Applied Acoustics 135 (2018) 85–90

Contents lists available at ScienceDirect

Applied Acoustics

journal homepage: www.elsevier.com/locate/apacoust

Direction of arrival estimation of multiple acoustic sources using a maximum likelihood method in the spherical harmonic domain

Yuxiang Hu[a], Jing Lu[a,*], Xiaojun Qiu[b]

[a] Key Lab of Modern Acoustics, Institute of Acoustics, Nanjing University, Nanjing 210093, China
[b] Centre for Audio, Acoustics and Vibration, Faculty of Engineering and IT, University of Technology Sydney, NSW 2007, Australia

ARTICLE INFO

Keywords:
Multiple source DOA estimation
Maximum likelihood
Spherical harmonic domain
Alternating projection

ABSTRACT

Direction of arrival estimation (DOA) of multiple acoustic sources has been used for a wide range of applications, including room geometry inference, source separation and speech enhancement. The beamformer-based and subspace-based methods are most commonly used for spherical microphone arrays; however, the former suffers from spatial resolution limitations, while the later suffers from performance degradation in noisy environment. This letter proposes a multiple source DOA estimation approach based on the maximum likelihood method in the spherical harmonic domain and implements an efficient sequential iterative search of maxima on the cost function in the spherical harmonic domain. The proposed method avoids the division of the spherical Bessel function, which makes it suitable for both rigid-sphere and open-sphere configurations. Simulation results show that the proposed method has a significant superiority over the commonly used frequency smoothing multiple signal classification method. Experiments in a normal listening room and a reverberation room validate the effectiveness of the proposed method.

1. Introduction

The rotationally symmetric spatial directivity makes the spherical microphone array an appealing structure in many audio applications, among which the acoustic source localization, or the direction of arrival (DOA) estimation, plays an important role in speech enhancement [1], room impulse response analysis [2], and room geometry inference [3].

Various DOA estimation methods have been proposed, which can be generally classified as beamformer-based [2–5] and subspace-based [6–8]. The beamformer-based methods, such as those based on plane-wave decomposition (PWD) [4] and the minimum variance distortionless response (MVDR) beamformer [3], have the benefit of straightforward implementation, but suffer from low spatial resolution. The subspace-based methods, such as the multiple signal classification (MUSIC) [7], provide a high spatial resolution; however, they suffer from severe performance degradation when the signal-to-noise ratio (SNR) is low [6,9]. In order to improve the robustness of the DOA estimation of coherent sources, wideband expansion based on focusing matrices or frequency smoothing (FS) techniques has to be employed [8].

We proposed a maximum likelihood DOA estimation method in the spherical harmonic domain (SHMLE) recently, which is an attractive alternative DOA estimation method with advantages of high spatial resolution, strong robustness and straightforward wideband implementation [10]. The proposed SHMLE method only considered one source situation, while two or more sources often need to be localized in many practical applications. In this letter, the SHMLE method is extended to estimate the DOA of multiple sources. Generally speaking, the DOAs can be determined by searching maxima on the maximum likelihood (ML) cost function. However, the commonly used grid search method is only effective in finding the global maximum, which restricts its applicability in one source situation. To achieve effective DOA estimation of multiple sources, an efficient sequential iterative search method is introduced in the spherical harmonic (SH) domain. Experiments using a 32-element spherical microphone array validate the feasibility and superiority of the proposed method.

2. Method

2.1. Signal model in the spherical harmonic domain

The standard spherical coordinate system is utilized with r, θ and ϕ representing the radius, the elevation angle and the azimuth, respectively. The sound field is assumed to be composed of plane waves from L sources with $\Psi_l = (\theta_l, \phi_l)$ $(l = 1, 2, \ldots, L)$ being the DOA of the l-th plane wave and $s_l(k)$ being its amplitude, where k denotes the wave

* Corresponding author.
E-mail address: lujing@nju.edu.cn (J. Lu).

https://doi.org/10.1016/j.apacoust.2018.02.005
Received 27 September 2017; Received in revised form 17 January 2018; Accepted 7 February 2018
Available online 13 February 2018

number. The Q element spherical microphone array is distributed uniformly on a sphere with a radius of a centred at the origin of the coordinate system, and $\Omega_q = (\theta_q,\ \phi_q)$ is the angle position of the q-th microphone [11].

The sound pressure of the q-th microphone for the incident waves can be expressed as [12]

$$p(k,\Omega_q) = \sum_{l=1}^{L} s_l(k)e^{\mathrm{i}\mathbf{k}_l^{\mathrm{T}}\mathbf{r}_q} \approx \sum_{l=1}^{L} s_l(k) \sum_{n=0}^{N} \sum_{m=-n}^{n} b_n(k)Y_{n,m}^{*}(\Psi_l)Y_{n,m}(\Omega_q), \tag{1}$$

where $\mathbf{k}_l = -k(\cos\phi_l\sin\theta_l,\ \sin\phi_l\sin\theta_l,\ \cos\theta_l)^{\mathrm{T}}$ and $\mathbf{r}_q = a(\cos\phi_q\sin\theta_q,\ \sin\phi_q\sin\theta_q,\ \cos\theta_q)^{\mathrm{T}}$ denote the wave vector of the plane wave and position of the q-th microphone in the Cartesian coordinate. $Y_{n,\ m}$ is the spherical harmonic of order n and degree m, N is the highest order number for the plane wave decomposition and satisfies $(N+1)^2 < Q$. The superscript $(*)$ denotes complex conjugation. $b_n(k)$ is a function of array configuration [11]. Eq. (1) can be expressed in matrix form as

$$p(k,\Omega_q) \approx \mathbf{y}^{T}(\Omega_q)\mathbf{B}(k)\mathbf{Y}^{\mathrm{H}}(\boldsymbol{\Psi})\mathbf{s}(k), \tag{2}$$

with

$$\mathbf{y}(\Omega_q) = [Y_{0,0}(\Omega_q),\ Y_{1,-1}(\Omega_q),\ Y_{1,0}(\Omega_q),\ Y_{1,1}(\Omega_q),\ \ldots,\ Y_{N,N}(\Omega_q)]^{\mathrm{T}}, \tag{3}$$

$$\mathbf{y}(\Psi_l) = [Y_{0,0}(\Psi_l),\ Y_{1,-1}(\Psi_l),\ Y_{1,0}(\Psi_l),\ Y_{1,1}(\Psi_l),\ \ldots,\ Y_{N,N}(\Psi_l)]^{\mathrm{T}}, \tag{4}$$

$$\mathbf{Y}(\boldsymbol{\Psi}) = [\mathbf{y}(\Psi_1), \mathbf{y}(\Psi_2), \ldots, \mathbf{y}(\Psi_L)\ ]^{\mathrm{T}}, \tag{5}$$

$$\mathbf{B}(k) = diag\{b_0(k), b_1(k), b_1(k)\ , b_1(k), \ldots, b_N(k)\}, \tag{6}$$

$$\mathbf{s}(k) = [s_1(k), s_2(k), \ldots, s_L(k)]^{T}, \tag{7}$$

$$\boldsymbol{\Psi} = [\Psi_1, \Psi_2, \ldots, \Psi_L], \tag{8}$$

where the superscript $(^{\mathrm{T}})$ denotes the transpose.

In the presence of additive noise, the sound pressure at all Q microphones can be expressed as

$$\mathbf{p}(k,\boldsymbol{\Omega}) \approx \mathbf{Y}(\boldsymbol{\Omega})\mathbf{B}(k)\mathbf{Y}^{\mathrm{H}}(\boldsymbol{\Psi})\mathbf{s}(k) + \boldsymbol{\nu}(k), \tag{9}$$

where

$$\mathbf{Y}(\boldsymbol{\Omega}) = [\mathbf{y}(\Omega_1), \mathbf{y}(\Omega_2), \ldots, \mathbf{y}(\Omega_Q)]^{T}, \tag{10}$$

$\mathbf{p}(k,\boldsymbol{\Omega}) = [p(k,\Omega_1),\ p(k,\Omega_2),\ \ldots,\ p(k,\Omega_Q)]^{\mathrm{T}}$ is the vector of the sound pressure of Q microphones, and $\boldsymbol{\nu}(k) = [\nu_1(k),\ \nu_2(k),\ \ldots,\ \nu_Q(k)]^{\mathrm{T}}$ is the additive sensor noise added to the system. The noise is assumed to be complex Gaussian, to be uncorrelated with the signal, to have zero mean, and for simplicity, to be spatially white with a covariance matrix $\mathbf{R}_{\boldsymbol{\nu}}(k) = \sigma_\nu^2\mathbf{I}_Q$, where σ_ν^2 is the unknown noise variance and $\mathbf{I}_Q$ is the identity matrix of order $Q \times Q$.

For the uniformly spatial sampling configuration used in this letter, the following orthogonal relation holds (note that $(N+1)^2 \le Q$) [11]

$$\frac{4\pi}{Q}\mathbf{Y}^{H}(\boldsymbol{\Omega})\mathbf{Y}(\boldsymbol{\Omega}) = \mathbf{I}_{(N+1)^2}. \tag{11}$$

The SH transform can be carried out by multiplying both sides of Eq. (9) from the left by $\frac{4\pi}{Q}\mathbf{Y}^{\mathrm{H}}(\boldsymbol{\Omega})$, which yields

$$\mathbf{p_{nm}}(k) \approx \mathbf{B}(k)\mathbf{Y}^{\mathrm{H}}(\boldsymbol{\Psi})\mathbf{s}(k) + \boldsymbol{\nu}_{\mathbf{nm}}(k), \tag{12}$$

where $\mathbf{p_{nm}}(k)$ is a vector containing $(N+1)^2$ SH domain coefficients, i.e.,

$$\mathbf{p_{nm}}(k) = [p_{0,\,0}(k),\ p_{1,\,-1}(k),\ p_{1,\,0}(k),\ p_{1,\,1}(k),\ \ldots,\ p_{N,\,N}(k)]^{\mathrm{T}}. \tag{13}$$

The second term on the right side of Eq. (12) is the noise expressed in the SH domain, i.e. $\boldsymbol{\nu}_{\mathbf{nm}}(k) = \frac{4\pi}{Q}\mathbf{Y}^{\mathrm{H}}(\Omega)\boldsymbol{\nu}(k)$, with the mean

$$E[\boldsymbol{\nu}_{\mathbf{nm}}(k)] = \frac{4\pi}{Q}\mathbf{Y}^{\mathrm{H}}(\boldsymbol{\Omega})E[\boldsymbol{\nu}(k)] = \mathbf{0}, \tag{14}$$

and the covariance matrix

$$\mathbf{R_{nm}}(k) = E\left[\frac{4\pi}{Q}\mathbf{Y}^{\mathrm{H}}(\boldsymbol{\Omega})\boldsymbol{\nu}(k)\boldsymbol{\nu}^{\mathrm{H}}(k)\mathbf{Y}(\boldsymbol{\Omega})\frac{4\pi}{Q}\right] = \frac{4\pi}{Q}\cdot\sigma_\nu^2\mathbf{I}_{(N+1)^2}, \tag{15}$$

where $E(\cdot)$ denotes the statistical expectation. Apparently, the noise model in the SH domain is also zero-mean complex Gaussian.

2.2. Sound source DOA estimation in the spherical harmonic domain

Define $\boldsymbol{\Theta} = [\boldsymbol{\Psi}^{\mathrm{T}},\mathbf{S}^{\mathrm{T}},\sigma_n^2]^{\mathrm{T}}$ as the vector of all the unknown parameters, where $\mathbf{S} = [\mathbf{s}(k_{\min})^{\mathrm{T}},\ldots,\mathbf{s}(k_{\max})^{\mathrm{T}}\ ]^{\mathrm{T}}$ contains the amplitudes of the source signals with k_{min} and k_{max} representing the minimum and maximum wave numbers and satisfying $ka \le N$. Throughout this paper, $\boldsymbol{\Psi}$, s and σ_ν^2 are assumed to be deterministic and unknown, while the observed data $\mathbf{p_{nm}}$ is considered random [13]. The likelihood function of $\mathbf{p_{nm}}$ given $\boldsymbol{\Theta}$ in the SH domain can be expressed as [9,13]

$$f(\mathbf{p_{nm}};\boldsymbol{\Theta}) = \frac{\exp\left\{-\sum_{k=k_{min}}^{k_{max}} [\mathbf{p_{nm}}(k) - \mathbf{V_{nm}}(k,\boldsymbol{\Psi})\mathbf{s}(k)]^{\mathrm{H}}\mathbf{R}_{\mathbf{nm}}^{-1}[\mathbf{p_{nm}}(k) - \mathbf{V_{nm}}(k,\boldsymbol{\Psi})\mathbf{s}(k)]\right\}}{(\pi^{(N+1)^2}|\mathbf{R_{nm}}|)^{k_{max}-k_{min}}}, \tag{16}$$

where $\mathbf{V_{nm}}(k,\Psi) = \mathbf{B}(k)\mathbf{Y}^{\mathrm{H}}(\boldsymbol{\Psi})$ and $|\cdot|$ denotes the matrix determinant. The solution to Eq. (16) is given by [10]

$$\hat{\boldsymbol{\Psi}} = \mathrm{argmin}_{\boldsymbol{\Psi}} \sum_{k=k_{min}}^{k_{max}} \|\mathbf{p_{nm}}(k) - \mathbf{V_{nm}}(k,\boldsymbol{\Psi})\mathbf{V_{nm}}(k,\boldsymbol{\Psi})^{\dagger}\mathbf{p_{nm}}(k)\|^2, \tag{17}$$

where $(\cdot)^{\dagger}$denotes pseudo-inverse operation.

Define the cost function as

$$J(\boldsymbol{\Psi}) = -10\log_{10}\left(\sum_{k=k_{min}}^{k_{max}} \|\mathbf{p_{nm}}(k) - \mathbf{V_{nm}}(k,\boldsymbol{\Psi})\mathbf{V_{nm}}(k,\boldsymbol{\Psi})^{\dagger}\mathbf{p_{nm}}(k)\|^2\right), \tag{18}$$

then the wideband estimator can be described as

$$\hat{\boldsymbol{\Psi}} = \underset{\boldsymbol{\Psi}}{\mathrm{argmax}}\,J(\boldsymbol{\Psi}). \tag{19}$$

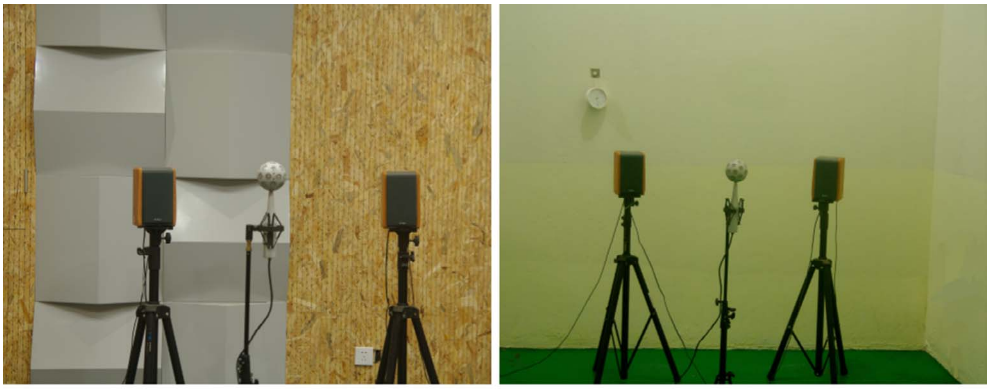
The SHMLE has the remarkable benefit of easy wideband implementation as described in Eqs. (17–19). This is superior over the other methods in the spherical harmonic domain, which usually require a quite cumbersome frequency smoothing (FS) technique to realize wideband DOA [8]. Compared with the maximum likelihood method in Ref. [14], the division of $b_n(k)$ is avoided, which makes the method proposed in this letter suitable for both rigid-sphere and open-sphere arrays. Note that for open-sphere arrays, $b_n(k)$ is close to 0 at frequencies corresponding to the zeros of the spherical Bessel functions.

2.3. DOA estimation of multiple sources

For one source situation, Eq. (19) can be solved using the grid search method. For P grid points and L sources situation, the computational load of Eq. (19) is $\mathrm{O}(P^L)$, which is computationally prohibitive. Moreover, effective discrimination of the multiple maxima in the cost function is very difficult even if repetitive traversal is feasible. To alleviate these problems, a nonlinear optimization algorithm is applied in the SH domain with implementation of the alternating projection method [15]. The alternating projection method avoids the multidimensional search by estimating the location of one source sequentially while fixing the estimates of other source locations from the previous iteration.

For nonlinear optimization methods, the initial locations of the sound sources is critical to reach the global maximum. In this letter, the simplified grid search method is adopted to find initial locations, and the procedure of the method is described as follows.

(1) Estimate the location of the first source s_1 on a single source grid with



(a) (b)

Fig. 1. Eigenmike® and two sound sources in (a) normal listening room (b) reverberation room.

$$\Psi_1^{(0)} = \underset{\Psi_1}{\operatorname{argmax}}\, J(\Psi_1). \tag{20}$$

(2) For $l = 2, \ldots, L$, estimate the location of the lth source s_l, assuming locations of the first $l-1$ sources are fixed by using

$$\Psi_l^{(0)} = \underset{\Psi_2}{\operatorname{argmax}}\, J([\boldsymbol{\Psi}_{l-1}^{(0)}, \Psi_l]), \tag{21}$$

where $\boldsymbol{\Psi}_{l-1}^{(0)}$ denotes the initial locations of the first $l-1$ sources, i.e.

$$\boldsymbol{\Psi}_{l-1}^{(0)} = [\Psi_1^{(0)},\Psi_2^{(0)},\cdots,\Psi_{l-1}^{(0)}]. \tag{22}$$

It should be noted that Step (2) is an iterative process, and the first $l-1$ source locations are fixed for the estimation of the lth source location, as depicted in Eq. (21).

In steps (1) and (2), Eq. (19) only needs to be calculated $P \times L$ times and the effective initial location information can be obtained. In some cases, this initialization process is not necessary since a good initial location estimate is available, for example, from the estimate of the previous data for slowly moving sources.

After initialization, the accurate locations can be estimated using a nonlinear optimization algorithm with implementation of the alternating projection method [15]. The location of Ψ_l at the $(i+1)$ th iteration can be estimated by solving the one-dimensional maximization problem

$$\Psi_l^{(i+1)} = \underset{\Psi_l}{\operatorname{argmax}}\, J([\,\Psi_l,\boldsymbol{\Psi}_{\mathbf{s}}^{(i)},]), \tag{23}$$

where $\boldsymbol{\Psi}_{\mathbf{s}}^{(i)}$ denotes the estimated locations of other $L-1$ sources, i.e.

$$\boldsymbol{\Psi}_{\mathbf{s}}^{(i)} = [\Psi_1^{(i+1)},\cdots,\Psi_{l-1}^{(i+1)},\Psi_{l+1}^{(i)},\cdots,\Psi_L^{(i)}]. \tag{24}$$

In Eq. (23), the Quasi-Newton (QN) method with Broyden-Fletcher-Goldfarb-Shanno algorithm [16] is used, and the QN method is available in MATLAB as in the *fminunc* function. For the beamformer-based and subspace-based methods, the DOA estimation results are acoustic maps [8]. An extra multiple-peak searching procedure is needed to identify the location of the sound sources from the acoustic maps, and the iterative optimization process as described in this manuscript cannot be utilized. On the contrary, the method proposed in this letter can automatically provide the DOA estimation results, as described in Eq. (23). Furthermore, the DOA estimation precision of the proposed method can be extremely high, as will be demonstrated in the following simulations and experiments.

DOA estimation requires knowledge on the number of sound sources L. Numerous methods have been proposed for estimating the number of sources, and the reader can refer to Ref. [17] for an overview. In this paper, we do not focus on the estimation of the sound source number, but assume that L already is known as a prior knowledge.

3. Simulation and experiments

Beamformer-based DOA estimation methods suffer from spatial resolution limitations and are unable to localize spatially adjacent sources [7]. Therefore, in this section, the performance of the proposed method, i.e., the multiple source SHMLE (MS-SHMLE), is investigated and only compared to the FSMUSIC method [2], which has the benefits of high spatial resolution and easy implementation. The Eigenmike® [18] microphone array model, with $Q = 32$ microphones arranged uniformly on a sphere with radius $a = 4.2$ cm (depicted in Fig. 1), was used in both simulations and experiments. Only two source cases were considered in simulations and experiments, and the proposed method can be easily utilized to other scenarios with more sources as described in Section 2.3. The source signals are independent white Gaussian noise sampled at a sampling rate of $f_s = 16$ kHz. The DOA estimation frequency range is $ka \in [2.53.5]$, which avoids both low $b_n(k)$ value and spatial aliasing, leading to robust DOA estimation [8]. The grid resolution of FSMUSIC is 1°.

Root mean squared error (RMSE) is used to assess the performance of the DOA estimation results, which is defined as

$$\mathrm{RMSE} = \sqrt{E\{(\boldsymbol{\Psi}-\hat{\boldsymbol{\Psi}})(\boldsymbol{\Psi}-\hat{\boldsymbol{\Psi}})^T\}}. \tag{25}$$

3.1. DOA estimation performance versus SNR

Fig. 2 depicts the RMSE of the FSMUSIC and the MS-SHMLE as a function of SNR in a room with reverberation time of 0.3 s. In this simulation, the room dimensions are $6 \times 7 \times 5\,\mathrm{m}^3$, the microphone array is located at [3, 2.5, 1.5] m and the speakers are placed 1.0 m away from the array center. Sound sources incident from directions of (90°, 180°) and (90°, 120°). The RMSE is averaged over 100 different trials with sampling length $K = 1024$. The room impulse responses between the sound sources and the microphones positioned on the rigid sphere are simulated using the method proposed in Ref. [19]. It can be seen that the RMSE of the MS-SHMLE is better than that of FSMUSIC especially for the low SNR situations, and FSMUSIC fails to present meaningful DOAs when the SNR is lower than 2 dB.

Note that the RMSE difference between the SHMLE and the FSMUSIC is larger than 1° in low SNR conditions. Subspace-based DOA estimation methods are based on the property that any vector lying in the signal subspace is orthogonal to eigenvectors in the noise subspace [9]. However, in low SNR situations, the estimated noise subspace has considerable bias, which undermines the orthogonality between the signal subspace and the estimated subspace. Therefore, the performance of these methods rapidly breaks down when the SNR falls below a certain threshold.

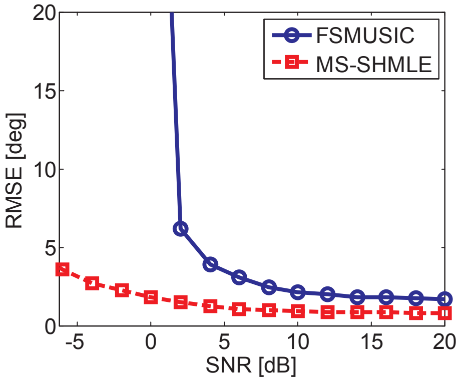


Fig. 2. DOA estimation RMSE of the FSMUSIC and the MS-SHMLE versus SNR in a room with reverberation time of 0.3 s.

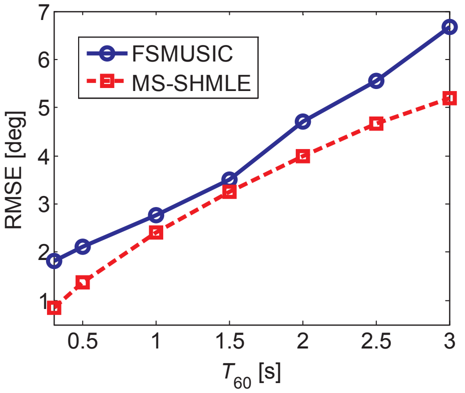


Fig. 3. DOA estimation RMSE of the FSMUSIC and the MS-SHMLE versus reverberation time with an SNR of 15 dB.

3.2. DOA estimation performance versus reverberation time

When the reverberation time is higher than 1.5 s, the method proposed in Ref. [19] is not suitable to simulate the room impulse responses because of its high computational burden and memory requirement. Therefore, an open-sphere array configuration is used in this simulation and the room impulse responses are simulated using the method proposed in Ref. [20]. When the open-sphere configuration is used, the FSMUSIC method suffers from ill-conditioning around the zeros of the spherical Bessel function, and a mitigation method proposed in Ref. [21] is utilized.

Fig. 3 depicts the DOA estimation RMSE of the FSMUSIC and the MS-SHMLE as a function of the reverberation time T_{60}. In this simulation, the SNR is fixed at 15 dB and the sampling length is 1024. It can be seen that the RMSE of the MS-SHMLE is better than that of the FSMUSIC under different reverberation time. Note that the RMSE difference between the SHMLE and the FSMUSIC is larger than 1° when the reverberation time is extremely high.

3.3. DOA estimation performance versus sampling length

Fig. 4 depicts the DOA estimation RMSE of the FSMUSIC and the MS-SHMLE as a function of the sampling length K when the reverberation time is 0.3 s and the SNR is 15 dB. It can be seen that the RMSE of both methods becomes lower with increase of the sampling length. When K is larger than 1024, the performance does not vary significantly. Therefore, in the following sections, the sampling length K is fixed at 1024. When K is smaller than 256, the MS-SHMLE method still works well while the performance of the FSMUSIC method deteriorates drastically.

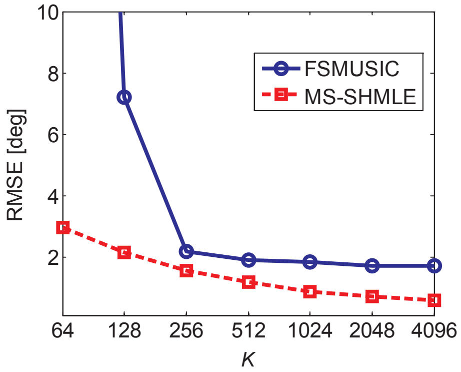


Fig. 4. DOA estimation RMSE of the FSMUSIC and the MS-SHMLE with different sampling length when reverberation time is 0.3 s and the SNR is 15 dB.

3.4. Two-source experiments in a listening room

The experiments for DOA estimation of two sources were carried out in a listening room with background noise less than 30 dBA as depicted in Fig. 1(a). The room dimensions are 5 × 8 × 4 m^3 and the reverberation time is around 0.3 s. The microphone array was located at [2.5, 3, 1.5] m. Two sound sources were placed 1.5 m away from the array with 7 different angle differences $\Delta\phi$. The sound pressure level (SPL) difference of the two sources at the array center is around 2 dB.

Fig. 5 depicts the DOA estimation results for two sources case using the FSMUSIC methods. The DOA of the sound source is denoted by a solid black circle in all these figures. It can be found that FSMUSIC can distinguish both sources when the angle difference between the two sources is larger than 20°, as depicted in Fig. 5(a) and (b). When the separation angle between sources is close to 20°, FSMUSIC fails to identify two sources, as depicted in Fig. 5(c), because the peak corresponding to the weaker source is totally merged into the peak corresponding to the stronger one. It should be noted that $\mathbf{V_{nm}}(k,\Psi)$ in Eq. (17) contains the steering vector of all sources. Therefore, the acoustic maps depicted in Fig. 5 are not suitable for the MS-SHMLE method.

Fig. 6 depicts the convergence curve of the SHMLE cost function, i.e. Eq. (18), using the same data as those used in Fig. 5. In the first and second iterations, the DOA of s_1 and s_2 are estimated using a rough grid search method, and the cost function are given by Eqs. (20) and (21). After the first two iterations, the DOA are estimated using the nonlinear optimization method and the cost function are calculated using Eq. (23). It can be seen that $J(\boldsymbol{\Psi})$ converges quickly to a stable maximum with only 3 or 4 iterations.

Table 1 shows the RMSE of the FSMUSIC and the MS-SHMLE for a 10 s recorded data. Although the RMSE of these two methods are close, the FSMUSIC can only locate the stronger one when the separation angle between the two sources is close to 20° although 1° seems to be a significantly high resolution, while the MS-SHMLE can distinguish both sources. The average RMSE of the MS-SHMLE is comparatively lower than FSMUSIC, which is in consistence with the simulations results.

3.5. Two-source experiments in a reverberation room

To further validate the robustness of the proposed algorithm in high reverberant environments, the experiments for DOA estimation of two sources were also carried out in a reverberation room as depicted in Fig. 1(b). The room dimensions are 5.9 × 7.35 × 5.22 m^3. The reverberation time is around 3 s at frequency range $ka \in [2.5\ 3.5]$. In the experiments, the microphone array was located at [3 2.5 1.5] m, and the sound sources were placed 1.5 m away from the array.

Fig. 7 depicts the DOA estimation results for the two sources case using the FSMUSIC method. Similar to the results in the listening room, when the angle difference between the two sources is larger than 20°, FSMUSIC can distinguish both sources as depicted in Fig. 7(a) and (b).

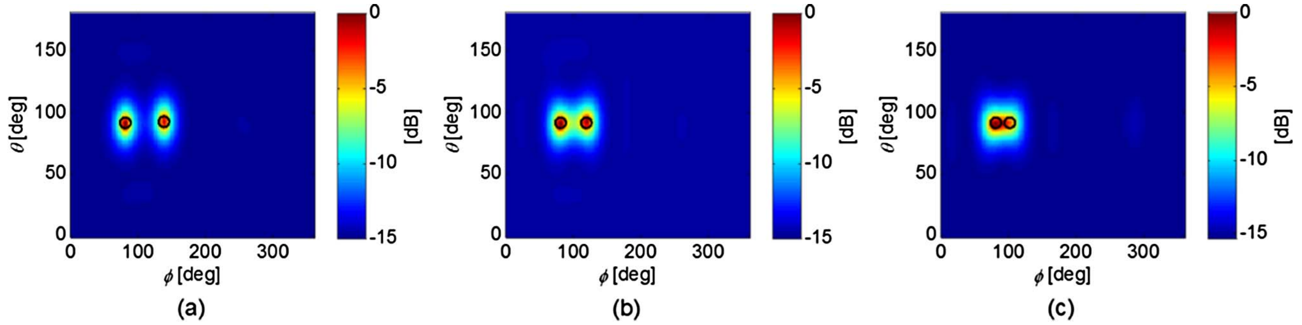


Fig. 5. DOA estimation results for two sources case using FSMUSIC method in a listening room with (a) $\Delta\phi = 60^\circ$, (b) $\Delta\phi = 40^\circ$ and (c) $\Delta\phi = 20^\circ$.

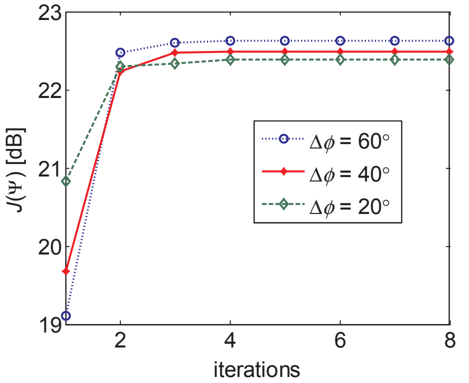


Fig. 6. Convergence curve of the SHMLE cost function using the same data as those used in Fig. 5.

Table 1
RMSE of the FSMUSIC and the MS-SHMLE for two sources case in a listening room.

Angle difference	RMSE of the FSMUSIC			RMSE of the SHMLE		
	Strong	Weak	Total	Strong	Weak	Total
180°	0.73°	1.19°	0.99°	0.32°	1.66°	1.20°
120°	0.59°	0.85°	0.73°	0.32°	0.91°	0.68°
90°	0.76°	0.89°	0.83°	0.42°	0.90°	0.70°
60°	0.77°	0.71°	0.74°	0.32°	0.79°	0.60°
40°	0.83°	0.78°	0.81°	0.34°	1.05°	0.78°
30°	1.08°	1.65°	1.39°	0.44°	1.76°	1.28°
20°	1.66°	–	–	0.63°	1.54°	1.18°

When the separation angle between sources is close to 20°, FSMUSIC can only locate the stronger source while fail to identify the weaker one although 1° seems to be a significantly high resolution, as depicted in Fig. 7(c).

Fig. 8 depicts the convergence curve of the SHMLE cost function using the same data as those used in Fig. 7. Similar to the experimental results in the listening room, $J(\boldsymbol{\Psi})$ converges to a stable maximum after 3 or 4 iterations. Compared with Fig. 6, the converged $J(\boldsymbol{\Psi})$ in Fig. 8 is much smaller, which indicates that the DOA estimation results in the listening room is better than that in the reverberant room.

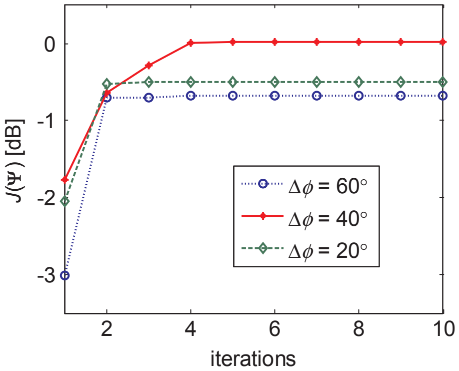


Fig. 8. Convergence curve of the SHMLE cost function using the same data as those used in Fig. 7.

Table 2 shows the RMSE of the FSMUSIC and the MS-SHMLE for a 10 s recorded data. It can be seen that the DOA estimation RMSE in the reverberation room is higher than that in the listening room. The RMSE of these two methods increase significantly when the angle difference between the two sources is close to or lower than 30°. When the separation angle between the two sources is close to 20°, the FSMUSIC can only locate the stronger one, while the MS-SHMLE can distinguish both sources. The superiority of the MS-SHMLE in high reverberant environment coincides well with the simulations presented in Section 3.2.

White Gaussian noise is used in all simulations and experiments, which has the benefits of flat spectrum. However, common sources, such as speech and music, also need to be localized in practical situations. In order to locate the direction of the common sources, voice activity detection is needed to select the effective signals. Furthermore, speech or music in a single frame may not contain enough information

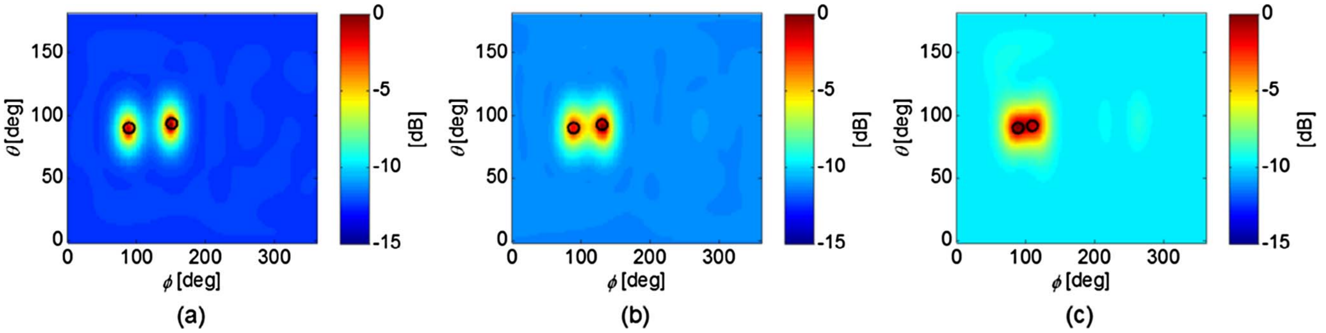


Fig. 7. DOA estimation results for two sources case using FSMUSIC method in a reverberation room with (a) $\Delta\phi = 60^\circ$, (b) $\Delta\phi = 40^\circ$ and (c) $\Delta\phi = 20^\circ$.

Table 2
RMSE of the FSMUSIC and the MS-SHMLE for two sources case in a reverberation room.

Angle difference	RMSE of the FSMUSIC			RMSE of the SHMLE		
	Strong	Weak	Total	Strong	Weak	Total
180°	1.38°	1.38°	1.38°	1.42°	1.22°	1.32°
120°	1.25°	1.17°	1.21°	1.30°	1.19°	1.25°
90°	1.49°	1.17°	1.34°	1.40°	1.07°	1.24°
60°	1.54°	1.20°	1.38°	1.29°	0.98°	1.15°
40°	1.53°	1.34°	1.44°	1.73°	1.42°	1.59°
30°	4.22°	3.46°	3.86°	1.99°	2.02°	2.00°
20°	4.57°	–	–	3.56°	4.02°	3.80°

in the DOA estimation frequency range. Therefore, a time-domain smoothing technique is needed to improve the DOA estimation robustness [14].

4. Conclusion

This letter proposes a multiple source DOA estimation method in the spherical harmonic domain using the maximum likelihood strategy. To avoid high-dimensional grid search with extremely high computational burden, a nonlinear optimization algorithm with implementation of the alternating projection method is introduced, leading to an efficient MS-SHMLE method. The proposed method avoids the division of the spherical Bessel function, which makes it suitable for both rigid-sphere and open-sphere configurations. Simulations and experiments on a 32-microphone model demonstrate that the proposed MS-SHMLE method has very good spatial resolution and can distinguish two sources with 20° angle difference in both normal listening room and reverberation room. Furthermore, the performance is stable in low SNR environment, circumventing the problem faced by the subspace-based method.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant Nos. 11374156 and 11474163) and the Fundamental Research Funds for the Central Universities.

References

[1] Kumatani K, McDonough J, Raj B. Microphone array processing for distant speech recognition: from close-talking microphones to far-field sensors. IEEE Signal Process Mag 2012;29(6):127–40.

[2] Khaykin D, Rafaely B. Acoustic analysis by spherical microphone array processing of room impulse responses. J Acoust Soc Am 2012;132(1):261–70.

[3] Mabande E, Kowalczyk K, Sun H, Kellermann W. Room geometry inference based on spherical microphone array eigenbeam processing. J Acoust Soc Am 2013;134(4):2773–89.

[4] Park M, Rafaely B. Sound-field analysis by plane-wave decomposition using spherical microphone array. J Acoust Soc Am 2005;118(5):3094–103.

[5] Torres AM, Cobos M, Pueo B, Lopez JJ. Robust acoustic source localization based on modal beamforming and time–frequency processing using circular microphone arrays. J Acoust Soc Am 2012;132(3):1511–20.

[6] Li X, Yan S, Ma X, Hou C. Spherical harmonics MUSIC versus conventional MUSIC. Appl Acoust 2011;72(9):646–52.

[7] Nadiri O, Rafaely B. Localization of multiple speakers under high reverberation using a spherical microphone array and the direct-path dominance test. IEEE/ACM Trans Audio Speech Lang Process 2014;22(10):1494–505.

[8] Sun H, Mabande E, Kowalczyk K, Kellermann W. Localization of distinct reflections in rooms using spherical microphone array eigenbeam processing. J Acoust Soc Am 2012;131(4):2828–40.

[9] Mestre X, Lagunas MÁ. Modified subspace algorithms for DoA estimation with large arrays. IEEE Trans Signal Process 2008;56(2):598–614.

[10] Hu Y, Lu J, Qiu X. A maximum likelihood direction of arrival estimation method for open-sphere microphone arrays in the spherical harmonic domain. J Acoust Soc Am 2015;138(2):791–4.

[11] Rafaely B. Analysis and design of spherical microphone arrays. IEEE Trans Speech Audio Process 2005;13(1):135–43.

[12] Rafaely B. Fundamentals of spherical array processing. Berlin: Springer; 2015. p. 57–99.

[13] Chen CE, Lorenzelli F, Hudson RE, Yao K. Maximum likelihood DOA estimation of multiple wideband sources in the presence of nonuniform sensor noise. EURASIP J Adv Signal Process 2007;2008(1):1–12.

[14] Tervo S, Politis A. Direction of arrival estimation of reflections from room impulse responses using a spherical microphone array. IEEE/ACM Trans Audio Speech Lang Process 2015;23(10):1539–51.

[15] Chen JC, Hudson RE, Yao K. Maximum-likelihood source localization and unknown sensor location estimation for wideband signals in the near-field. IEEE Trans Signal Process 2002;50(8):1843–54.

[16] Dennis Jr JE, Moré JJ. Quasi-Newton methods, motivation and theory. SIAM Rev 1977;19(1):46–89.

[17] Krim H, Viberg M. Two decades of array signal processing research: the parametric approach. IEEE Signal Process Mag 1996;13(4):67–94.

[18] Meyer J, Elko G. A highly scalable spherical microphone array based on an orthonormal decomposition of the soundfield. In Proceedings of the IEEE international conference on acoustics, speech, and signal processing (ICASSP), Orlando, FL, USA, May 2002; p. 1781–1781.

[19] Jarrett DP, Habets EAP, Thomas MRP, Naylor PA. Rigid sphere room impulse response simulation: algorithm and applications. J Acoust Soc Am 2012;132(3):1462–72.

[20] Allen JB, Berkley DA. Image method for efficiently simulating small-room acoustics. J Acoust Soc Am 1979;65(4):943–50.

[21] Rafaely B. Bessel nulls recovery in spherical microphone arrays for time-limited signals. IEEE Trans Audio Speech Lang Process 2011;19(8):2430–8.