 60°}. It should be noted that in this section, the minimum difference threshold used in SDSCR and post-processing changes with the separation of sources to simulate the changing of the pre-estimated minimum angle spacing between sources. The minimum difference threshold is set 10° lower than the separation between sources which means that the minimum difference threshold is select from {20°, 30°, 40°, 50°} corresponding to the separation between sources. The results are shown in Fig. 10:

From both figures in Fig. 10, the proposed method still has the lowest RMSE in all the scenarios. Apart from that, another trend in Fig. 10 is that in complex acoustic environments, the difference of RMSE for the proposed method and the reference methods is lower than that in Fig. 9. For example, in the room with a reverberation time of 600 ms, the difference of RMSE for the proposed method and the SSP is 6.58 while in the scenario that the separation between sources is 30°, the difference between the proposed method and the SSP is just 0.25. It should be noted that the reduction of separation between sources does not boost the reverberation components as the operation of increasing reverberation time does. The problem caused by the reduction of separation mainly lies in the difficultly in the peak searching part where the distance between adjacent peaks is too close and hard to be detected. That means a reasonable pre-estimated minimum angle spacing can help with this problem. Since peak searching is a necessary part of the post-processing, both the proposed method and reference method could benefit from the suitable minimum difference threshold setting. That lowers the RMSE difference between the proposed method and reference methods. Even so, the proposed method still has the lowest RMSE compared with both reference methods and the proposed method with one proposed weight removed, which demonstrates the validity of the proposed method in a variety of situations.

5.5 The evaluation of proposed method using noisy signals and actual recorded signals

The proposed method is further verified by using noisy signals and signals recorded in real environments.

For the experiments using noisy signals, the reverberation time is set as 450 ms. Three sources with a separation of 90° sound simultaneously in the simulated room and the minimum difference threshold is set as 70°. According to the setup of experiments, white Gaussian

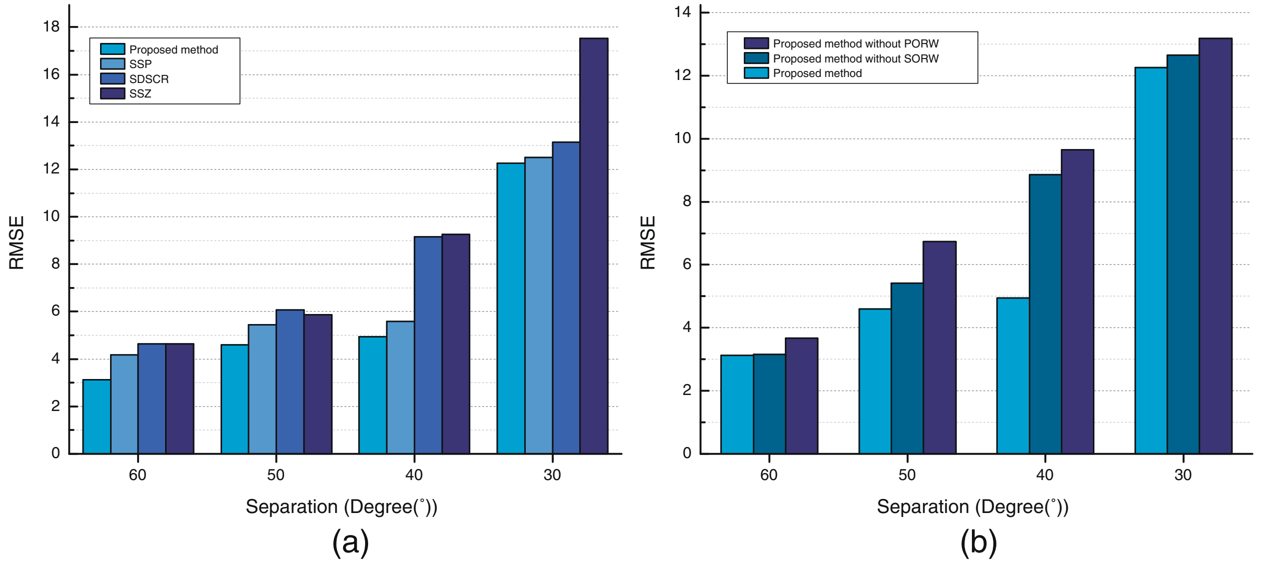

Fig. 10 RMSE of localization results with different separation among sources, three sources, RT60 = 150 ms. **a** The RMSE of localization results using different methods. **b** The RMSE of localization results using different proposed weights

noise is added to all the channels of the sound field microphone to create signals with different signal to noise ratio (SNR). In this set of experiments, the SNR is chosen from {15dB, 25 dB, 35 dB} to evaluate the performance of proposed method for different recorded signals. The results are shown in Fig. 11 below:

It can be found from Fig. 11 that with the increase of SNR, the RMSE of all the methods decline, while their decline rate varies. Figure 11a shows that when the SNR increase from 15 to 25 dB, the RMSE of the proposed method declines sharply from 10.9 to 6.8. Then the RMSE remains relatively stable when the SNR increases from 25 to 35 dB, which means that when the SNR is equal or lower than 25 dB, noise has little effect on the localization results. While for the other methods, it can be observed that their RMSE decreases constantly with the increase of the SNR, which indicates that among all the experimental situations, noise always has a great impact on the localization accuracy of the reference methods. A similar conclusion can be also obtained from Fig. 11b. In summary, the proposed method can achieve a better performance in processing noisy signals. More specifically, when the noise is moderately added into the signals (i.e., when SNR = 25 dB or 35 dB), the proposed method can effectively reduce the impact of noise, even the localization accuracy of the proposed method is degraded by the harsh effects of lower SNR (i.e., SNR = 15 dB), it still achieves better performance than the reference methods.

To further evaluate the performance of the proposed method, a series of experiments using actual recorded signals are conducted. The information about the acoustic chamber used in these experiments is given in the first part of the section. The locations of the speakers are set in advance, and the angle difference between sources is chosen from {60°, 90°, 120°, 150°, 180°}. The results are shown in Fig. 12:

Similar to the experiments conducted using simulation signals, the RMSE of all the methods increases with the decline of separation between sources. The proposed method also achieves better performance in all the experimental situations compared with reference methods and incomplete versions of the proposed method. It should be noted that the RMSE of all the methods using actual recorded signals is much larger than that using simulation signals. This is not only because both noise components and reverberation components exist in the actual recorded signals but also due to the unpredictable noise caused by the incomplete sound insulation of acoustic chamber or the stronger reflections from the objects such as devices in the room. Even so, the proposed method still has the lowest RMSE, which demonstrates the validity of the proposed method in actual environments.

6 Conclusions

In this paper, the unavoidable problem of outliers' existence in the detected SSZs has been explained and verified. To solve this problem, an assumption based on the characteristic of the whole SSZ has been proposed. Refer to this assumption, each T-F point within SSZ is weighted by the PORW and SORW proposed in this

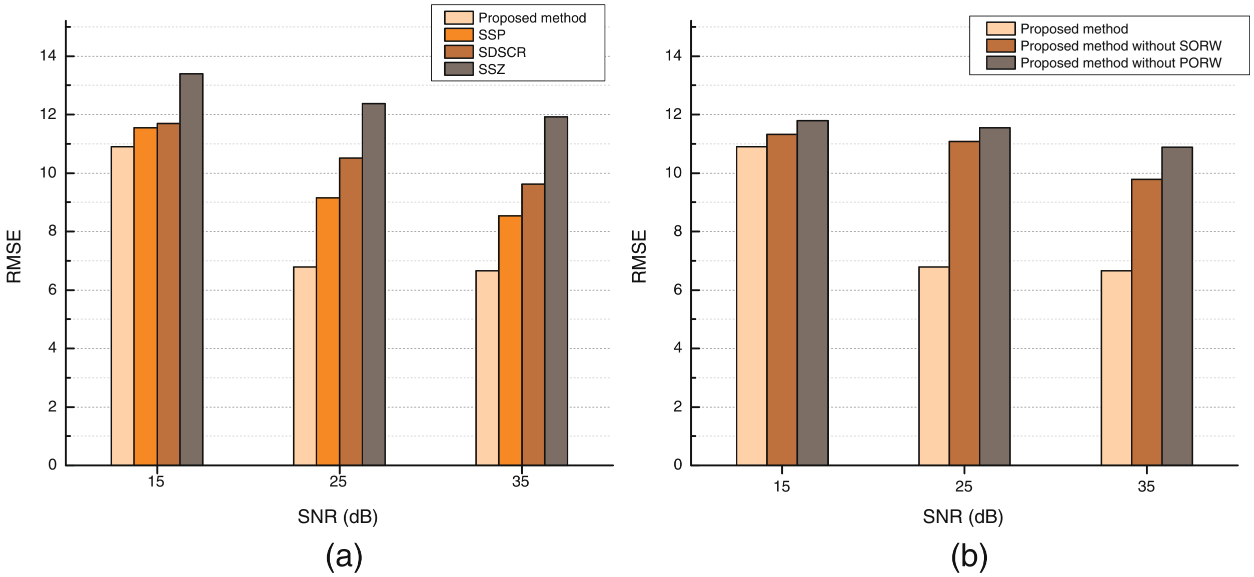

Fig. 11 RMSE of localization results by using signals with different SNR, RT60 = 450ms, three sources, separation between adjacent sources is 60°. **a** The RMSE of localization results using different methods. **b** The RMSE of localization results using different proposed weights

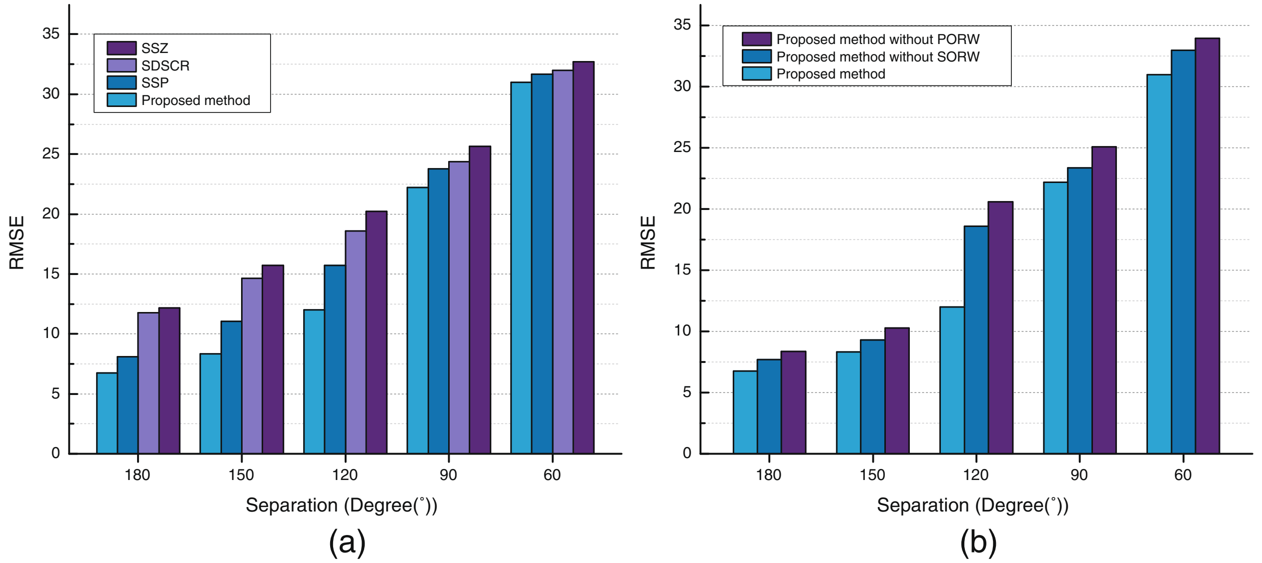

Fig. 12 RMSE of localization results by using actual recorded signals, two sources. **a** The RMSE of localization results using different methods. **b** The RMSE of localization results using different proposed weights

paper to weaken the effect of outliers. Finally, the proposed PORW and SORW are combined as a composite weight to maximize the difference between the weight applied to the outliers and desired points, then the outliers are removed by an empirical threshold. The proposed method has been proved to achieve better performance over various experimental environments compared with the reference methods. Besides, the proposed method can be integrated into other localization frameworks making use of DOA histograms plotted by SSZ-based methods. The future work lies in the scientific selection of thresholds instead of using empirical thresholds.

Abbreviations

SSZ: Single source zone; T-F: Time-frequency; PORW: Point offset residual weight; SORW: Source offset residual weight; TDOA: Time difference of arrival; MUSIC: Multiple signal classification; ESPRIT: Estimation of signal parameters via rotational invariance; DOA: Direction of arrival; DPD: Direct path dominance; ICA: Independent component analysis; SCA: Sparse components analysis; SSP: Single source point; LRSS: Low-reverberant single-source; SWS: Statistically weak source; SDS: Statistically dominant sources; SDSCR: Statistically dominant source component removal; FLU: Front left up; FRD: Front right down; BLD: Back left down; BRU: Back right up; STFT: Short time Fourier transform; NCC: Normalized cross-correlation; RMSE: Root mean squared error; DPDS: Desired points in the desired SSZ; DPFS: Desired points in falsely detected SSZ; ODS: Outliers in desired SSZ; OFDS: Falsely detected SSZ; SSCE: Statistical source component equalization

Acknowledgements

Not applicable.

Authors' contributions

Gao S performed the whole research and wrote the paper. Jia M provided support to the writing and experiments. Bao C supervised the research. The authors read and approved the final manuscript.

Funding

This work was supported by the National Natural Science Foundation of China under Grants (No. 61971015, 61831019) and the Cooperative Research Project of BJUT-NTUT (No. NTUT-BJUT-110-05).

Availability of data and materials

Not applicable.

Declarations

Competing interests

The authors declare that they have no competing interests.

Received: 30 November 2020 Accepted: 23 May 2021
Published online: 24 June 2021

References

1. J.M. Valin, F. Michaud, B. Hadjou, J. Rouat, Localization of simultaneous moving sound sources for mobile robot using a frequency-domain steered beamformer approach. Proc. IEEE. Int. Conf. Robot. Automat. **1**, 1033–1038 (2004)
2. X. Chen, W. Wang, Y. Wang, X. Zhong, A. Alinaghi, Reverberant speech separation with probabilistic time-frequency masking for b-format recordings. Speech Comm. **68**, 41–54 (2015)
3. Y. Yu, W. Wang, P. Han, Localization based stereo speech source separation using probabilistic time-frequency masking and deep neural networks. EURASIP J. Audio, Speech, Music Process **2016**(1), 1–18 (2016)
4. T.V. Bogaert, E. Carette, J. Wouters, Sound source localization using hearing aids with microphones placed behind-the-ear, in-the-canal, and in-the-pinna. Int. J. Audiol. **50**(3), 164–176 (2011)
5. T. Latif, E. Whitmire, T. Novak, A. Bozkurt, Sound localization sensors for search and rescue biobots. IEEE Sensors J. **16**(10), 3444–3453 (2016)
6. C. Knapp, G. Carter, The generalized correlation method for estimation of time delay. IEEE Trans. Acoust. Speech Signal Process. **24**(4), 320–327 (1976)
7. J. Benesty, J. Chen, Y. Huang, Time-delay estimation via linear interpolation and cross correlation. IEEE Trans. Speech Audio Process **12**(5), 509–519 (2004)
8. F. Nesta, M. Omologo, Generalized state coherence transform for multidimensional TDOA estimation of multiple sources. IEEE Trans. Audio Speech Lang. Process. **20**(1), 246–260 (2012)

9. J. Chen, J. Benesty, Y. Huang, Time delay estimation in room acoustic environments: an overview. EURASIP J. Adv. Signal Process, **26503**, 1-19 (2006)
10. P. Annibale, R. Rabenstein, *Accuracy of Time-difference-of-arrival based source localization algorithms under temperature variations, 2010 4th International Symposium on Communications, Control and Signal Processing (ISCCSP)* (Limassol, 2010), pp. 1–4. https://doi.org/10.1109/ISCCSP.2010.5463320
11. R. Schmidt, Multiple emitter location and signal parameter estimation. IEEE Trans. Antennas Propag. **34**(3), 276–280 (1986)
12. J.P. Dmochowski, J. Benesty, S. Affes, in *Applications of signal processing to audio and acoustics, 2007 IEEE workshop on*. Broadband music: opportunities and challenges for multiple source localization (2007), pp. 18–21
13. C.T. Ishi, O. Chatot, H. Ishiguro, N. Hagita, in *Intelligent robots and systems, 2009. IROS 2009. IEEE/RSJ international con-ference on*. Evaluation of a music-based real-time sound localization of multiple sound sources in real noisy environments (2009), pp. 2027–2032
14. R. Roy, T. Kailath, Esprit-estimation of signal parameters via rotational invariance techniques. IEEE Trans. Acoust. Speech Signal Process. **37**(7), 984–995 (1989)
15. Argentieri, S.; Danes, P. Broadband variations of the MUSIC high-resolution method for sound source localization in robotics. In Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems, San Diego, CA, USA, 29 October–2 November 2007; Volume 11, pp. 2009–2014.
16. Dmochowski, J.P.; Benesty, J.; Affes, S. Broadband MUSIC: opportunities and challenges for multiple source localization. In Proceedings of the IEEE Workshop on Applications of Signals Processing to Audio and Acoustics, New York, NY, USA, 21–24 October 2007; pp. 18–21.
17. B. Rafaely, K. Alhaiany, Speaker localization using direct path dominance test based on sound field directivity. Signal Process. **143**, 42–47 (2018)
18. Z. I. Khan, M. M. Kamal, N. Hamzah, K. Othman and N. I. Khan, “Analysis of performance for multiple signal classification (MUSIC) in estimating direction of arrival,” 2008 IEEE International RF and Microwave Conference, Kuala Lumpur, 2008, pp. 524-529, doi: 10.1109/RFM.2008.4897465.
19. S. Haykin, *Unsupervised Adaptive Filtering: Volume I Blind Source Separation* (Wiley Interscience, New York, 2000)
20. A. Hyvarinen, J. Karhunen, E. Oja, *Independent component analysis* (Wiley, New York, 2001)
21. T.-W. Lee, *Indepndent component analysis: theory and ap-plications* (Kluwer Academic, Boston, 1998)
22. P. Comon, C. Jutten, *Handbook of blind source separation* (Academic Press, Salt Lake City, 2010)
23. N. Epain, C.T. Jin, Independent component analysis using spherical microphone arrays. Acta. Acust. United Acust. **98**, 91–102 (2012)
24. Noohi, T.; Epain, N.; Jin, C.T. Direction of arrival estimation for spherical microphone arrays by combination of independent component analysis and sparse recovery. In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, Vancouver, BC, Canada, 26–30 May 2013; Volume 32, pp. 346–349.
25. Noohi, T.; Epain, N.; Jin, C.T. Super-resolution acoustic imaging using sparse recovery with spatial priming. In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, South Brisbane, Australia, 19–24 April 2015; pp. 2414–2418.
26. D. Pavlidi, A. Griffin, M. Puigt, A. Mouchtaris, Real-time multiple sound source localization and counting using a circular microphone array. IEEE Trans. Audio Speech Lang. Process. **21**(10), 2193–2206 (2013)
27. O. Yilmaz, S. Rickard, Blind separation of speech mixtures via time-frequency masking. IEEE Transact. signal Process **52**(7), 1830–1847 (2004)
28. M. Jia, J. Sun, C. Bao, Real-time multiple sound source localization and counting using a soundfield microphone. J. Ambient. Intell. Humaniz. Comput. **8**(6), 829–844 (2017)
29. S. Mohan, M.E. Lockwood, M.L. Kramer, D.L. Jones, Localization of multiple acoustic sources with small arrays using a coherence test. J. Acoust. Soc. Am. **123**(4), 2136–2147 (2008)
30. C. Avendano and J. M. Jot, “Ambience extraction and synthesis from stereo signals for multi-channel audio up-mix,” in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'02), Orlando, Florida, USA, May 2002.
31. A. Aissa-El-Bey, N. Linh-Trung, K. Abed-Meraim, A. Belouchrani, Y. Grenier, Underdetermined blind separation of nondisjoint sources in the time-frequency domain. IEEE Transact. Signal Process **55**(3), 897–907 (2007). https://doi.org/10.1109/TSP.2006.888877
32. H. Zhang, G. Hua, L. Yu, Y. Cai, G. Bi, Underdetermined blind separation of over-lapped speech mixtures in time-frequency domain with estimated number of sources. Speech Comm. **89**, 1–16 (2017)
33. K. Wu, V. G. Reju and A. W. H. Khong, “Multisource DOA estimation in a reverberant environment using a single acoustic vector sensor,” in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 26, no. 10, pp. 1848-1859, 2018, doi: https://doi.org/10.1109/TASLP.2018.2845121.
34. M. Cobos, J.J. Lopez, D. Martinez, Two-microphone multi-speaker localization based on a Laplacian mixture model. Digital Signal Process **21**(1), 66–76 (2011)
35. M. Jia, Y. Wu, C. Bao, J. Wang, Multiple sound sources localization with frame-by-frame component removal of statistically dominant source. Sensors **18**(11), 3613 (2018) 1-21
36. NTT database, 2008. More information see as http://www.ntt-at.com/product. Accessed 2009
37. Gao S, Jia M, Wu Y, et al. Multiple sound sources localization by using statistical source component equalization[C]// ICCPR ‘19: 2019 8th International Conference on Computing and Pattern Recognition. 2019.
38. B. Gunel, H. Hacihabiboglu, A.M. Kondoz, Acoustic source separation of convolutive mixtures based on intensity vector statistics. IEEE Trans. Audio Speech Lang. Process. **16**(4), 748–756 (2008)
39. D.R. Campbell, K.J. Palomki, G.J. Brown, A matlab simulation of “shoebox” room acoustics for use in research and teaching. Comput Inf Syst J **9**, 48–51 (2005)

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.