J Ambient Intell Human Comput
DOI 10.1007/s12652-016-0388-x

ORIGINAL RESEARCH

Real-time multiple sound source localization and counting using a soundfield microphone

Maoshen Jia[1] · **Jundai Sun**[1] · **Changchun Bao**[1]

Received: 25 March 2016 / Accepted: 13 June 2016

Abstract In this work, a multiple sound source localization and counting method based on a relaxed sparsity of speech signal is presented. A soundfield microphone is adopted to overcome the redundancy and complexity of microphone array in this paper. After establishing an effective measure, the relaxed sparsity of speech signals is investigated. According to this relaxed sparsity, we can obtain an extensive assumption that "single-source" zones always exist among the soundfield microphone signals, which is validated by statistical analysis. Based on "single-source" zone detecting, the proposed method jointly estimates the number of active sources and their corresponding DOAs by applying a peak searching approach to the normalized histogram of estimated DOA. The cross distortions caused by multiple simultaneously occurring sources are solved by estimating DOA in these "single-source" zones. The evaluations reveal that the proposed method achieves a higher accuracy of DOA estimation and source counting compared with the existing techniques. Furthermore, the proposed method has higher efficiency and lower complexity, which makes it suitable for real-time applications.

Keywords Multiple source localization · Direction of arrival estimation · Sparsity · Soundfield microphone

✉ Maoshen Jia
jiamaoshen@bjut.edu.cn

Jundai Sun
sunjundai@emails.bjut.edu.cn

Changchun Bao
baochch@bjut.edu.cn

[1] Speech and Audio Signal Processing Lab, School of Electronic Information and Control Engineering, Beijing University of Technology, Beijing, China

1 Introduction

Multiple sound source localization is a hot topic in audio signal processing over recent decades, in which accurate estimation of the direction of arrival (DOA) is a vital issue indisputably. It is widely utilized in many application areas, such as teleconference, speech enhancement, hearing aid, target tracking and so on (Argentieri and Danes 2007; Tim et al. 2011; Nakadai et al. 2003; Bechler et al. 2004; Su et al. 2015; Yi and Kuroda 2014). Moreover, the information obtained by sound source localization could be widely used for parameter coding and reconstruction of the sound scene (Asaei et al. 2016; Jia et al. 2015; Zheng et al. 2016).

In the early years, research in the field of DOA estimation mainly focused on the time difference of arrival (TDOA) combined with generalized cross-correlation phase transform (GCC-PHAT) based on different microphone arrays, such as linear, circular microphone array and so on (Knapp and Carter 1976). The idea of the TDOA is to acquire the direction of the sound source by estimating the relative delay between the microphones. A series of improvements are presented to the TDOA method, in which both the multi-path and the so-far unexploited information among the microphone pairs are considered (Benesty et al. 2004). Some further works have been conducted to change the geometry of the array aiming to get a more accurate estimation of DOA (Karbasi and Sugiyama 2007; Dmochowski et al. 2007a). However, these just work for single sound source localization system. And the performance will be severely decreased if they are used for the localization of multiple sound sources. For multiple simultaneously active source localization, an extension of the GCC-PHAT method has been proposed where the second peak is considered as an indicator of a

Published online: 23 June 2016

possible second source in (Bechler and Kroschel 2003). Therefore, the problem of sources' overlap is solved effectively. Nevertheless, most of the localization methods based on TDOA still require excessive microphones to improve the reliability of the TDOA estimates, whereas they are not appropriate in the case where a limited amount of microphones is available (Nesta and Omologo 2012).

The well-known multiple signal classification (MUSIC) is one of the earliest methods for multiple source DOA estimation (Schmidt 1986). Spatial spectrum function is acquired by using the orthogonality of the signal and the noise subspace, then the DOA of the source signal will be detected through spectral peak searching. In addition, a lot of researches and improvements are conducted on the MUSIC algorithm for DOA estimation (Dmochowski et al. 2007b; Belloni and Koivunen 2003; Zhang et al. 2009; Ishi et al. 2009a; Shiiki and Suyama 2015). Nevertheless, there are two issues regarding the MUSIC algorithm, one is the computational cost, while the other is the requirement of previous knowledge about the number of actual sources (Ishi et al. 2009b).

Independent component analysis (ICA) (Comon and Jutten 2010) relies on the statistical independence of the individual sources. It has been widely used in BSS to separate the sources from mixtures in a determined case, which could be employed for multiple sound source localization as well (Loesch et al. 2009; Lombard et al. 2011; Sawada et al. 2005). An assumption that the number of dominant sources does not exceed the number of microphones in each time-frequency region has been proposed in (Nesta and Omologo 2012). Similarly, sparse component analysis (SCA) approaches (Swartling et al. 2011; Blandin et al. 1950; Pavlidi et al. 2012) are under the assumption that there is always one source active respecting to the others in most time-frequency zones (i.e., the energy of it is much stronger that the others'). As a result, the problem of the multiple source localization might be solved by single source DOA estimation approaches. For example, a recent method based on the assumption has been proposed which achieved a high estimation accuracy by using a pair of coincident soundfield microphones (Zheng et al. 2013).

It should be noted that most of the SCA methods are dependent on the w-disjoint orthogonally (W-DO) property of sound sources meaning that respective time-frequency representations of sources are located in different time-frequency bands. However, if the number of simultaneously occurring sources are four or above, speech sources are more likely to overlap in the time-frequency domain, in other words, more than one source are active in a time-frequency bin with high probability. It shows that this assumption is less accurate when the number of sound source increases, which would also affect the localization accuracy of the DOA estimation.

Considering this situation, a multiple sound source localization method that applied the relaxed sparsity constraints on the source signals has been proposed in (Pavlidi et al. 2013). The method is based on an assumption that there always exist several regions among the signals recorded by the circular microphone array (CMA-SCM in this paper), in which one source is dominant over others. It implies that DOA estimation could be proceeded in these regions to improve the localization accuracy. Some performance comparisons show that the method is slightly better than other localization and source counting methods, both in accuracy and computational complexity; in addition, it is suitable for real-time applications due to the low computational complexity and the low requirement of historical samples. However, there are two issues of the method should be noted, one is the horizontal restriction and the other is the excessive number of microphones. Moreover, the high accuracy of the method is based on the case where the number of sources is no more than the number of microphones.

In this paper, we present a novel method of multiple sound source localization and counting using a soundfield microphone. The proposed method explores a new perspective for the sparsity of speech sources in company with the superiority (each channel is co-located and spatially separated, so the influence of sources with different DOAs is converted to the relative amount of each source recorded by each channel) of the soundfield microphone. Unlike the existing methods based on W-DO, the proposed method employs a relaxed sparsity of sources, which leads to the advantage compared to most of the existing DOA techniques. The relaxed sparsity means that the multiple simultaneously occurring speech sources do not need to be W-DO ones, as long as they have several time-frequency components staggered. This assumption makes it possible to estimate the DOA of multiple sound source at a specific time-frequency region called "single-source" zone. In this work, we make an investigation about the sparsity of speech sources and the percent of "single-source" region among signals recorded via a soundfield microphone by statistical analysis. Through experiments, it is confirmed that more than one source is active in time-frequency region with the increasing number of simultaneously occurring speech signals. However, the "single-source" region still exist among the soundfield microphone signals. This implies that for simultaneously occurring speech sources, since we have detected the "single-source" zones of all sources, we can apply any known single source DOA algorithm over these zones. Subsequently, a method combining the source counting with DOA estimation is proposed, which is based on the following steps: (a) finding the "single-

source" zones in the time-frequency domain, where one source is clearly dominant over others; (b) performing DOA estimation by using the mathematical properties of soundfield microphone; (c) collecting all the results on statistical histogram; (d) jointly performing the DOA estimation and source counting through the post-processing of the histogram. The relaxed sparsity of the speech and the advantages of soundfield microphone are proved to be an exceptional choice to simplify the complexity of the DOA estimation problem sharply while maintaining a high accuracy. Among the results, we provide a series of performance comparisons of our proposed method with the DOA estimation approach in CBSS (collaborative blind source separation) and the method mentioned in (Pavlidi et al. 2013). The results show that the proposed method achieves a high accuracy of multiple source localization using a soundfield microphone, and can be used for real-time processing due to its low computational complexity and low requirement of historical samples. We also present the robustness of the proposed method in different reverberation conditions.

The key contributions of this paper can be summarized as follows. The relaxed sparsity of the speech signals is investigated and validated. Based on this property, a new approach to estimating the DOAs of simultaneously occurring speech sources is proposed. Compared to the existing techniques, the proposed method achieves better efficiency only via a soundfield microphone with four channels.

The remainder of the paper is organized as follows: Section 2 introduces the soundfield microphone and investigates the sparsity of speech signals. Section 3 introduces the proposed localization method. Experimental results are presented in Sect. 4, while conclusions are drawn in Sect. 5.

2 Problem statement

In this section, the soundfield microphone and a new sparsity assumption of speech sources are introduced.

2.1 Overview of soundfield microphone

The soundfield microphone consists of four channels, which are called front left up (FLU), front right down (FRD), back left down (BLD), back right up (BRU), respectively (Sound 2015). The conceptual architecture of the soundfield microphone is shown in Fig. 1 (x, y, z is the coordinate axis in the Descartes coordinate system).

Each channel of the structure can be placed by microphone with different directional characteristics called Polar Patterns scientifically. It means that microphones with different polar patterns could be employed in the process of recording. For better directivity, super-cardioid microphones are chosen in this paper. The polar response of super-cardioid microphone is shown in Fig. 2.

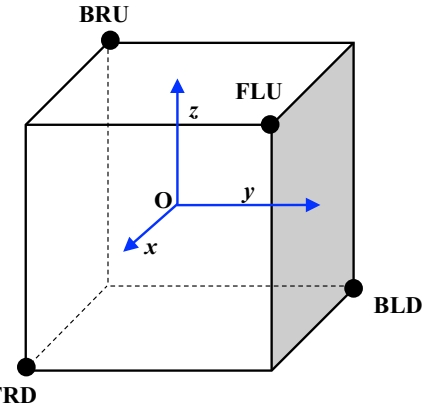

Fig. 1 Conceptual architecture of the soundfield microphone

In addition, the raw recording of the soundfield microphone is called the A-format, i.e. $\{S_{FLU}, S_{FRD}, S_{BLD}, S_{BRU}\}$. From the A-format recordings, the B-format signal could be derived for post-processing by (Gunel et al. 2008; Pulkki 2007):

$$\begin{cases} S_W = S_{FLU} + S_{FRD} + S_{BLD} + S_{BRU} \\ S_X = S_{FLU} + S_{FRD} - S_{BLD} - S_{BRU} \\ S_Y = S_{FLU} - S_{FRD} + S_{BLD} - S_{BRU} \\ S_Z = S_{FLU} - S_{FRD} - S_{BLD} + S_{BRU} \end{cases}, \tag{1}$$

where the W is an omnidirectional channel and the others namely, X, Y, and Z are three Cartesian bi-directional channel, in later section we take $\{S_1, S_2, S_3, S_4\}$ to present $\{S_{FLU}, S_{FRD}, S_{BLD}, S_{BRU}\}$ for simplicity. Besides, the spatial information of source S is preserved in the B-format signal by:

$$\begin{cases} S_W = \dfrac{\sqrt{2}}{2} \cdot S \\ S_X = \cos\mu \cdot \cos\eta \cdot S \\ S_Y = \sin\mu \cdot \cos\eta \cdot S \\ S_Z = \sin\eta \cdot S \end{cases}, \tag{2}$$

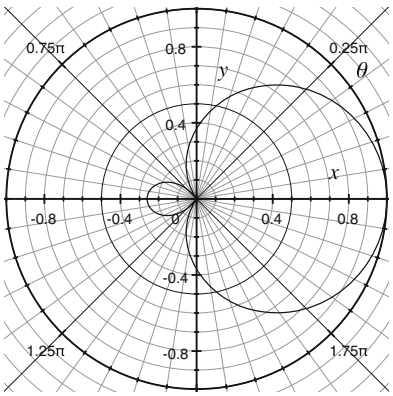

Fig. 2 Polar response of super-cardioid

where S is the sound source signal, μ and η are the azimuth and elevation of the sound source respecting to the center of the soundfield microphone. In Sect. 3, the proposed method estimates the DOAs of the active sound sources by using the B-format signal of sources.

2.2 Exploring sparsity of speech sources

A speech signal is known to be sparse in the short-term time-frequency domain (Zheng 2013). It means that a large percentage of energy is concentrated in a small number of time-frequency instants. An assumption has been proposed based on the sparsity: a time-frequency zone can be always detected where one source is dominant over others when multiple speech sources are occurring simultaneously (Pavlidi et al. 2013). Following this assumption, we define a time-frequency region as a "single-source" zone where the energy of one source is far stronger than any other sound sources; furthermore, we suppose that each source itself will always have several corresponding "single-source" zones in the time-frequency domain. In this section, the existence of the "single-source" region is analyzed and validated respectively.

In general, a speech signal has a time-frequency representation obtained by a linear time-frequency transform (Jia et al. 2015). Specifically, for a general dictionary of atoms $D = \{\phi_{n,k}\}$ satisfied $\phi_{n,k} \in k^2(\mathbb{Z})$ and $||\phi_{n,k} = 1||$, there exists a linear time-frequency transform for a speech signal $s[m]$ defined by:

$$S(n,k) = \sum_{m=-\infty}^{\infty} s[m]\phi^*_{n,k}[m] = <s, \phi_{n,k}> \tag{3}$$

where $s[m]$ and $S(n, k)$ denote the representation of signal s in time domain and time-frequency domain, respectively. Moreover, n and k represent the frame number and frequency index, respectively, and $\phi^*_{n,k}$ represents the conjugate function of $\phi_{n,k}$. The time-frequency atoms such as short-time Fourier transform (STFT) basis functions and discrete cosine transform (DCT) basis functions are commonly used in speech signal processing. In this work, STFT is chosen as the time-frequency atom.

From (Jia et al. 2015), we can find that W-DO reveals the sparsity of multiple simultaneous speech signal. The W-DO means that different time-frequency representations of speech signals locate in different bands of the domain. However, simultaneously occurring speech signal is likely to overlap in this domain, which means that more than one speech signals are active in a time-frequency band with certain probability. Hence, multiple sources localization based on W-DO assumption cannot always obtain the accurate estimation of DOA.

In this work, a new perspective for the sparsity of speech signal is explored compared to W-DO. We define a "time-frequency analysis zone" $(\mathcal{N}, \mathcal{K})$ as a set of adjacent time-frequency points (n, k), and omit $\mathcal{N}$ in the $(\mathcal{N}, \mathcal{K})$ for simplicity.Then the sparsity assumption of speech signal proposed in this work is that there is at least one "time-frequency analysis zone" for each sound source where it is "isolated" or absolutely dominant over the other sources. The limit of this assumption is much weaker than the WDO's, because in most cases, the majority of multiple sound sources in the time-frequency domain is overlapped except in a few "single-source" zones.

To validate this assumption, we define the region energy of speech $S(n, k)$ in a given time-frequency region $\mathcal{K}$ as follow:

$$\sum_{(n,k)\in\mathcal{K}} |S(n,k)|^2. \tag{4}$$

Hence, for source i among M multiple simultaneously occurring speech sources, inter source energy ratio (ISER) is defined by:

$$ISER(S_i(n,k))_{\mathcal{K}_i} = \frac{\sum_{(n,k)\in\mathcal{K}_i} |S_i(n,k)|^2}{\sum_{j=1}^{M} \sum_{(n,k)\in\mathcal{K}_i} |S_j(n,k)|^2}, \tag{5}$$

where n and k represent time and frequency index, respectively, $S_i(n,k)$ and $S_j(n,k)$ are the time-frequency representation of signal S_i and S_j, respectively. If there exists a $\mathcal{K}_i$ for an arbitrary source $i(i = 1, 2, \ldots, M)$:

$$ISER(S_i(n,k))_{\mathcal{K}_i} > \zeta, \tag{6}$$

where ζ is a user-defined threshold between [0,1] which is used for the "single-source" zone determination. Then $\mathcal{K}_i$ is the "single-source" zone of source i.

In order to verify the existence of single sound source region further, a total of 36 sentences (the sampling frequency is 16 kHz) from the NTT speech database were used for testing. In the following evaluation, all the test data is from the NTT database unless otherwise state. Each sentence was divided into a group with the other $M - 1(2 \le M \le 7)$ sentences in the time domain resulting in M simultaneously occurring speech conditions. For $M = 2$, each sentence was divided into a group with each of the remaining 35 sentences resulting $36 \times 35 = 1260$ combinations. For $M > 2$, each sentence was randomly grouped 35 times with $M - 1$ other sentences to give the same number (1260) combinations as for $M = 2$. Here, STFT basis is applied to form the time-frequency atoms, i.e., $\phi_{n,k}$ in 3 is:

$$\phi_{n,k}[m] = g_{n,k}[m] = g[n-m] \exp\left(\frac{i2\pi km}{N}\right) \tag{7}$$

where n $(1 \le n \le N)$ and k $(1 \le k \le \mathcal{L})$ are frame number and frequency index, respectively, and $g[\cdot]$ is the window function. In this paper, a sine window is chosen to meet the demand of speech processing. The number of STFT points is 2048 in each frame, i,e., $\mathcal{L} = 2048$.

We search for the "single-source" zone where the instantaneous normalized energy of one source is much stronger than the others'; more specifically, we collect the percentage of "single-source" zone over the total time-frequency analysis area, and the results are show in Fig. 3, where the width of $\mathcal{K}$ is 128 and the threshold is set by:

$$\zeta = 0.9. \tag{8}$$

From Fig. 3, it can be observed that more than 1/4 regions among all the time-frequency zones of each signal are "isolated" for $M = 2$, and there still exist "single-source" zones when there are seven signals occurring simultaneously, i.e. $M = 7$. However, the percent of "single-source" region will decline rapidly with the increasing number of simultaneously occurring speech signals; especially for $M \ge 5$, the percent will below 5 % and "single-source" zones will hardly be detected ultimately. In other words, it shows that there must be several "single-source" zones for each speech signal. However, this phenomenon gets less evident when the number of speech signals increases.

2.3 Exploring sparsity of microphone signals

The sparsity described above is among the speech signals themselves. Then we aim to investigate the existence of "single-source" zone among the signals recorded by soundfield microphone. It should be mentioned that a B-format recording assumes the microphones are co-located or co-incident as opposed to a uniform linear array where microphones are spatially separated and this may also be providing an additional advantage (single sources signals recorded by adjacent microphones in a reverberant room

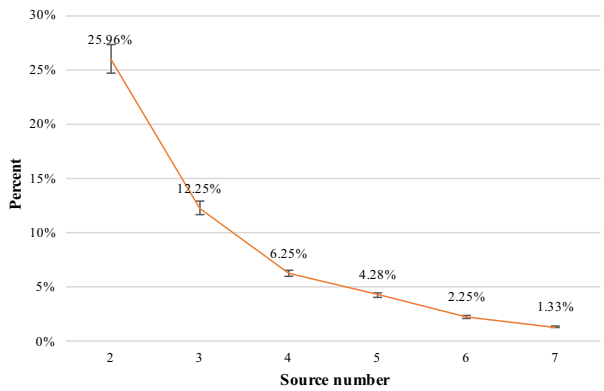

Fig. 3 The percent of "single-source" zone among sources with 95 % confidence intervals

may be more highly correlated than in a uniform linear array). Due to the use of super-cardioid microphones which has a higher directivity, when there is only one active source in some time-frequency regions, the signals recorded by soundfield microphone have a strong correlation. On the contrary, if multiple active sources are occurring simultaneously, the polar pattern that influences the relative amount of each source recorded by each microphone. It means that the correlation between each recorded signal will get weaker. Therefore, we define a cross-correlation function of time-frequency coefficients over an analysis zone to detect the "single-source" zone. More specifically, for any pair of soundfield microphone recorded signals $S_i(n,k)$ and $S_j(n,k)$, the function is defined as:

$$R_{i,j}(\mathcal{K}) = \sum_{(n,k)\in\mathcal{K}} |S_i(n,k) \cdot S_j(n,k)|, \tag{9}$$

where $i \ne j$, $S_i(n,k), S_j(n,k) \in \{S_1(n,k), S_2(n,k), S_3(n,k), S_4(n,k), \}$. The normalized cross-correlation coefficient can be obtained by:

$$r_{i,j}(\mathcal{K}) = \frac{R_{i,j}(\mathcal{K})}{\sqrt{R_{i,i}(\mathcal{K}) \cdot R_{j,j}(\mathcal{K})}}. \tag{10}$$

A necessary and sufficient condition for a zone $\mathcal{K}$ to be a "single-source" zone is

$$r_{i,j}(\mathcal{K}) = 1, \tag{11}$$

where $i \in \{1, 2, \ldots, 4\}$, $j = (i+1) mod\ 4$, and mod is a function for remainder operation. We search for "single-source" zones that satisfy the following inequality:

$$r_{i,j}(\mathcal{K}) \ge 1 - \varepsilon, \tag{12}$$

where ε is a sufficient small threshold value for the user to define.

In order to examine the existence of "single-source" zone among sound field microphone signals, a statistical analysis is taken. The speech signals from the NTT database were chosen as sound sources for analysis. The separation of two adjacent sources were set as γ= {10°, 20°, 30°, 40°}, and the source number was ranging in $\{2, 3, \ldots, 7\}$. The recording was conducted in an anechoic condition simulated by Roomsim (Campbell et al. 2005). Four different widths of $\mathcal{K}$, i.e. {256, 128, 64, 32}, were adopted for "single-source" zone detecting. We collect the percentage of "single-source" zone over the total time-frequency analysis zones, and the statistical results with 95 % confidence intervals are shown in Fig. 4a–d.

From Fig. 4, it can be observed that if the sound source is distributed in the space, the "single-source" zone could be detected with a higher probability by using soundfield microphone signals compared to the percent of "single-source" zone among source signals, the proportion

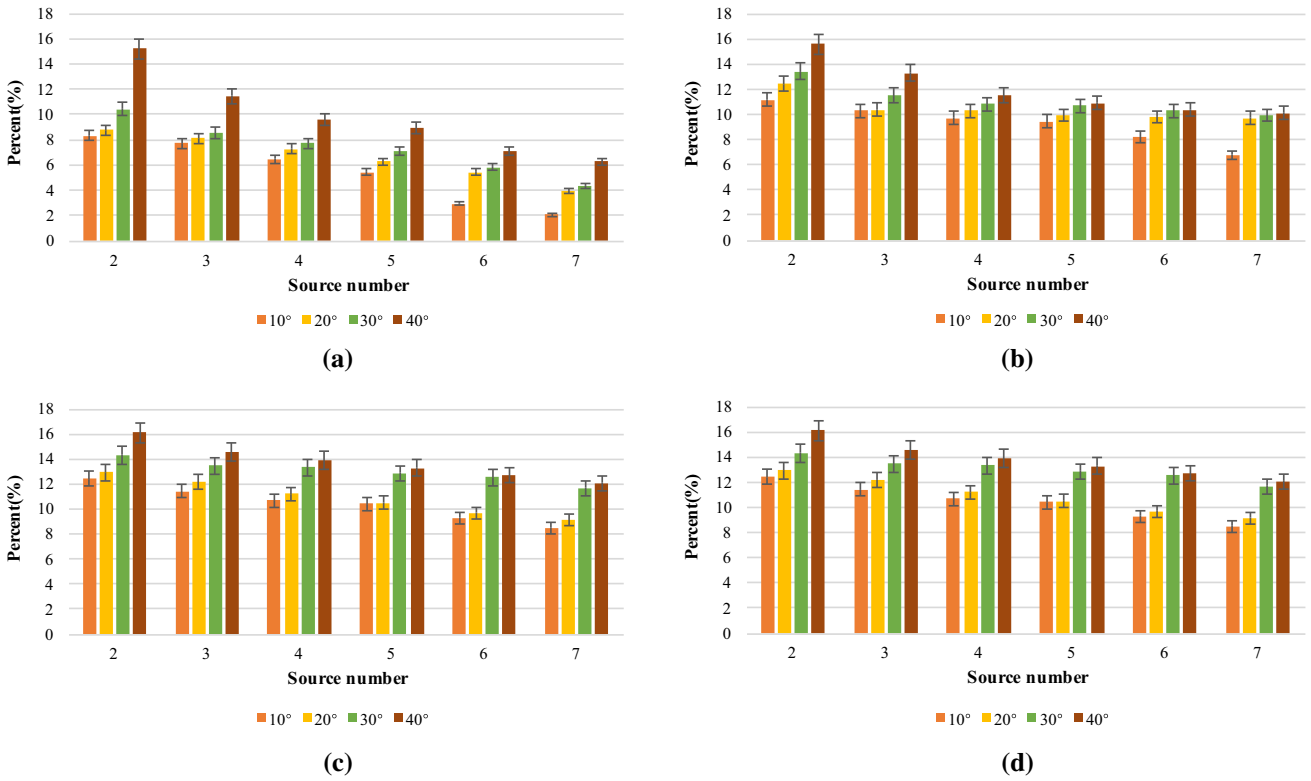

Fig. 4 The percent of "single-source" zones among soundfield microphone signals with 95 % confidence intervals, **a–d** represent the results of the width of $\mathcal{K}$ is 256, 128, 64, 32

increases which may be caused by the spatial recording procedure. In addition, three conclusions can be drawn as follows: (a) The percent of "single-source" zones become higher gradually with the increasing separation between two adjacent sources; (b) The percent of "single-source" zones decline with the increasing number of simultaneously occurring sources; (c) The percent of "single-source" zones increases with the decrease of the region's width in a certain range, while the percent will slowly converge to a constant, i.e. the percent will keep stable eventually.

In detail, the percent can still keep more than 5 % when $M = 7$, the region width is 128, 64 or 32. It also implies that the ways to find "single-source" zone based on cross-correlation among soundfield microphone recording signals are reasonable. Further, we calculate the average percent of the "single-source" zone and the approximate number of operation under four different region widths. The results are shown in Table 1.

Conclusion can be obtained from the Table 1 that the narrower width of $\mathcal{K}$ will guarantee a higher utilization ratio of "single-source" zone in a certain range, however, it is in price of the increasing complexity. Consequently, in the following experiments, we choose 128 as the width of $\mathcal{K}$ for the case when $M \leq 5$; if the number of active sources is 6 or above, we need to choose 64 as the region width for higher accuracy.

Table 1 Computational complexity

Width of $\mathcal{K}$	"Single-source" zone percent (%)	Approximate number of operation
256	7.30	536,575
128	10.68	719,360
64	11.96	1,101,400
32	12.05	1,803,876

In summarize, the assumption that there exists at least a "single-source" zone for each sound source is proved reasonable in this section.

3 Proposed Method

From the conclusion of Sect. 2, it can be obtained that there is at least one "single-source" zone for each source in the scenario where multiple sound sources occur simultaneously. The phenomenon is particularly evident for the signals recorded by soundfield microphone. In this section, a method of DOA estimation and source counting is presented, which is based on "single-source" zone detecting. In this method, the DOAs of the sound sources are estimated over these zones.

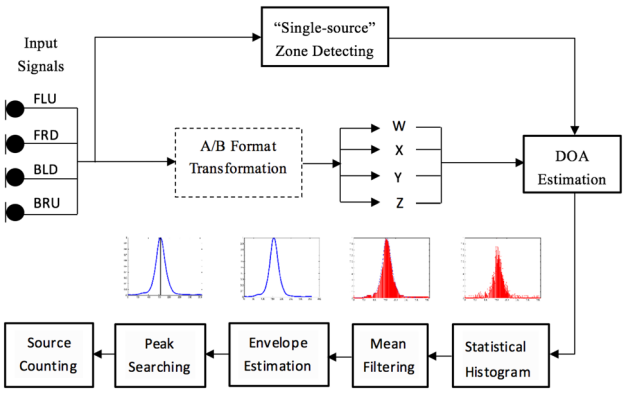

Fig. 5 System block diagram of the proposed method

The following process is performed on a frame-by-frame basis. As illustrated in Fig. 5, input signals are transformed into the time-frequency domain for "single-source" zone detecting. If there exists "single-source" zones in the current frame, the four input signals need to be converted to B-format by 1, and the processed signals, S_W, S_X, S_Y, S_Z are transformed into the time-frequency domain for DOA estimation. The histogram of DOA is obtained by statistical analysis. Subsequently, the histogram is smoothed via a mean filter and then the envelope of the histogram can be drawn by using kernel density estimation (KDE). Finally, we use the peak searching method to get the estimated DOA and the number of sound sources. More details of these processes will be described below.

3.1 "Single-source" analysis zones detection

The soundfield microphone recording signals $\{S_1, S_2, S_3, S_4\}$ are transformed into the time-frequency domain using $\mathcal{L}$-point STFT. In each frame, the frequency band is divided into C regions and the width of the region is $\mathbb{K}$, the following constraints are required to be satisfied:

$$C = \left\lceil \frac{\mathcal{L}}{\mathbb{K}} \right\rceil. \tag{13}$$

For recording signals $S_i(n,k)$ and $S_j(n,k)$, the normalized correlation coefficient $r_{i,j}(\mathcal{Z}_c)$ in time-frequency region $\mathcal{Z}_c, c \in [1, C]$ is calculated by 10. We search for the "single-source" zone by 12, then a set of $\mathcal{Z}_l$ can be obtained as follow:

$$\bigcup\{\mathcal{Z}_l\} = \{Z_l | r_{i,j}(Z_l) \geq 1 - \varepsilon, (i,j) \in D, l \in [1, C]\}, \tag{14}$$

where $D = \{(1,2),(2,3),(3,4),(4,1)\}$, the set $\bigcup\{\mathcal{Z}_l\}$ are the "single-source" zones.

3.2 DOA estimation in a "Single-source" zone

Since all the "single-source" zones have been obtained, any method of single source DOA estimation can be used. Subsequently, a method of DOA estimation over the "single-source" zones is proposed in this section.

The DOA estimation based on soundfield microphone will be proceeded in time-frequency domain. The time-frequency DOA analysis aims to obtain the spatial location (azimuth μ) at each time-frequency instant, which is based on calculating the instantaneous intensity of the signals in B-format. The sound intensity is defined by:

$$I = p \cdot u, \tag{15}$$

where p is the sound pressure and u is the particle velocity. The sound intensity can be split into two parts, namely, the active intensity and reactive intensity. The active intensity can be calculated in the time-frequency domain by (Gunel et al. 2008; Pulkki 2007):

$$I(n,k) = 2[Re\{P^*(n,k) \cdot U(n,k)\}]. \tag{16}$$

Thus, in B-format, for the Cartesian directional channel (X, Y, Z channels), the active intensity can be found as (Gunel et al. 2008; Pulkki 2007):

$$\begin{cases} I_X(n,k) = \dfrac{\sqrt{2}}{\rho c}[Re\{S_W^*(n,k) \cdot S_X(n,k)\}], (n,k) \in \bigcup Z_l \\ I_Y(n,k) = \dfrac{\sqrt{2}}{\rho c}[Re\{S_W^*(n,k) \cdot S_Y(n,k)\}], (n,k) \in \bigcup Z_l \\ I_Z(n,k) = \dfrac{\sqrt{2}}{\rho c}[Re\{S_W^*(n,k) \cdot S_Z(n,k)\}], (n,k) \in \bigcup Z_l \end{cases} \tag{17}$$

where ρ is the density of the medium, c is the speed of sound, * denotes conjugation and $Re\{\cdot\}$ denotes taking the real part of the argument. Hence, the DOA estimation $\hat{\mu}$ in "single-source" zone $\mathcal{Z}_l$ can be calculated by (Gunel et al. 2008; Pulkki 2007):

$$\hat{\mu}_{\mathcal{Z}_l} = \tan^{-1}\left(\frac{-I_Y(n,k)}{-I_X(n,k)}\right). \tag{18}$$

We can get the estimated azimuth of the active source in a certain "single-source" zone by 18, and we get the histogram through the statistics of all the DOA estimates obtained in several frames. It has to be mentioned that higher accuracy of DOA estimation will be obtain over more frames, we should balance the accuracy and delay according to the actual applications.

An example of the normalized histogram of six sources at $45°, 90°, 135°, 180°, 240°$, and $340°$ is shown in Fig. 6. Where the soundfield microphone signal was recorded in the anechoic room, the width of "single-source" zone was 64, and 9 (4 look-ahead, 1 current and 4 look-back) frames are used for statistic. The histogram of DOA estimation in all time-frequency bands is shown in Fig. 6a, while the one in "single-source" zones is shown in Fig. 6b.

It can be seen from Fig. 6a, there are only two obvious peaks around 50° and 100°, while the other four peaks are lost which implies that the mutual distortion among the multiple sound sources is very evident. However, in Fig. 6b, there are six obvious peaks whose corresponding angles are particularly close to the ones of real sound sources, i.e., all peaks representing six sources can be identified. The peak value of sound source depends on its sparsity respecting to the other sources. Specifically, if the time-frequency components of a certain source are isolated with a higher probability, its peak will be more significant; on the contrary, if the time-frequency components are overlapped with the other sources, it is difficult to find a corresponding "single-source" region that will lead a lower peak value. Above all, it can be concluded that the effect of proceeding DOA estimation in "single-source" zones is excellent from Fig. 6. In detail, the distortion between multiple sound sources will be greatly weakened. As a result, the accuracy of DOA estimation is improved, besides, it means that the source number estimated by this method increases as well.

3.3 Histogram smoothing and envelope estimation

Aiming to remove the burr around the peak and make the histogram more regular for an accurate envelope drawing, a mean filter is employed for histogram smoothing, the process can be expressed as:

$$y_i = \frac{\sum_{j=-W_h}^{W_h} x(i+j)}{2 \cdot W_h + 1}, i = 1, 2, \ldots, L, \tag{19}$$

where, $x(i)$ is the corresponding value at i in histogram, $y(i)$ is the result after mean filtering, W_h is the order of the filter and L is the number of total bins in histogram. As shown in Fig. 7, all the peaks are more identifiable compared to the DOA histogram of Fig. 6b.

It can be seen that the peaks are more evident by mean filtering and the result will be utilized for post processing. Specifically, a series of approaches can be used for envelope estimating, such as histogram equalization, LPC, kernel density estimation (KDE) and so on. We select KDE in this paper but the method is not specific to the chosen one.

KDE belongs to a non-parametric estimation method for probability density function estimation. A smooth estimated curve can be obtained in a series of observation points. In simple terms, KDE can estimate the probability density function of the corresponding probability by a random variable. If we suppose $\{y_1, y_2, \ldots, y_L\}$ represents an independent and identically distributed sample which belongs to a certain distribution with a density function f, the estimated shape of this density function f can be obtained from its kernel density estimator, which is given by:

$$\hat{f}(y) = \frac{1}{Lh} \sum_{i=1}^{L} K\left(\frac{y - y_i}{h}\right), \tag{20}$$

where $K(\cdot)$ is kernel function and $h(h > 0)$ is a smoothing parameter usually referred as the bandwidth. The estimated envelope proceeded by KDE is shown in Fig. 8.

It can be observed that we get smooth curve which contains 6 peaks around $45°, 90°, 135°, 180°, 240°$, and $340°$, respectively. The subsequent process is to search the peaks of the curve to obtain the accurate source number and its corresponding DOAs.

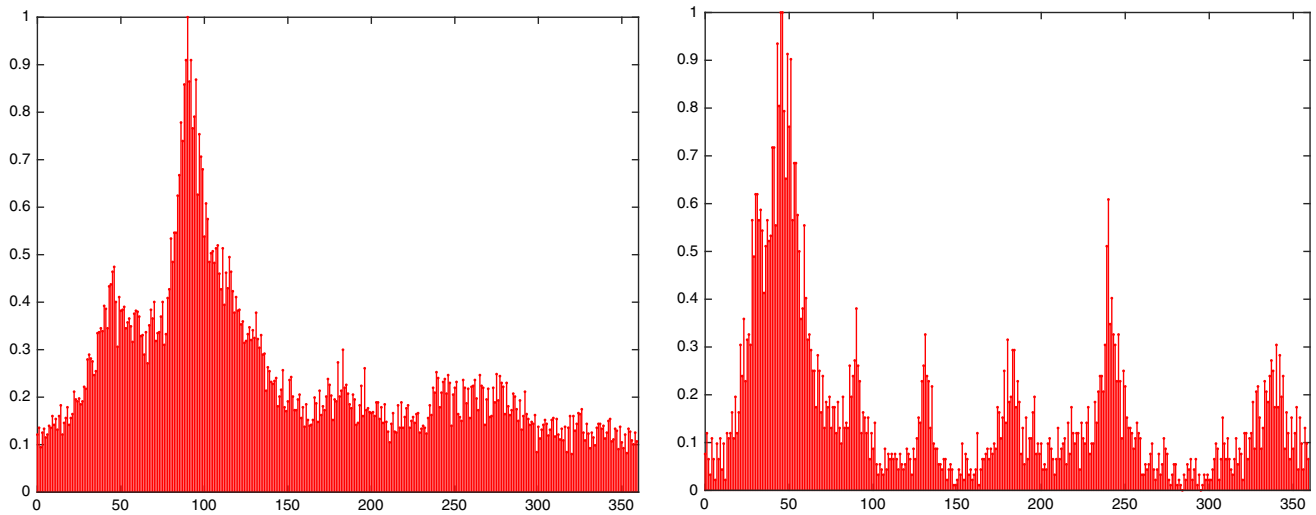

Fig. 6 Normalized histogram of six sources. **a** DOA estimation in all time-frequency bands; **b** DOA estimation in "single-source" zones

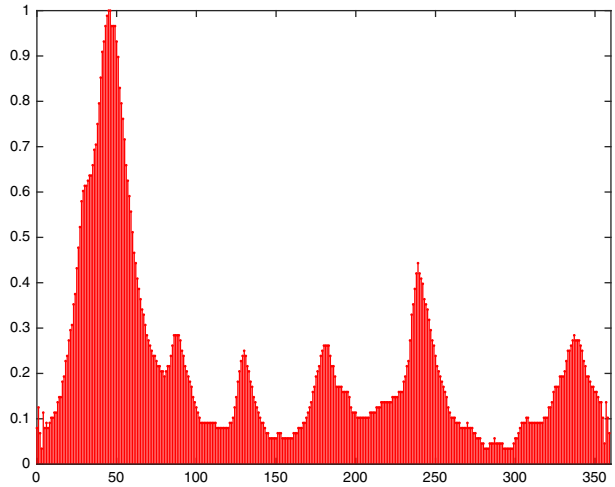

Fig. 7 Normalized histogram after mean filtering

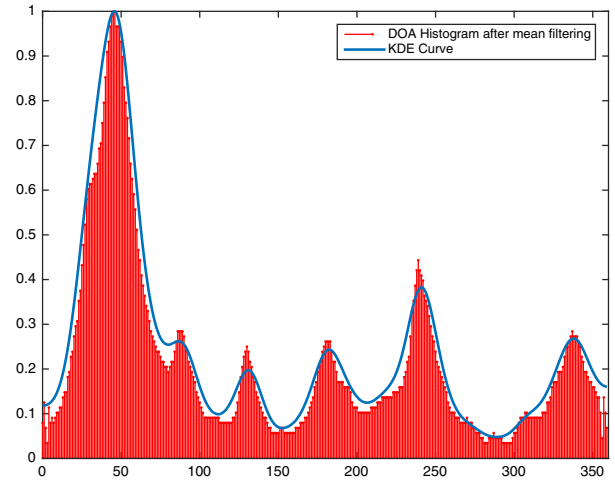

Fig. 8 Estimated envelope of histogram by KDE

3.4 Peak searching and source counting

The core idea of peak searching is to find the local maximum value in the estimated envelope of histogram (i.e., KDE curve in this paper). There are some common approaches can be used in peak searching like matching pursuit, maximum likelihood and so on. In this paper, we present a new peak searching method for low-delay algorithms. To obtain a more accurate peak searching, two important parameters are contained in the proposed method which are minimum difference η and minimum distance σ,respectively. We can set the local minimum difference for the peak as well as the minimum distance between the two peaks according to the actual applications. Then, the peak searching proceeds as follows:

1. Set the loop index $i = 1$
2. Let vector $\boldsymbol{\mathcal{P}} = \{\mathcal{P}(\boldsymbol{i}), \boldsymbol{i} = \boldsymbol{1}, \boldsymbol{2}, \ldots, \boldsymbol{L}\}$ represent the existence of peak, the initial value of $\boldsymbol{\mathcal{P}} = \boldsymbol{0}$, $\hat{f}_i$ donates the value of KDE curve at i
3. If $(\hat{f}_i - \hat{f}_{i-1})(\hat{f}_i - \hat{f}_{i+1}) > \eta : \mathcal{P}(i) = 1$
4. Increment i
5. If $i \leq L$ go to step 3
6. $\boldsymbol{D} = \{i | \mathcal{P}(i) = 1\}$donates the locations of peaks, $\boldsymbol{\mathcal{P}}' = \{\hat{f}_i, i \in \boldsymbol{D}\}$ donates the value of peaks, and set the loop index $j = 2$
7. If $D(j) - D(j-1) < \sigma$: remove $min\{\mathcal{P}'(j), \mathcal{P}'(j-1)\}$ from $\boldsymbol{\mathcal{P}}'$, and remove the corresponding index from $\boldsymbol{D}$ respectively, decrement j
8. Increment j
9. If $j \leq length(\boldsymbol{\mathcal{D}})$ go to step 7
10. $\boldsymbol{\mathcal{D}}$ is the set of locations of the all detected peaks.

It should be noted that the peak searching process is accurate and computational-efficient so that it can be done in real-time. The result of peak searching is shown in Fig. 9.

It can be seen that there are six accurate peaks around the DOAs of real sources, i.e. all the estimated DOAs are almost equal with the real ones. In addition, it is signed for source counting.

4 Results and discussion

In this section, the experiments were conducted to evaluate the performance of the proposed approach. Specifically, the whole procedure was proceeded in terms of source localization and source counting, in which several aspects were considered to assess the robustness in company with the low-delay evaluation of the proposed method as well.

For the DOA estimation case, both number and separation of sound sources were considered to value the accuracy of the proposed method. First, we conducted the evaluation on the accuracy by mean absolute estimated error (MAEE) where different number and separations of sound sources in anechoic room were tested. Then a series of processes were conducted to examine the robustness in reverberant environment. For the source counting case, the percent of correct estimated number in different scenarios was calculated. The CMA-SCM—which belongs to the family of SCA approaches, was employed as a reference method, because it have achieved an approximately perfect performance in source counting (Pavlidi et al. 2013). Subsequently, the real-time evaluation was conducted between the proposed method and the DOA estimation method in CBSS—which is a low delay approach for localization (Zheng et al. 2013).

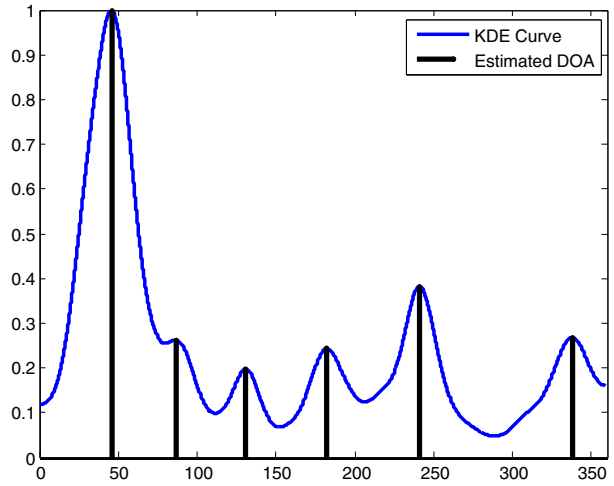

Fig. 9 The result of peak searching over KDE curve

Table 2 Experimental parameters

Parameter	Notation	Value
Sampling frequency of speech source	fs	16 kHz
Source distance	r	1 m
Time-frequency zones width		128 or 64
Overlapping in frequency		50 %
"Single-source" threshold	ϵ	0.1
Number of bins in the histogram	L	720
Length of data		1 s

The performance of the proposed method was investigated in anechoic and reverberant condition. For different test conditions, the experimental parameters and their corresponding values are listed in Table 2.

4.1 The evaluation of DOA estimation

The evaluation of DOA estimation was conducted in 5 different scenarios, i.e., Anechoic room, Quiet room, Room 1, Room 2, and Room 3. The parameters of these rooms are listed in Table 3.

NTT is a speech database containing various speakers of different countries. The Chinese speech sub-data-base in NTT was employed as the test database in this section. We used Roomsim (Campbell et al. 2005) to simulate a room of $6.25 \times 3.75 \times 2.5$ m, where $\{6.25, 3.75, 2.5\}$ represented the length, width, height and we took them as the x, y, z axis, respectively. Besides, the speed of sound was $c = 340$ m/s. The soundfield microphone was placed in the center of the room paralleling with z-axis and the power of sound sources from different directions was equal in each simulation. It should be noted that the soundfield microphone was simulated via Roomsim which was completed by simulating the recording condition of each channels. The polar patterns of the four microphones are super-cardioid. In addition, the radius of the cube in Fig.1 was 12 mm.

Table 3 Parameters of testing room

Simulated room	Absorption of the wall	T_{60} (ms)	Reflection order
Anechoic room	1	0	0
Quiet room	1	0	0
Room1	0.75	250	18
Room2	0.75	450	18
Room3	0.75	600	18

MAEE is utilized to measure the performance of the proposed method, which calculates the difference between the true DOA and the estimated DOA. The *MAEE* can be calculated by:

$$MAEE = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{P_{\max\_i}}\sum_{j=1}^{P_{\max\_i}}|\mu_{ij} - \hat{\mu}_{ij}|, \tag{21}$$

where $P_{max\_i} = max\{P_i, \hat{P}_i\}$, $\mu_{ij}$is the true DOA of the jth active source in the ith experiment, and $\hat{\mu}_{ij}$ is the corresponding estimated DOA. P_i is the number of active sources, while $\hat{P}_i$ is the estimated number of sources in the ith experiment and N is the total number of experiments. The following cases should be noted in 21:

(a) if $P_i > \hat{P}_i$: $\hat{\mu}_{ik} = 0^\circ, k \in [\hat{P}_i + 1, P_i]$
(b) if $P_i < \hat{P}_i$: $\mu_{ik} = 0^\circ, k \in [P_i + 1, \hat{P}_i]$

The *MAEE* is collected in a series of experiments proceeded in different scenarios.

4.1.1 DOA estimation in anechoic room

We measured the *MAEE* of the estimated DOA among different source numbers and separations in anechoic room. More specifically, we conducted 72 different orientations of each source number and separation. To investigate the localized accuracy, the separation was set as $\{10^\circ, 20^\circ, 30^\circ, 40^\circ\ 50^\circ\}$, the source number was $\{2, 3, 4, 5\}$, and the time-frequency zones width was 128 aiming to reduce the computational complexity. However, when the *source number* $= 6$ or 7, the time-frequency zone width was 64 to improve the accuracy. Then the *MAEE* results are shown in Fig. 10. It should be noted that the *MAEE* is larger than the separation when source number is 6 or 7 and separation is 10° which means the localization method become invalid, so the corresponding results are omitted for better resolution of the results.

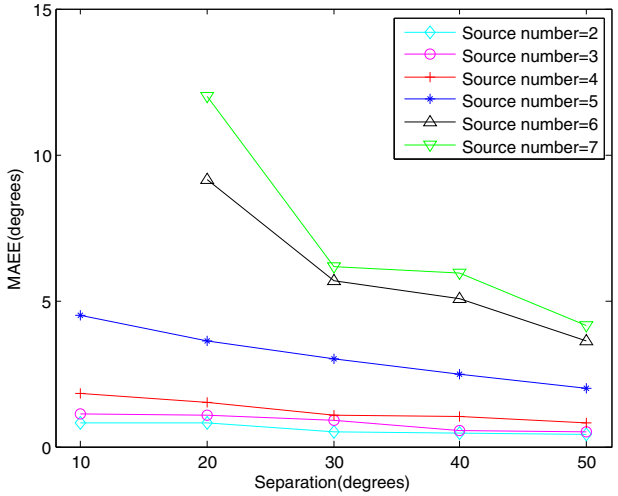

Fig. 10 *MAEE* versus separation between adjacent sources. Source number $= \{2, 3, 4, 5\}$, the width of $\mathcal{K}$ is 128; Source number $= \{6, 7\}$, the width $\mathcal{K}$ of is 64

It can be concluded that *MAEE* declines with the increase of separation between the adjacent sources. Specifically, the *MAEE* is more than 8° when the separation is 20° and *source number* = 6 or 7, but it come down sharply when the separation is 30° or above. Ultimately, the *MAEE* is about 5° when the separation was 50° for all source numbers. In addition, it should be noted that for the case where *source number* was $\{2, 3, 4, 5\}$, the width of $\mathcal{K}$ was 128 while the width of $\mathcal{K}$ was 64 when *source number* > 5 in the following subsection.

In addition, to show the consistent behavior of the proposed method no matter where the sources are located, we simulated 72 different orientations of each source number and separation in anechoic room. It should be mentioned that the initial orientation was 0° of the first source. In detail, for the *separation* = 30°, all the sources were separated by 30° and then shifted by 5° in next experiment aiming to test the performance all around the microphone. In addition, the *source number* was ranging in $\{2, 3, \ldots, 7\}$. The *MAEE* scatter of it is shown in Fig. 11.

It can be seen that the *MAEE* is always below 10° for any *source number* in $\{2, 3, \ldots, 7\}$. Moreover, the *MAEE* is approximately stable in all consistent source degrees and the variation of it converges to a constant. The result implies that there are no evident blind areas for our localization method.

4.1.2 DOA estimation in quiet room and reverberant room

To investigate the robustness of the proposed method in reverberant environments, a set of experiments were proceeded in this section. The evaluation was conducted in four simulated rooms {Quiet room, Room1, Room2, Room3}, and the parameters of them are listed in Table 3. In each scenario, we set 36 group orientations around the soundfield microphone for each source number, source number $P \in \{2, 3, 4, 5, 6\}$. It must be noted that all the orientations covered the whole 360 degrees and the separation was 50° for all the groups. For example, if the source number was 6, in the first group, the sources located at $\{0°, 50°, 100°, 150°, 200°, 250°\}$, and then shifted by 10 steps in next group, i.e. the orientations were $\{10°, 60°, 110°, 160°, 210°, 260°\}$ in second group, but remained the same separation. The *MAEE* versus source number in four rooms is shown in Fig. 12.

It can be observed that the *MAEE* is higher in reverberant environment compared with the case in anechoic room. However, the *MAEE* is still below 7°, which implies that the proposed method remains a high accuracy in reverberant environments. In addition, the *MAEE* will declined with the decrease of source count, while accumulate with the increasing T_{60} which may be interfered by the reflected version of sources. Above all, the proposed method achieves a high localized accuracy of any

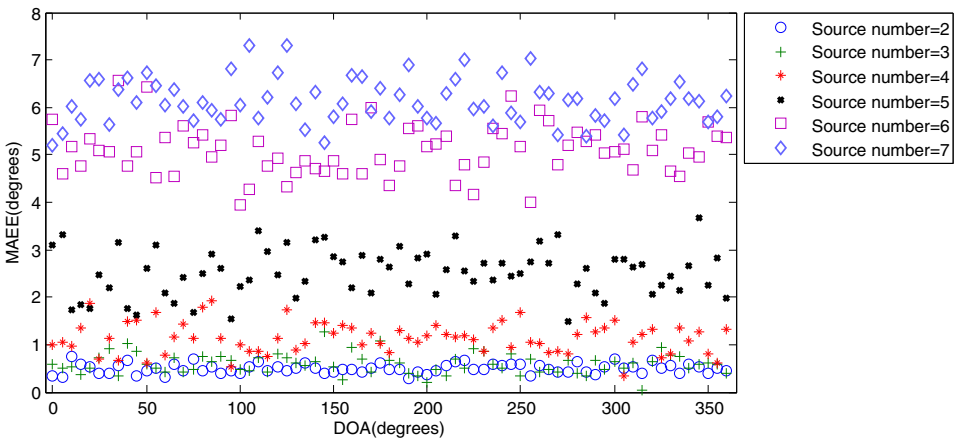

Fig. 11 DOA estimation error versus the true DOA. *Different markers*correspond to different number of speakers: separation $= 30°$

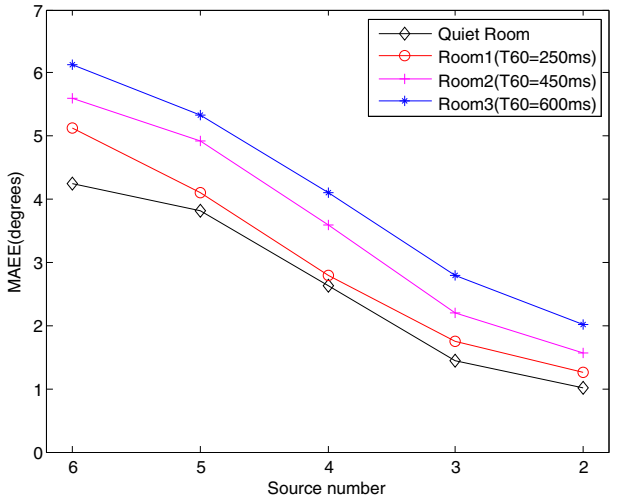

Fig. 12 DOA estimation error vs source number in reverberant environment

positioning sources around the soundfield microphone, especially when source number is less than 4, the *MAEE* will be below 3°. It implies that the proposed method is robust in reverberant environments.

4.2 The evaluation of source counting

Aiming to evaluate the accuracy of source counting, the evaluation was conducted in five different simulated scenarios, i.e., Anechoic room, Quiet room, Room1, Room2, Room3. In each simulated room, we tested a certain number of simultaneously active sources ranging from one to seven. The separation of two arbitrary adjacent sound sources was 50°. We present the results in terms of a confusion matrix in Tables 4, 5, 6, 7 and 8. These matrixes show the confusion percent of source counting in five scenarios respectively, where P is the actual number of speech sources and $\hat{P}$ is the estimated number.

Table 4 Confusion matrix for the proposed source counting method in anechoic room

Percent (%)	$\hat{P}$	1	2	3	4	5	6	7	8
	1	100	0	0	0	0	0	0	0
	2	0	100	0	0	0	0	0	0
	3	0	0	100	0	0	0	0	0
P	4	0	0	2.00	96.6	1.4	0	0	0
	5	0	0	0.1	4.9	91.2	3.80	0	0
	6	0	0.02	0.08	2.90	5.87	89.8	1.33	0
	7	0	0	0.13	0.55	2.20	11.2	82.7	3.22

It could be seen from Tables 4, 5, 6, 7 and 8 that the accurate estimation of source counting is always 100 % in five scenarios when $P \leq 2$, and the confusion will be more significant with the increase of source number especially when $P > 5$. Moreover, the accuracy of source counting will decline slightly in reverberant environment, specifically, the confusion gets more evident when T_{60} is longer. Nevertheless, the accuracy of source counting is still more than 77.43 % in different scenarios.

Subsequently, we compared the proposed method (Pro Method) with the CMA-SCM (it was simulated by Roomsim, the array radius was 0.05 mm and the number of microphones were 8) under the same conditions described above in this subsection, and we made a statistic about the counting accuracy of the two methods. The results are shown in Fig. 13 with 95 % confidence intervals.

As we can see, for the case *source number* ≤ 2, both of the two method achieve an accuracy of 100 %; for the case *source number* > 2, the accuracy of the proposed method is slightly higher than CMA-SCM. It should be mentioned that number of microphone used in the CMA-SCM is eight which is the twice number of that in the proposed method. Besides, the counting accuracy is affected by the reverberation slightly. However, the accuracy of two methods is all more than 85 %.

4.3 Measuring the low-delay performance

The average error between the estimated and true DOA of the sources (Gunel et al. 2008; Shujau et al. 2011; Cobos et al. 2011; Ren and Zou 2012) was usually used to evaluate the DOA estimation algorithms. However, it was useful for non-real evaluation while not suitable for measuring the performance of such low-delay algorithms. Considering this problem, (Zheng 2013) proposed a new evaluation scheme. This method contains two important parameters, one is source detection ratio (SDR) and the other is correct-to-incorrect frame ratio (CIFR), respectively. The former is used to calculate the ratio of the detected number and the actual number of all active sound sources in N frames by:

$$SDR = \frac{\sum_{n=1}^{N} S_d(n)}{\sum_{n=1}^{N} S_t(n)}, \tag{22}$$

where $n(1 \leq n \leq N)$ is the frame index, N is the total number of frames involved in statistics, $s_d(n)$ is the number of correct sources detected for frame n and $s_t(n)$ is the total number of active sources for frame n. SDR can reflect the ability to detect the DOA, while the second parameter, namely, *CIFR* is used to test the error rate of the seized DOA. The *CIFR* can be calculated by:

Table 5 Confusion matrix for the proposed source counting method in Quiet room

Percent (%)	$\hat{P}$								
		1	2	3	4	5	6	7	8
	1	100	0	0	0	0	0	0	0
	2	0	100	0	0	0	0	0	0
	3	0	0	100	0	0	0	0	0
P	4	0	0	2.53	96.0	1.47	0	0	0
	5	0	0	0.10	5.30	90.1	4.50	0	0
	6	0	0.02	0.15	3.12	6.32	88.60	1.79	0
	7	0	0	0.11	0.82	2.67	11.8	81.2	3.40

Table 6 Confusion matrix for the proposed source counting method in Room1

Percent (%)	$\hat{P}$								
		1	2	3	4	5	6	7	8
	1	100	0	0	0	0	0	0	0
	2	0	100	0	0	0	0	0	0
	3	0	0.41	98.20	1.39	0	0	0	0
P	4	0	0.04	2.73	95.5	1.67	0	0	0
	5	0	0	0.54	4.70	89.80	4.87	0	0
	6	0	0.02	0.15	3.22	6.52	88.2	1.89	0
	7	0	0	0.11	0.92	2.98	12.30	80.10	3.59

Table 7 Confusion matrix for the proposed source counting method in Room2

Percent (%)	$\hat{P}$								
		1	2	3	4	5	6	7	8
	1	100	0	0	0	0	0	0	0
	2	0	100	0	0	0	0	0	0
	3	0	2.22	96.90	0.88	0	0	0	0
P	4	0	0.14	3.69	91.5	3.98	0.69	0	0
	5	0	0	1.19	4.04	89.30	5.47	0	0
	6	0	0.02	0.57	3.63	5.71	88.0	2.07	0
	7	0	0	0	1.07	3.19	12.60	79.20	3.94

Table 8 Confusion matrix for the proposed source counting method in Room3

Percent(%)	$\hat{P}$								
		1	2	3	4	5	6	7	8
	1	100	0	0	0	0	0	0	0
	2	0	100	0	0	0	0	0	0
	3	0	2.54	96.20	1.26	0	0	0	0
P	4	0	0.11	4.12	90.1	5.18	0.49	0	0
	5	0	0	1.39	5.32	88.1	5.19	0	0
	6	0	0.02	0.56	3.17	6.11	87.9	2.24	0
	7	0	0	0	1.38	3.49	13.18	77.4	4.55

$$CIFR = \frac{\sum_{n=1}^{N} c(n)}{N - \sum_{n=1}^{N} c(n)}, \tag{23}$$

where $c(n)$ is defined by:

$$c(n) = \begin{cases} 0, & \text{false DOA is found} \\ 1, & \text{false DOA is not found} \end{cases}. \tag{24}$$

The low-delay DOA estimation measurement (Zheng 2013) is given by:

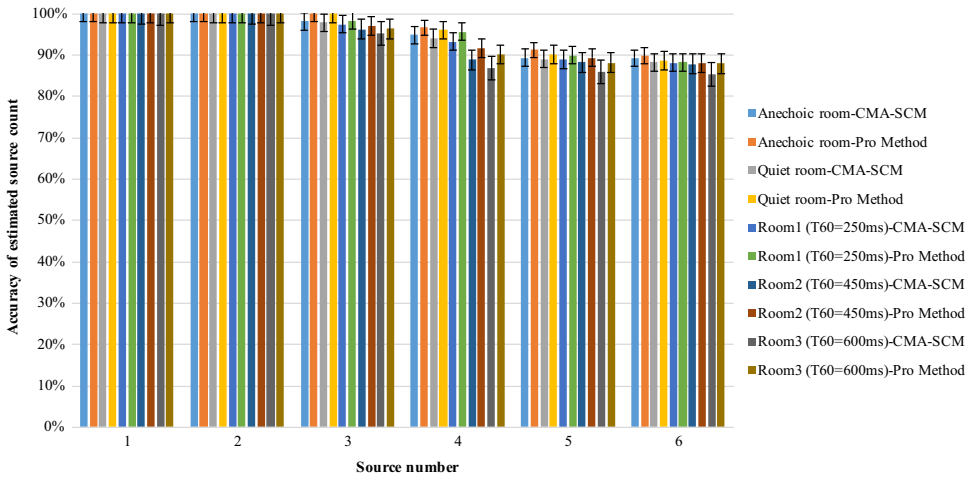

Fig. 13 Objective comparison on accuracy among different source counting algorithms under anechoic, quiet room and reverberant conditions with 95 % confidence intervals

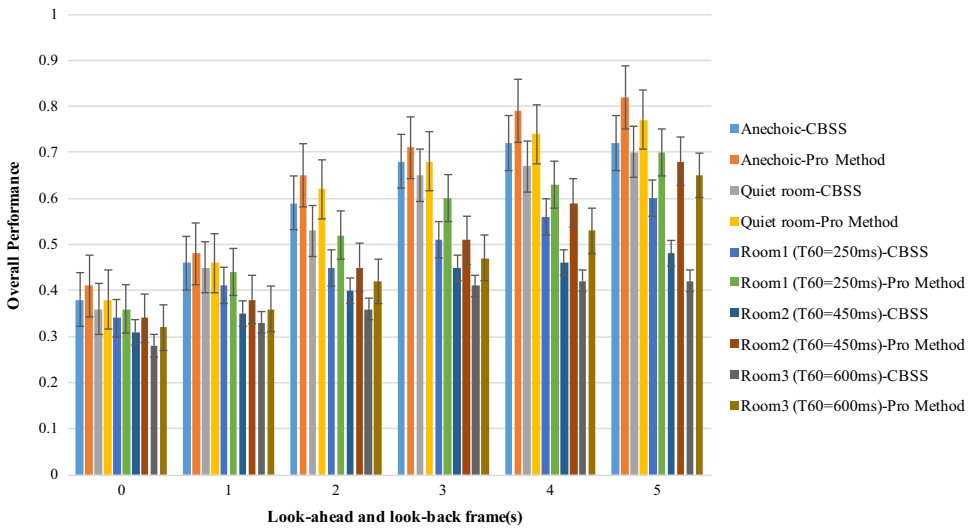

Fig. 14 Objective comparison on low-delay property among different DOA estimation algorithms under anechoic, quiet room and reverberant conditions with 95 % confidence intervals

$$P_{score} = SDR - \frac{SDR}{CIFR}, \tag{25}$$

where *Pscore* is the overall performance of a DOA estimation method. For the ideal situation, $SDR = 1$, and $CIFR = \infty$, thus $Pscore = 1$. So *Pscore* is a number less than 1. In order to make *Pscore* close to 1, we have to ensure that our algorithm has the characteristics of low-delay, in addition, the source counting and DOA estimation must be accuracy.

The objective evaluation was also proceeded under the same conditions described in Sect. 4.2 except for the number of frames which was an important factor in this subsection. We have compared the proposed method with the DOA estimation method in CBSS (Zheng et al. 2013) by using 25, and the result is presented in Fig. 14 with 95 % confidence intervals. It should be noted that the condition CBSS means the DOA estimation method in (Zheng et al. 2013).

From Fig. 14, the following observation can be made:

(a) As expected, the performance score increases along with the increase in number of look-ahead and look-back frames, as more data becomes available to be used in the statistical modeling;
(b) Condition Pro Method achieve a higher score than CBSS for all cases;
(c) The DOA estimation performance degrades as the reverberation time increases which may be affected by the reflected version of sources.

So the proposed method can obtain a more accuracy DOA estimation based on less frames of data. In addition, there are no complex calculation in the proposed method which leads a low computational complexity. Above all, it means that the proposed method has a good performance in low delay and is suitable for real-time applications.

5 Conclusion

In this paper, we proposed a relaxed sparsity assumption among the speech itself as well as the recorded signals and the assumption was proved to be reasonable according to the statistical results. Then, we presented an approach of DOA estimation that imposed the relaxed sparsity constrains on the spatial source signals. A soundfield microphone was utilized to overcome the horizontal restriction and excessive number of microphones in array. The proposed method is based on detecting the "single-source" zones where one source is dominant over the others. Subsequently, in each "single-source" zone, we can obtain the DOA estimation and the source number by using the algorithm described in Sect. 3. Estimating the DOAs of sources in the "single-source" zone is proved to be an efficient process in improving the accuracy of localization. Meanwhile, owning to the unique internal structure of soundfield microphone, the complexity of the method is reduced significantly. Then we performed extensive simulations experiments for various numbers of sources and separations in quiet room and reverberation. The results show that the method has a high accuracy of DOA estimation as well as source counting, and is robust in reverberant environment. Moreover, it is suitable for real-time processing due to its low computational complexity. The results also confirm that the soundfield microphone is suitable for sound source localization and the accuracy of DOA estimation is improved greatly by "single-source" detecting. In future work, we will investigate the performance of the proposed method in various scenarios involving more sources with closer DOAs. We also plan to improve the accuracy of real-time estimation.

Acknowledgments This work has been supported by the National Natural Science Foundation of China (Nos. 61231015, 61201197), Specialized Research Fund for the Doctoral Program of Higher Education of the People's Republic of China (No. 20121103120017), the Project supported by Beijing Postdoctoral Research Foundation.

References

Argentieri S, Danes P(2007) Broadband variations of the music high-resolution method for sound source localization in robotics. In: IEEE/RSJ International Conference on Intelligent Robots and Systems, 2007. IROS 2007. pp 2009–2014

Asaei A, Taghizadeh MJ, Haghighatshoar S, Raj B, Bourlard H, Cevher V (2016) Binary sparse coding of convolutive mixtures for sound localization and separation via spatialization. IEEE Trans Signal Process 64(3):567–579

Bechler D, Kroschel K (2003) Considering the second peak in the gcc function for multi-source tdoa estimation with a microphone array. In: Proceedings of the International Workshop on Acoustic Echo and Noise Control (IWAENC '03), pp 315–318

Bechler D, Schlosser MS, Kroschel K (2004) System for robust 3d speaker tracking using microphone array measurements. In: Proceedings 2004 IEEE/RSJ International Conference on Intelligent Robots and Systems, 2004. (IROS 2004), vol 3, pp 2117–2122

Belloni F, Koivunen V (2003) Unitary root-music technique for uniform circular array. In: Proceedings of the 3rd IEEE International Symposium on Signal Processing and Information Technology, 2003. ISSPIT 2003. pp 451–454

Benesty J, Chen J, Huang Y (2004) Time-delay estimation via linear interpolation and cross correlation. IEEE Trans Speech Audio Process 12(5):509–519

Blandin C, Ozerov A, Vincent E (1950) Multi-source tdoa estimation in reverberant audio using angular spectra and clustering. Signal Process 92(8):1950–1960

Campbell DR, Palomki KJ, Brown GJ (2005) A matlab simulation of "shoebox" room acoustics for use in research and teaching. Comput Inf Syst J 9(3):48–51

Cobos M, Lopez JJ, Martinez D (2011) Two-microphone multi-speaker localization based on a Laplacian mixture model. Digit Signal Process 21(1):66–76

Dmochowski J, Benesty J, Affes S (2007a) Direction of arrival estimation using the parameterized spatial correlation matrix. IEEE Trans Audio Speech Lang Process 15(4):1327–1339

Dmochowski JP, Benesty J, Affes S (2007b) Broadband music: Opportunities and challenges for multiple source localization. In: IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, 2007, pp 18–21

Gunel B, Hacihabiboglu H, Kondoz AM (2008) Acoustic source separation of convolutive mixtures based on intensity vector statistics. IEEE Trans Audio Speech Lang Process 16(4):748–756

Ishi CT, Chatot O, Ishiguro H, Hagita N (2009a) Evaluation of a music-based real-time sound localization of multiple sound sources in real noisy environments. In :IEEE/RSJ International Conference on Intelligent Robots and Systems, 2009. IROS 2009. pp 2027–2032

Ishi CT, Chatot O, Ishiguro H, Hagita N (2009b) Evaluation of a music-based real-time sound localization of multiple sound sources in real noisy environments. In: IEEE/RSJ International Conference on Intelligent Robots and Systems, 2009. IROS 2009. pp 2027–2032

Jia M, Yang Z, Bao C, Zheng X, Ritz C (2015) Encoding multiple audio objects using intra-object sparsity. IEEE/ACM Trans Audio Speech Lang Process 23(6):1082–1095

Karbasi A, Sugiyama A (2007) A new DOA estimation method using a circular microphone array. In: Signal Processing Conference, 2007 15th European, pp 778–782

Knapp C, Carter G (1976) The generalized correlation method for estimation of time delay. IEEE Trans Acoustics Speech Signal Process 24(4):320–327

Loesch B, Uhlich S, Yang B (2009) Multidimensional localization of multiple sound sources using frequency domain ica and an extended state coherence transform. In: IEEE/SP 15th Workshop on Statistical Signal Processing, 2009. SSP '09. pp 677–680

Lombard A, Zheng Y, Buchner H, Kellermann W (2011) Tdoa estimation for multiple sound sources in noisy and reverberant environments using broadband independent component analysis. IEEE Trans Audio Speech Lang Process 19(6):1490–1503

Nakadai K, Matsuura D, Okuno HG, Kitano H (2003) Applying scattering theory to robot audition system: robust sound source localization and extraction. In: Proceedings 2003 IEEE/RSJ International Conference on Intelligent Robots and Systems, 2003. (IROS 2003). vol 2, pp 1147–1152

Nesta F, Omologo M (2012) Generalized state coherence transform for multidimensional tdoa estimation of multiple sources. IEEE Trans Audio Speech Lang Process 20(1):246–260

Comon P, Jutten C (2010) Handbook of blind source separation: independent component analysis and applications. Academic Press, Elsevier, Burlington

Pavlidi D, Griffin A, Puigt M, Mouchtaris A (2013) Real-time multiple sound source localization and counting using a circular microphone array. IEEE Trans Audio Speech Lang Process 21(10):2193–2206

Pavlidi D, Puigt M, Griffin A, Mouchtaris A (2012) Real-time multiple sound source localization using a circular microphone array based on single-source confidence measures. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2012. pp 2625–2628

Pulkki V (2007) Spatial sound reproduction with directional audio coding. J Audio Eng Soc 55(6):503–516

Ren M, Zou YX (2012) A novel multiple sparse source localization using triangular pyramid microphone array. IEEE Signal Process Lett 19(2):83–86

Sawada H, Mukai R, Araki S, Malcino S (2005) Multiple source localization using independent component analysis. In: Antennas and Propagation Society International Symposium, 2005 IEEE, vol 4B, pp 81–84

Schmidt R (1986) Multiple emitter location and signal parameter estimation. IEEE Trans Antennas Propag 34(3):276–280

Shiiki Y, Suyama K (2015) Omnidirectional sound source tracking based on sequential updating histogram. In: 2015 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA), pp 1249–1256

Shujau M, Ritz CH, Burnett IS (2011) Separation of speech sources using an acoustic vector sensor. In: IEEE 13th International Workshop on Multimedia Signal Processing (MMSP), 2011, pp 1–6

Sound C (2015) Core sound TetraMic. http://www.core-sound.com/TetraMic/1.php. Online; Accessed 25 Sep 2015

Su D, Miro JV, Vidal-Calleja T (2015) Real-time sound source localisation for target tracking applications using an asynchronous microphone array. In: IEEE 10th Conference on Industrial Electronics and Applications (ICIEA), 2015, pp 1261–1266

Swartling M, Sllberg B, Grbi N (2011) Source localization for multiple speech sources using low complexity non-parametric source separation and clustering. Signal Process 91(8):1781–1788

Tim VDB, Evelyne C, Jan W (2011) 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

Yi Z, Kuroda T (2014) Wearable sensor-based human activity recognition from environmental background sounds. J Ambient Intell Humaniz Comput 5(1):77–89

Zhang JX, Christensen MG, Dahl J, Jensen SH, Moonen M (2009) Robust implementation of the music algorithm. In: IEEE International Conference on Acoustics, Speech and Signal Processing, 2009. ICASSP 2009., pp 3037–3040

Zheng X (2013) Soundfield navigation: separation, compression and transmission, doctoral dissertation. University of Wollongong, Wollongong

Zheng X, Ritz C, Xi J (2013) Collaborative blind source separation using location informed spatial microphones. IEEE Signal Process Lett 20(1):83–86

Zheng X, Ritz C, Xi J (2016) Encoding and communicating navigable speech soundfields. Multimed Tools Appl 75(9):5183–5204