doi:10.11959/j.issn.1000-436x.2017013

基于近似核密度估计的近场多声源定位算法

房玉琢，许志勇，赵兆
（南京理工大学电子工程与光电技术学院，江苏 南京 210094）

摘 要：针对混响环境下的近场多声源定位问题，提出了一种基于近似核密度估计（KDE）的算法模型。引入多阶段（MS）分频带处理有效解决宽间距时的空域模糊，同时，构建空域似然率函数（SLF）通过相加（S）及相乘（P）2 种算子进行多维融合，从而衍生出 S-KDE、P-KDE、S-KDEMS 和 P-KDEMS 4 种算法。通过对均方根误差（*RMSE*）以及表征辨识度的 SLF 百分比（*PSLF*）这 2 个统计指标的综合比较，证实了 P-KDEMS 是一种具有较高稳健性与辨识度的近场多声源定位算法。
关键词：麦克风阵列；近似核密度估计；多阶段分频带处理；空域似然率函数；数据融合
中图分类号：TN911.72 **文献标识码**：A

Near-field localization algorithm of multiple sound sources based on approximated kernel density estimator

FANG Yu-zhuo, XU Zhi-yong, ZHAO Zhao
(School of Electronic and Optical Engineering, Nanjing University of Science and Technology, Nanjing 210094, China)

Abstract: For near-field localization of multiple sound sources in reverberant environments, a algorithm model based on approximated kernel density estimator (KDE) was proposed. Multi-stage (MS) of sub-band processing was introduced to effectively solve the spatial aliasing by wide spacing. Spatial likelihood function (SLF) was built for multi-dimensional fusion by using two operators, sum (S) and prod (P). Then four algorithms, S-KDE, P-KDE, S-KDEMS, P-KDEMS, were derived. By the comprehensive comparison of the two statistical indicators root mean square error (*RMSE*) and percentage of SLF (*PSLF*) which denoted the recognition, P-KDEMS is confirmed as a near-field localization algorithm of multiple sound sources with high robustness and recognition.
Key words: microphone array, approximated kernel density estimator, multi-stage of sub-band processing, spatial likelihood function, data fusion

1 引言

随着人类生活水平的提高，大批新技术不断涌现。触摸屏的普及使传统的键盘输入方式逐步被语音声控取代；虚拟现实（virtual reality）、智能房间、无人驾驶等新技术的发展对于人机的交互性有了更高的需求。作为其中的核心技术，一套强大的声控系统首先需要判定声源的具体位置，从而为后续的机器语音增强与抗干扰识别提供准确的前端处理结果[1]，因此，基于麦克风阵列的近场声源定位显得尤为重要。在实际应用中，其主要面临将回声和频谱衰减引入观察信号的混响效应，以及宽间距时多个声源间的高频混叠这 2 个问题的综合影响。

传统的麦克风阵列声源定位方法包含 2 类。1) 基于时延估计（TDE, time delay estimation）的两步定位算法，首先由经典的相位变换（PHAT, phase transform）广义互相关法[2]求取阵列中各对麦克风的相对时延；再由线性交叉（LI, linear intersection）

收稿日期：2016-06-05；**修回日期**：2016-11-30
通信作者：许志勇，ezyxu@mail.njust.edu.cn
基金项目：国家自然科学基金资助项目（No.61171167，No.61401203）；江苏省自然科学基金资助项目（No.BK20130776）
Foundation Items: The National Natural Science Foundation of China (No.61171167, No.61401203), The Natural Science Foundation of Jiangsu Province (No.BK20130776)

闭式解[3]、最小均方误差（MMSE, minimal mean square error）搜索[4]求取声源位置，该类方法虽然计算量较小，但依托于第 1 步 TDE 的精度，单对或多对麦克风的估计偏差将对第 2 步的定位结果产生影响。2) 波束形成（BF, beam-forming）算法，计算空间中各点波束能量[5]，由峰值搜索直接估计出声源的位置，该类方法需要较大的运算量来获得理想的定位结果。

综合上述 2 类方法的优点，文献[6]将波束能量的求取近似转化为麦克风对 PHAT 函数的累加，从而提出了相位加权控制响应功率（SRP-PHAT, steered-response power PHAT）算法，该方法的性能较上述第 1 类方法稳健，同时计算量比第 2 类小。文献[7]引入空域似然率函数（SLF, spatial likelihood function）的概念，通过相加（S, sum）的融合方式，从 SLF 融合的角度提出了 SRP-PHAT，同时，由另一种相乘（P, prod）的融合方式提出了辨识度更高的 Multi-PHAT。文献[6,7]中的方法基于理想传播模型下的 PHAT 时延估计函数，混响增强时，性能下降明显[8]，且在满足一定时延分辨力要求的较宽间距下，由于高频混叠的增强，伪峰幅度明显变高[9]。要获得实用的定位性能，这种 PHAT 类算法需要包含较多数量麦克风的大型阵列，然而现实中许多声控系统由于应用空间的局限性，要求麦克风阵列不能过于复杂。

近似核密度估计（KDE, kernel density estimator）通过引入非线性核密度函数使其相较于 PHAT 具有更加良好的抗混响性能[10,11]，文献[10]中核密度谱的计算采用了独立变量分析（ICA, independent component analysis），复杂度较高且需要考虑分离矩阵的收敛问题；而文献[11]运用多声源时频稀疏假设[12]，由归一化互功率谱（NCS, normalized cross-power spectrum）计算求得，在简化计算量的同时获得了与文献[10]差别不大的估计性能[11]，但实际上，该假设并不能够严格满足[10]。上述 KDE 方法并没有重点考虑较宽的阵元间距，此时非线性函数中的频率加权因子不足以克服高频混叠所带来的谱模糊问题。因此，文献[13]引入多阶段（MS, multi-stage）环节对 KDE 进行分频带处理，提出的 KDEMS 较好地克服了宽间距时的模糊问题，然而该方法使用文献[10]中的 ICA，且仅考虑了单对麦克风，计算量及阵元数限制了其应用范围。

本文所提方法基于 KDE，主要贡献有以下 3 个方面：1) 引入相关性检测（CT, coherence test）[14]提取出单个声源能量占优的时频支撑域，确保多声源时频稀疏假设成立，进而由 NCS 计算 KDE，并通过多阶段 MS 分频带处理[13]对 KDE 做出改进，削减乃至消除高频混叠带来的多声源空域模糊问题；2) 由 S 或 P 2 种基本算子，借助 SLF 将 KDE 及改进的 KDEMS 进行多维（多对麦克风）融合，所得结果表征空间中声源出现在各点的似然率，从而在 KDE 类函数的基础上建立了一套近场多声源定位算法模型；3) 在传统均方根误差（RMSE, root mean square error）的基础上，提出了表征定位辨识度的 SLF 百分比（PSLF, percentage of SLF）指标，从稳健性和辨识度 2 个方面综合比较该模型中各种算法的性能。

2　常规 KDE 算法原理

假设$\{s_n(i),\ n = 1,2,\cdots,N\}$表示 N 个声源所产生的声源信号，$\{x_m(i), m = 1,2,\cdots,M\}$表示 M 个阵元所组成的麦克风阵列观测到的混合信号，采样率为 f_s。则离散时域传播模型如式(1)所示。

$$x_m(i) = \sum_{n=1}^{N} \boldsymbol{h}_{m,n}^{\mathrm{T}} \boldsymbol{s}_n(i) + w_m(i) \tag{1}$$

其中，$\boldsymbol{h}_{m,n} = [h_{m,n}(0),\cdots,h_{m,n}(N_h-1)]^{\mathrm{T}}$ 表示第 n 个声源到第 m 个麦克风之间长度为 N_h 的冲激响应，$\boldsymbol{s}_n(i) = [s_n(i),\cdots,s_n(i-N_h+1)]^{\mathrm{T}}$ 表示第 n 个声源的信号矢量，$w_m(i)$表示与声源信号及冲激响应不相关的第 m 个麦克风的加性高斯白噪声。

通过 N_{FFT} 点短时傅里叶变换（STFT, short-time Fourier transform）将式(1)转换到离散时频域，可得

$$\begin{aligned}\boldsymbol{X}(r,k) &= \sum_{n=1}^{N} \boldsymbol{H}_n(k) S_n(r,k) + \boldsymbol{W}(r,k) \\ &= \left[X_1(r,k), X_2(r,k), \cdots, X_M(r,k)\right]^{\mathrm{T}}\end{aligned} \tag{2}$$

其中，$X_m(r,k)$、$S_n(r,k)$分别表示与观察信号 $x_m(i)$、声源信号 $s_n(i)$相对应的第 k 个离散频率、第 r 帧处的 STFT 系数；$\boldsymbol{W}(r,k) \in C^{M\times 1}$ 表示加性复噪声矢量；$\boldsymbol{H}_n(k) = [H_{1,n}(k),\cdots,H_{m,n}(k),\cdots,H_{M,n}(k)]^{\mathrm{T}}$，$H_{m,n}(k)$表示第 n 个声源到第 m 个麦克风的传递函数，在空间扩散噪声的假设下[15]，可以分解为直达波 $H_{m,n}^{\mathrm{DIR}}(k)$ 以及混响 $H_{m,n}^{\mathrm{REV}}(k)$ 2 个部分，即

$$\begin{aligned}H_{m,n}(k) &= H_{m,n}^{\mathrm{DIR}}(k) + H_{m,n}^{\mathrm{REV}}(k) \\ &= \left|H_{m,n}^{\mathrm{DIR}}(k)\right| \exp(-\mathrm{j}2\pi f_k T_{m,n}) + H_{m,n}^{\mathrm{REV}}(k)\end{aligned} \tag{3}$$

其中，$T_{m,n}$ 表示信号由第 n 个声源到达第 m 个麦克风的传播时间，$f_k=\frac{kf_s}{N_{\mathrm{FFT}}}$ 表示第 k 个离散频率值。

为了问题讨论的直观性，暂不考虑混响及噪声的影响，由多声源信号的时频稀疏假设[12]，每个时频支撑域$(r,\ k)$内最多只有一个声源能量占主导地位，此时，由式(2)、式(3)可得信号的传播模型

$$\boldsymbol{X}(r,k)=\begin{bmatrix}\left|H_{1,ND(r,k)}(k)\right|\exp(-\mathrm{j}2\pi f_k T_{1,ND(r,k)})\\ \vdots\\ \left|H_{M,ND(r,k)}(k)\right|\exp(-\mathrm{j}2\pi f_k T_{M,ND(r,k)})\end{bmatrix}\cdot S_{ND(r,k)}(r,k) \tag{4}$$

其中，$ND(r,k)\in\{1,\cdots,N\}$ 表示时频支撑域$(r,\ k)$处占主导地位的声源序号。考虑任意一对麦克风阵元$(a,\ b)$，在$(r,\ k)$处接收信号的归一化互功率谱 NCS 可写为[11]

$$NCS(r,k)=\frac{X_a(r,k)X_b^*(r,k)}{\left|X_a(r,k)X_b^*(r,k)\right|}=\exp(-\mathrm{j}2\pi f_k\tau_{(a,b)}^{ND(r,k)}) \tag{5}$$

其中，$(\cdot)^*$ 表示复共轭运算，$\tau_{(a,b)}^{ND(r,k)}=T_{a,ND(r,k)}-T_{b,ND(r,k)}$ 表示$(r,\ k)$处主导声源 $ND(r,\ k)$到麦克风对$(a,\ b)$的直达波到达时间差(后文简称时延)，为了估计上述时延，将其看作一种满足给定概率分布的随机变量，用 τ 表示，使用式(6)中的非线性变换

$$g(r,k,\tau)=\frac{1}{2\pi f_k}\exp\left(-\frac{\left|\exp(-\mathrm{j}2\pi f_k\tau)-NCS(r,k)\right|^2}{2B_d^2}\right) \tag{6}$$

由各个时频支撑域$(r,\ k)$处的$\exp(-\mathrm{j}2\pi f_k\tau)$与式(5)中 $NCS(r,\ k)$的绝对误差值构建近似高斯核密度函数[10,11]，$B_d=\frac{\tau_{\max}}{B}$ 为该核函数的带宽，其中，$\tau_{\max}=\frac{d}{c}$ 为最大可能的时延(d 为阵元间距，c 为大气声传播速度)，B 为影响时延域分辨力的因子。

常规 KDE 算法将式(6)中所得函数在离散时频域进行累加并求平均，得到关于 τ 的近似核密度谱函数[10]

$$\varphi_{\mathrm{KDE}}(\tau)=\frac{1}{N_rN_k}\sum_{(r,k)}g(r,k,\tau) \tag{7}$$

其中，N_k 和 N_r 分别表示累加所使用的离散频率及信号帧的数量。由于混响被视作空间散布的噪声，通过式(6)中近似核函数的非线性特性及式(7)的累加平均可以弱化混响对估计结果的影响，上述 KDE 算法能够提供较好的抗混响性能[10,11]。

3 基于 KDE 的近场多声源定位算法

3.1 相关性检测

式(4)的推导包含了声源的时频稀疏性假设，文献[12]通过大量实验证实了该假设的可靠性。然而实际中，对每个时频支撑域$(r,\ k)$，单个声源能量占优的假设并不能够严格满足，为了削弱其对最终定位结果的影响，本文引入如下 CT 环节[14]去除该假设条件的限制。阵元 a、b 间第$(r,\ k)$个时频支撑域的相关性参数可表示为

$$COH(r,k)=\left|\frac{\mathrm{E}(X_a(r,k)X_b^*(r,k))}{\sqrt{\mathrm{E}(X_a(r,k)X_a^*(r,k))}\sqrt{\mathrm{E}(X_b(r,k)X_b^*(r,k))}}\right| \tag{8}$$

其中，$\mathrm{E}(\cdot)$ 表示 $2C+1$ 个连续时间帧的近似平均数学期望，即

$$\mathrm{E}(X_{id1}(r,k)X_{id2}^*(r,k))=\frac{1}{2C+1}\sum_{r'=r-C}^{r+C}X_{id1}(r',k)X_{id2}^*(r',k),\ id1,id2\in\{a,b\} \tag{9}$$

根据实际应用环境设置经验门限 Thd，若该支撑域的 COH 高于给定阈值，则被认为仅包含一个主导声源，具体判决方法如式(10)所示。

$$W_{COH}(r,k)=\begin{cases}1, & COH(r,k)>Thd\\ 0, & \text{其他}\end{cases} \tag{10}$$

由相关性权值系数 $W_{COH}(r,k)$，可以判断并提取出单个声源能量占优的时频支撑域。

3.2 MS 分频带处理

式(6)计算近似核密度函数时，为了获得较高的 TDE 分辨力，需使用较宽的阵元间距 d，然而 d 的变宽势必会打破最小信号波长 $\lambda_{\min}$ 的限制。此时由空域奈奎斯特采样定理，超过最大不模糊频率（f_{UA}）的密度谱会产生混叠，对 KDE 的估计性能产生影响，式(6)中与频率相关的 $\frac{1}{2\pi f_k}$ 加权因子一定程度上能够抑制高频混叠，但并不能够完全消除这种影响[13]。

$$f_{\mathrm{UA}}=\frac{c}{2d} \tag{11}$$

假设所考虑观察信号的频率区间为$[f_{\mathrm{L}},\ f_{\mathrm{H}}]$，其

中，f_{L} 和 f_{H} 分别表示最低及最高频率，则 f_{UA} 将整个频率区间划分为$[f_{\mathrm{L}}, f_{\mathrm{UA}}]$和$[f_{\mathrm{UA}}, f_{\mathrm{H}}]$2 个子频带，其中第 1 个子频带并不存在高频混叠的影响，当 $f_{\mathrm{H}}>2f_{\mathrm{UA}}$ 时，第 2 个子频带又可分为$[f_{\mathrm{UA}}, 2f_{\mathrm{UA}}]$和$[2f_{\mathrm{UA}}, f_{\mathrm{H}}]$2 个部分，以此类推，整个频率区间将可拆分成 Q 个子频带，如式(12)所示。

$$Q=\mathrm{ceil}\left(\frac{f_{\mathrm{H}}-f_{\mathrm{L}}}{f_{\mathrm{UA}}}\right) \tag{12}$$

其中，ceil(·) 表示向上取整运算符。

据此，首先求得各个子频带内的密度谱函数

$$\varphi^{(q)}(\tau)=\frac{1}{(N_{q\mathrm{H}}-N_{q\mathrm{L}}+1)N_r}\sum_{k=N_{q\mathrm{L}}}^{N_{q\mathrm{H}}}\sum_{r}g(r,k,\tau),q=1,2,\cdots,Q \tag{13}$$

其中，$N_{q\mathrm{L}}$、$N_{q\mathrm{H}}$ 分别表示第 q 个子频带所对应的最低及最高离散频率的序号，其中，第 1 个子频带对应的谱密度函数 $\varphi^{(1)}(\tau)$ 不存在高频混叠的影响，因此，$\varphi^{(2)}(\tau)$ 与 $\varphi^{(1)}(\tau)$ 相乘将能有效抑制第 2 个子频带中高频混叠带来的模糊，此时将受到加权抑制混叠的 $\varphi^{(1)}(\tau)\cdot\varphi^{(2)}(\tau)$ 作为下一个子频带的加权因子，以此类推，直至整个讨论的频率范围，可得 KDEMS 的计算式如下。

$$\varphi_{\mathrm{KDEMS}}(\tau)=\prod_{q=1}^{Q}\varphi^{(q)}(\tau) \tag{14}$$

该方法考虑了高频混叠对 KDE 算法的影响，通过 MS 分频带处理，逐个频带顺次相乘达到抑制混叠的目的。

3.3　基于 SLF 的多维融合

上面讨论了对空间中任意一对麦克风(a, b)的 KDE 和 KDEMS 时延函数的求取。假设(a, b)为所有考虑的麦克风阵元对的集合 Ω_p 中的第 p 个元素，即 $(a,b)\in\Omega_p\subseteq\{(a,b)\,|\,(1\leqslant a<b\leqslant M)\}$，则 2 节中的 NCS 可扩展表示为 NCS_p，第 2 节及第 3.2 节中所求时延 τ 扩展为 τ_p，3.1 节中的 $W_{COH}(r,k)$ 扩展为 $W_{COH,\,p}(r,k)$。本文定义如下空域似然率函数 SLF[7] 来表征近场声源出现在空间中各个位置的可能性，第 p 对麦克风的 SLF_p 可表示为

$$\mathrm{SLF}_p(\boldsymbol{l})=\mathrm{Nl}[\boldsymbol{\Phi}(\tau_p(\boldsymbol{l}))] \tag{15}$$

其中，$\boldsymbol{\Phi}$ 表示 φ_{KDE} 或 φ_{KDEMS}，$\boldsymbol{l}=[x,y,z]$ 表示空间中各点的笛卡尔系位置坐标，Nl[·] 表示将函数值标准化到[0,1]区间的运算符。

$$\tau_p(\boldsymbol{l})=\frac{\|\boldsymbol{l}-\boldsymbol{l}_a\|-\|\boldsymbol{l}-\boldsymbol{l}_b\|}{c} \tag{16}$$

式(16)表示位置 $\boldsymbol{l}$ 到第 p 对麦克风的相对时延，其中，$\boldsymbol{l}_a$ 和 $\boldsymbol{l}_b$ 表示第 p 对麦克风 2 个阵元的位置，$\|\cdot\|$ 表示欧氏距离运算符。

上述时延函数 $\boldsymbol{\Phi}(\tau_p)$ 到 SLF 的映射过程如图 1 所示。其中，图 1(a)表示单对麦克风的时延函数（此处用 KDE 示意），横坐标 τ_p 相对于最大可能时延 $\tau_{\max}$ 做了归一化处理，图中箭头位置标示出了 2 个声源 S_1、S_2 的准确时延；图 1(b)为通过式(15)、式(16)映射后的关于二维平面坐标的定位函数 SLF_p，图中标示圆形、三角形的位置分别为声源 S_1、S_2 和麦克风对 a、b 的准确位置。

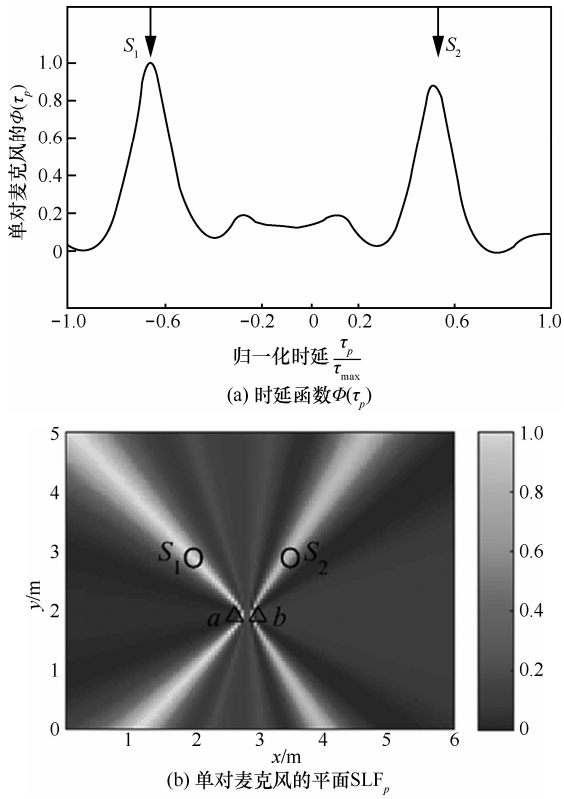

图 1　时延函数到 SLF 的映射示意

由图 1(b)可以看出，单对麦克风所对应的 $\mathrm{SLF}_p(\boldsymbol{l})$ 包含一系列由不同时延 $\tau_p(\boldsymbol{l})$ 处的函数值 τ_p 加权的双曲线，在正确的声源位置坐标处并没有显示出唯一的极值。要实现声源的定位，需要进一步对 Ω_P 中各对麦克风的 SLF 进行多维空域融合，利用阵列空间分布的多样性确定声源的位置[6,7]。具体的定位方法如式(17)所示。

$$\Omega(\hat{\boldsymbol{l}}_S)=\left\{\underbrace{\operatorname{argmax}}_{\boldsymbol{l}\in\Omega(\boldsymbol{l})}\left(\underset{p\in\Omega_P}{\otimes}\mathrm{SLF}_p(\boldsymbol{l})\right)\right\} \tag{17}$$

其中，$\Omega(\hat{\boldsymbol{l}}_S)$ 表示多个声源估计位置的集合，$\Omega(\boldsymbol{l})$ 表示空间中所有点位置的集合，$\otimes$ 表示满足式(18)所示运算准则[7]的 SLF 融合方式。

$$\begin{cases}\otimes(A,B)=\otimes(B,A)(\text{交换律})\\ \otimes(A,B)\leqslant\otimes(C,D),\text{如果}A\leqslant C\text{且}B\leqslant D(\text{单调性})\\ \otimes(\otimes(A,B),C)=\otimes(A,\otimes(B,C))(\text{结合律})\\ \text{其中，}A,B,C,D\in\{\mathrm{SLF}_p(\boldsymbol{l})\,|\,p\in\Omega_P\}\end{cases} \tag{18}$$

其中计算量较小且具有普适性的是相加(S)以及相乘(P) 2 种基本融合方式[7]，即

$$\begin{cases}\mathrm{SLF}_{\mathrm{ALL}}(\boldsymbol{l})|_{\mathrm{S}}=\dfrac{1}{P}\displaystyle\sum_{p\in\Omega_p}\mathrm{SLF}_p(\boldsymbol{l})\\ \mathrm{SLF}_{\mathrm{ALL}}(\boldsymbol{l})|_{\mathrm{P}}=\displaystyle\prod_{p\in\Omega_P}\mathrm{SLF}_p(\boldsymbol{l})\end{cases} \tag{19}$$

其中，$\mathrm{SLF}_{\mathrm{ALL}}(\boldsymbol{l})$表示融合 SLF，$P$=card(Ω_P)表示考虑麦克风对的总个数，card(·) 表示求取非空集合中元素数量的运算符。

图 2(a)、图 2(b)分别给出了 2 对麦克风 p_1、$p_2\in\Omega_p$ 在这 2 种融合方式下的计算结果，图中曲线左侧标示的数值为相应等高线的幅度。由图 2(a)可知，使用 S 时，当其中一对麦克风的 SLF 幅度较小时，只要另一对幅度足够大，融合后 SLF 的幅度仍然比较明显；而图 2(b)中，使用 P 时，只有 2 对麦克风的 SLF 幅度均比较大时，融合后的 SLF 才会比较明显。由此可见，对不同麦克风对之间的多样性差异，P 较 S 更加敏感。

3.4 基于 KDE 的多声源定位算法步骤及流程

综上所述，基于 KDE 的近场多声源定位算法的处理步骤如下。

步骤 1 对阵元数为 M 的麦克风阵列观察信号做 STFT，得到时频支撑域(r,k)处各个麦克风的 $X_m(r,k)$。

步骤 2 由式(5)求得集合 Ω_p 中各对麦克风在(r,k)处的 $NCS_p(r,k)$，并由式(8)、式(10)求得相应的相关性权值系数 $W_{COH,\,p}(r,k)$。

步骤 3 由式(7)或式(14)，将通过 $W_{COH,\,p}(r,k)$ 加权预处理的 $NCS_p(r,k)$进行 KDE 或 KDEMS 处理，得到关于时延的密度谱函数 $\Phi(\tau_p)$。

步骤 4 由式(15)和式(16)构建各对麦克风关于空间近场声源位置的 $\mathrm{SLF}_p(\boldsymbol{l})$。

步骤 5 由式(19)进行多维空域融合，得到 $\mathrm{SLF}_{\mathrm{ALL}}(\boldsymbol{l})$，再按式(17)进行空间峰值搜索，从而估计出各个声源的位置。

图 3 给出了相应的算法流程。本文将式(19)衍生出的 4 种 KDE 类算法分别称作 S-KDE、S-KDEMS、P-KDE 和 P-KDEMS，下面通过计算机仿真与统计分析来详细研究其定位性能。

4 计算机仿真与统计分析

上述基于 KDE 的近场多声源定位算法包含由时延函数构建 SLF 以及 SLF 融合 2 个部分。相应地，首先考察单对麦克风时 KDE 与 KDEMS 的时延估计结果，并与经典的 PHAT[2]进行比较；进而讨论 S 与 P 这 2 种融合方式下，4 种 KDE 类算法在不同混响时的近场定位性能，并与 SPR-PHAT、Multi-PHAT 这 2 种 PHAT 类算法进行综合比较；最后给出不同阵元间距时各种近场定位算法的统计分析。下面分节具体说明。

4.1 仿真基本参数设定

图4给出了计算机仿真中麦克风阵列与声源的室内平面位置示意。其中，房间尺寸为 6 m×5 m×3 m，

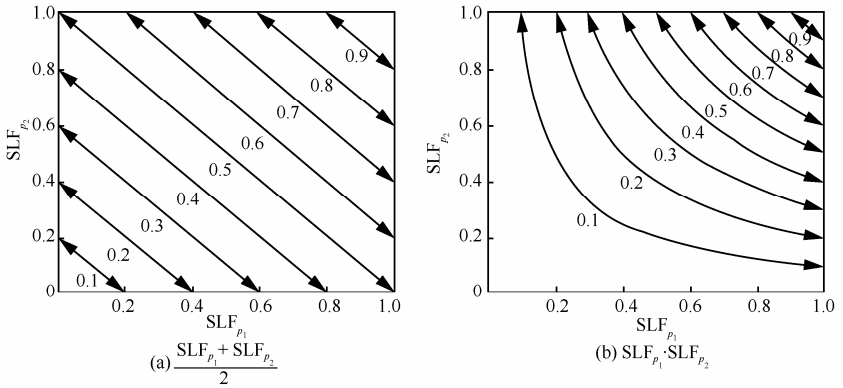

图 2 S 与 P 作用下 2 种融合方式的等高线

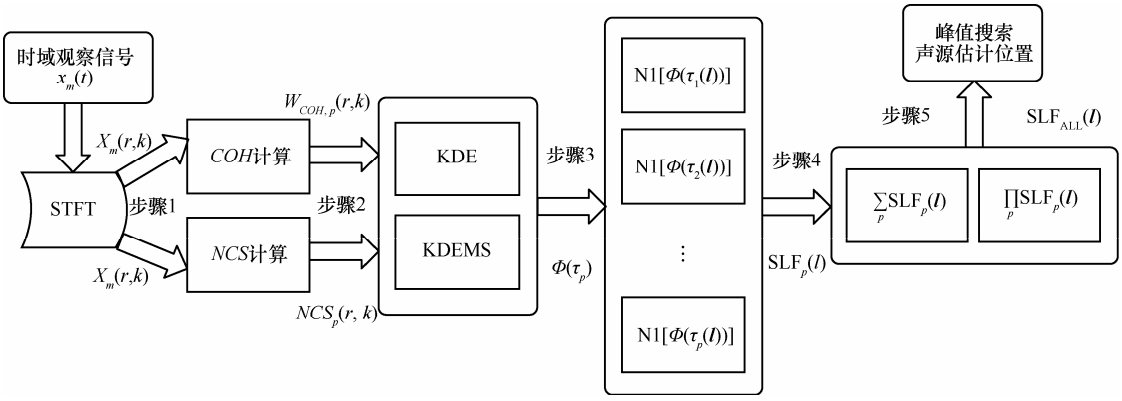

图 3　基于 KDE 的近场多声源定位算法流程

四元分布式阵列由序号 m 分别为 1~4 的全向阵元组成（由三角形表示），其中心位置 M_0 的全局坐标为 [3.00, 1.38, 1.50]，各阵元相对 M_0 的位置关系如图 4 所示，取阵元间距为 d 的 3 组麦克风对(1, 2)、(1, 3)、(2, 4)组成 Ω_P，各阵元对与水平方向的夹角如图所示。2 个声源 S_1、S_2（由圆形表示）位于以 M_0 为圆心，半径 1.35 m 的圆弧上，其坐标分别为[2.54, 2.65, 1.50]和[3.46, 2.65, 1.50]，以上位置坐标的单位均为 m。由于空间中各个麦克风与声源的 z 轴坐标均为 1.5 m，此时三维空间定位可简化为二维平面定位。

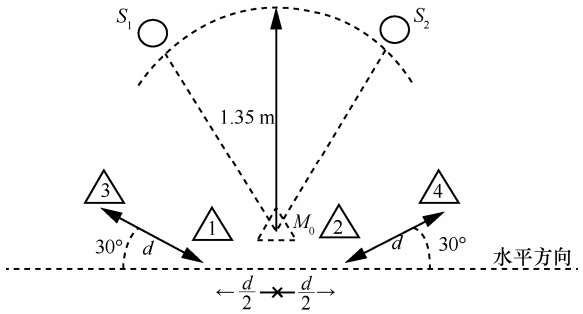

图 4　室内环境平面示意

声源信号来自由 12 个男声、12 个女声所组成的语音数据库，每次仿真从中取出长度约 2 s、采样率 f_s = 16 kHz、平均功率相同的 2 段信号，由镜像法[16]产生室内信道冲激响应，按式(1)卷积混合并加入固定 SNR = 20 dB 的白噪声形成观察信号，其中，声速 c 取 344 m/s，SNR 定义如下。

$$SNR = 10\lg \frac{\sum_{m=1}^{M} \sigma_m^2}{M\sigma_w^2} \tag{20}$$

其中，σ_m^2 为第 m 个阵元观察信号的平均功率，σ_w^2 为加性白噪声的平均功率。将观察信号通过 f_L=20 Hz，f_H=4 000 Hz 的带通滤波器，此时 $\lambda_{\min}=\frac{c}{f_H}$ =0.086 m，再将带通滤波后的信号做 1 024 点（合 64 ms）汉明窗加权、帧移 25%（合 16 ms）的 STFT。相关性检测中，C 取 2，Thd 取相应环境下的经验值为 0.8。近似核密度函数的计算中，B 取 20。平面定位中的网格宽度取 0.05 m。

4.2　单对麦克风时延估计性能比较

首先讨论混响时间 T_{60} 分别为 250、450 ms，阵元间距 d 分别取 $0.5\lambda_{\min}$、$4\lambda_{\min}$ 时 PHAT、KDE 以及 KDEMS 的时延估计性能。所用单对麦克风为图 4 阵列中第 1 组阵元(1, 2)，所用声源信号 S_1、S_2 时域波形及其时频谱如图 5 所示。由图 5(c)、图 5(d)可以看出 2 个语音信号的短时频谱稀疏性明显，各声源的能量都集中在比例很小的时频支撑域内。本文使用 3.1 节中的 CT 环节保证计算时的时频支撑域内单个声源能量占优。图 6、图 7 给出了阵元间距 d 分别为 $0.5\lambda_{\min}$、$4\lambda_{\min}$ 时的估计结果，各子图中左右 2 个箭头分别表示 S_1、S_2 正确的时延位置，横坐标 τ_1 以 $\tau_{\max}$ 为基准进行了归一化处理，纵坐标为[0, 1]区间的归一化幅度。

由图 6 可以看出，当 $d = 0.5\lambda_{\min}$ 时，虽然间距设定满足空域奈奎斯特采样定理，但由此产生的过低分辨率使 3 种算法的主瓣过胖，PHAT 算法仅有一个谱峰，而 KDE 以及 KDEMS 虽然形成了 2 个谱峰，峰值位置却与正确时延的位置存在不小的误差。因此，采用较窄的间距 d 会对算法的准确估计产生影响，需要加宽该参数来改善算法的性能。

图 7 中，当 $d = 4\lambda_{\min}$ 时，分辨率的提高使时延谱中主瓣过胖的现象基本消失。由图 7(a)可以看出，对于 PHAT 算法，在混响较弱时，宽间距引入的高频混叠使其伪峰幅度逼近正确时延位置的峰值，

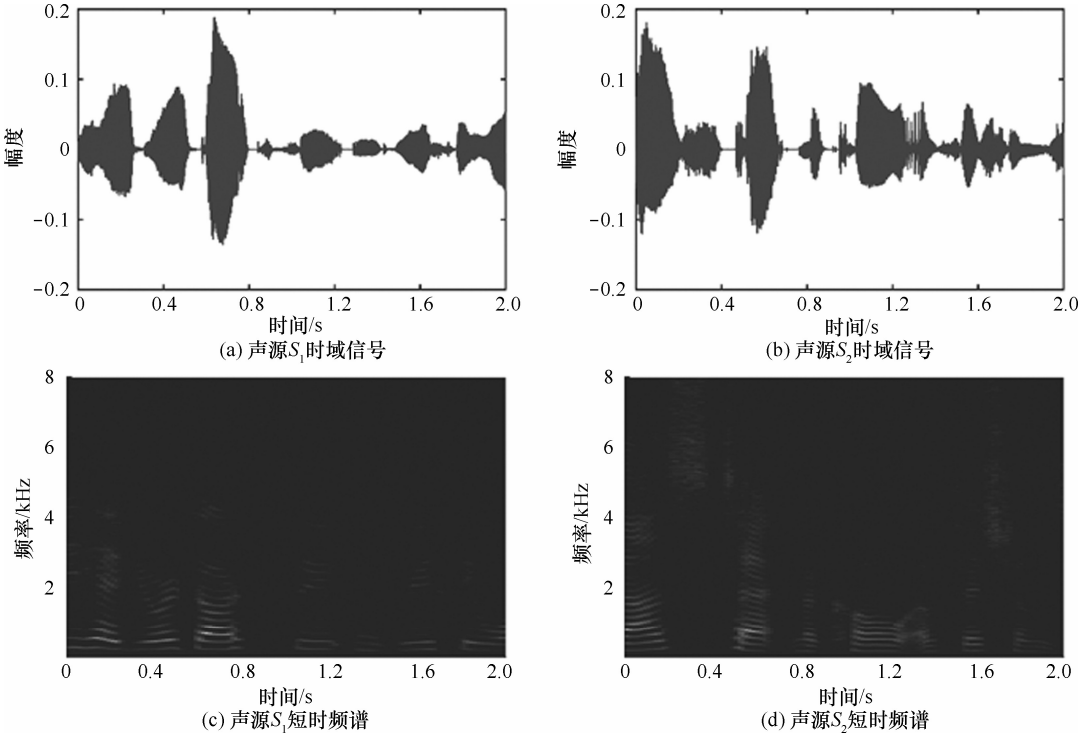

图 5　单次仿真所用声源 S_1、S_2 的时域波形及短时频谱

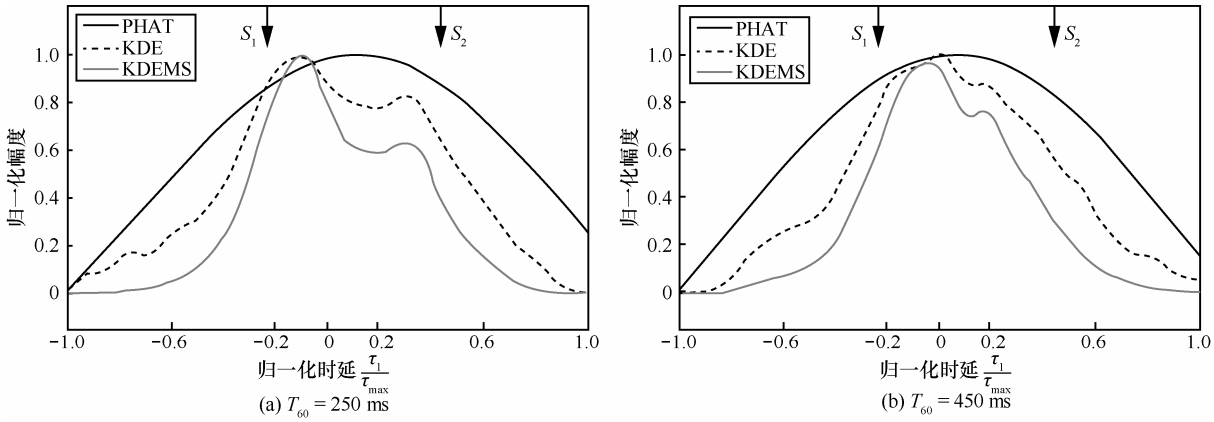

图 6　$d = 0.5\,\lambda_{\min}$ 时基于麦克风对 1 的 PHAT、KDE 以及 KDEMS 3 种算法的时延估计结果

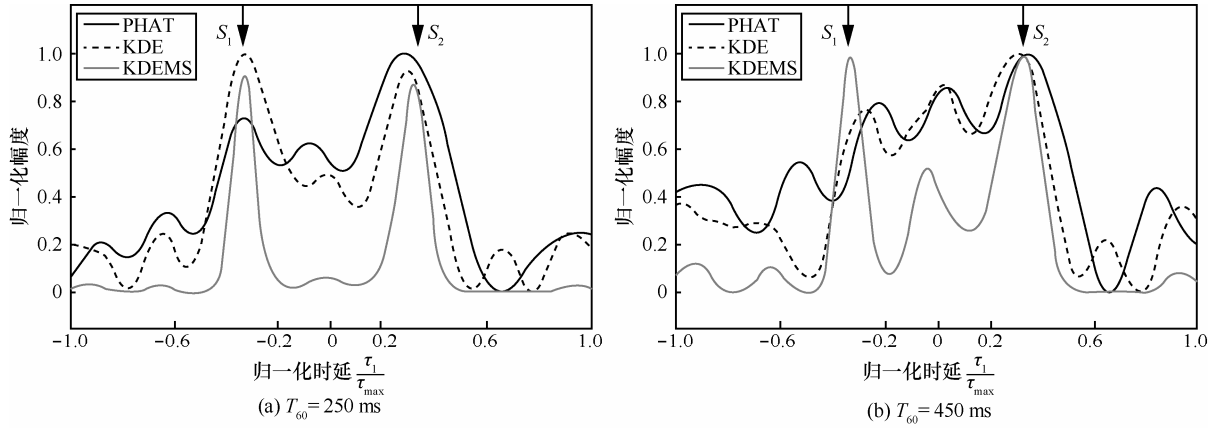

图 7　$d = 4\,\lambda_{\min}$ 时基于麦克风对 1 的 PHAT、KDE 以及 KDEMS 3 种算法的时延估计结果

有引起虚警的风险；而 KDE 算法由于$\frac{1}{2\pi f_k}$频率加权因子对高频混叠的抑制作用，伪峰幅度有所降低；相较于 KDE，可以发现 KDEMS 的分频带处理基本消除了高频混叠对谱基底的累加干扰，伪峰的削弱更加明显。图 7(b)中，在较强混响情况下，PHAT 与 KDE 时延谱中的伪峰幅度均超过了正确位置的峰值；而 KDEMS 由于 MS 强烈的抑制作用，伪峰幅度的提高不足以对声源的正确时延估计产生影响，结果仍在可控范围，这体现了该方法在较宽间距及较强混响综合作用下稳健的时延估计性能。

4.3　近场多声源定位性能比较

4.2 节中单对阵元的性能比较体现了宽间距时 KDEMS 算法相比 PHAT、KDE 的时延估计可靠性。下面给出上述 3 种方法构建的 SLF 在 S 和 P 这 2 种融合方式下的近场多声源定位性能。所用声源及房间参数设置与 4.2 节相同，取 Ω_p 中的 3 组麦克风对进行 SLF 融合。图 8 和图 9 分别给出了 T_{60} 分别为 250、450 ms 时的定位结果，其中“O”表示声源的正确位置，“X”表示超越正确位置幅度的伪峰。求取融合谱 $\mathrm{SLF_{ALL}}$ 中幅度最大的 10 个谱峰，将相邻（对应位置坐标的欧氏距离小于等于 0.2 m）谱峰划分为一组，其中幅度最大的位置作为这组谱峰的估计位置。

比较图 8(a)、图 8(c)和图 8(e)，可以看出，对于 SPR-PHAT 与 S-KDE，尽管混响较弱，但较宽的阵元间距 d 使各对阵元的 SLF 谱基底较高，经 S 融合后的 $\mathrm{SLF_{ALL}}$ 虽然在正确的声源位置处出现了最大的 2 个峰值，但正确位置周围的峰脊较宽较高，定位图像的辨识度较差，而 S-KDEMS 由于包含 MS 分频带处理环节，各对麦克风的 SLF 谱基底被大大削弱了，辨识度有了明显的提高；S-KDE 相比 SRP-PHAT 谱基底与伪峰幅度略低，但 2 种算法均在位置[3.05,2.15,1.50]附近出现了比较明显的伪峰群，有引起虚警的趋势。比较图 8(a)、图 8(c)、图 8(e)与图 8(b)、图 8(d)、图 8(f)，可以看出，由于各

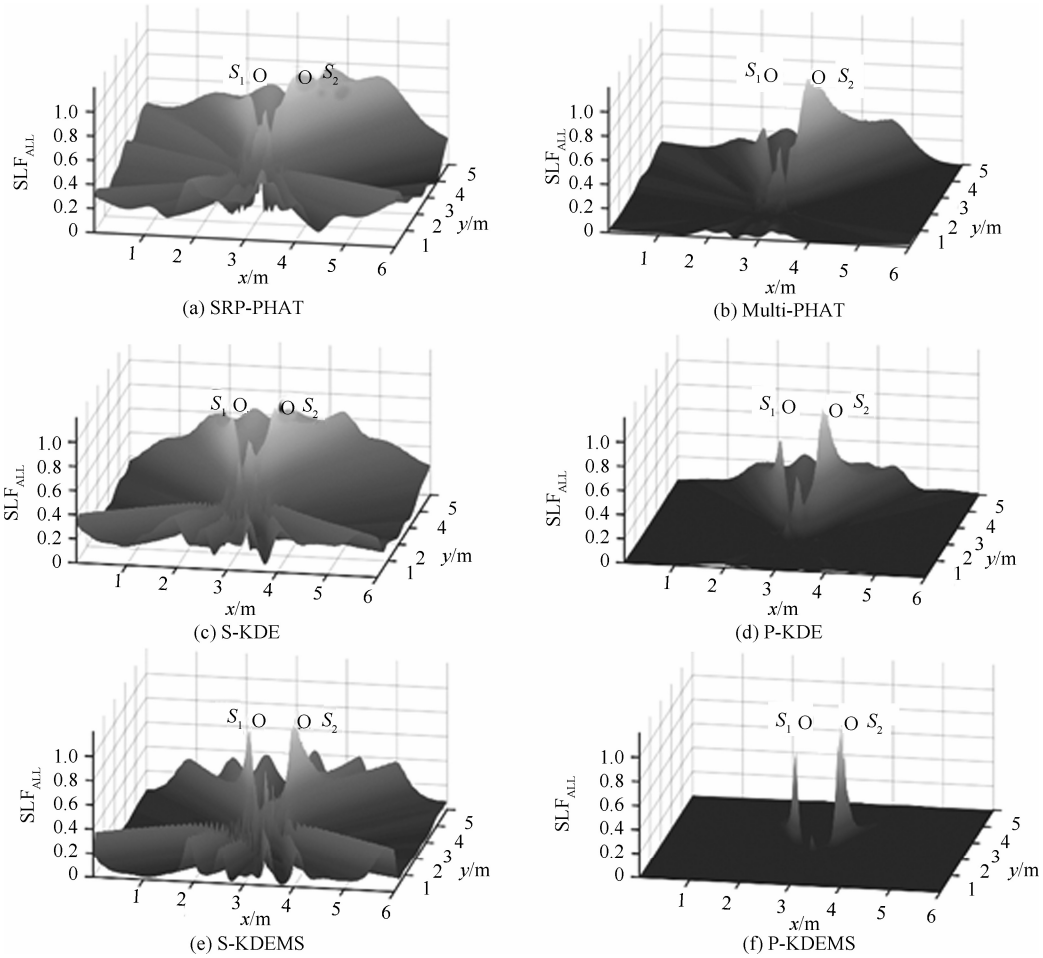

图 8　T_{60} = 250 ms，阵元间距 $d = 4\,\lambda_{\min}$ 时，6 种算法的定位结果

对麦克风分布的空域多样性，P 相比 S 可以进一步削弱 SLF 谱基底，且由混响拖尾与噪声引起的小伪峰（毛刺）基本被滤除，图 8(f)中，P-KDEMS 的融合谱 $\mathrm{SLF}_{\mathrm{ALL}}$ 中在正确的声源位置附近出现了辨识度非常高的 2 个峰值。

当混响增强时，由图 9(a)~图 9(d)可以看到，在宽间距引起的高频混叠与强混响的综合影响下，位置[3.05,2.15,1.50]附近的伪峰幅度超过了正确位置，因而产生了错误的定位结果；P 相比 S 虽可以降低谱基底的高度，但并不足以克服伪峰的干扰。而图 9(e)和图 9(f)中，MS 对 KDE 算法的改进可以很好地抑制伪峰的增长，从而避免错误的发生。比较图 8(f)和图 9(f)，可以发现，集成 MS 分频带处理与融合方式 P 2 个环节的 P-KDEMS 在混响由弱变强的过程中，性能并没有出现明显的恶化，均能够带来清晰明辨的定位图像。

4.4 指标验证与统计分析

4.3 节的单次仿真结果直观形象地比较了各种近场定位算法的性能，为更加全面地进行比较验证，本文引入如下均方根误差 *RMSE* 以及 SLF 百分比 *PSLF* 2 个量化指标，其单次计算的公式如下

$$RMSE=\sqrt{\frac{1}{2}\sum_{n=1}^{2}\left\|\hat{\boldsymbol{l}}_{S,n}-\boldsymbol{l}_{S,n}\right\|^{2}} \tag{21}$$

其中，$\boldsymbol{l}_{S,n}$、$\hat{\boldsymbol{l}}_{S,n}$ 分别表示声源 S_n, $n=1,2$ 的正确位置及估计位置的坐标，与 4.3 节类似，本文取融合谱 $\mathrm{SLF}_{\mathrm{ALL}}$ 中幅度最大的 10 个谱峰，将欧氏距离小于等于 0.2 m 的划分为一组，最大的谱峰位置作为该组位置的估计值，则前 2 组的估值即为声源估计位置的坐标。

$$PSLF=\frac{\mathrm{card}(\Omega(\hat{\boldsymbol{l}}_S))}{\mathrm{card}(\{(x,y)|\mathrm{SLF}_{\mathrm{ALL}}(x,y)>0.2\})} \tag{22}$$

PSLF 是一种量化表征定位辨识度的统计指标，式(22)中，分子表示位置估计正确的声源的个数（估计位置的坐标与正确位置的欧式距离小于等于 0.2 m 视为正确），分母表示 $\mathrm{SLF}_{\mathrm{ALL}}$ 中幅度大于 0.2 的谱峰个数。

从语音数据库 24 个声源中顺次取出 2 个不同的声源（取共计 500 组），T_{60} 取 250、450 ms 2 种情形，将 $0.5\lambda_{\min}$~$5\lambda_{\min}$ 范围内间隔 $0.5\lambda_{\min}$ 的 10

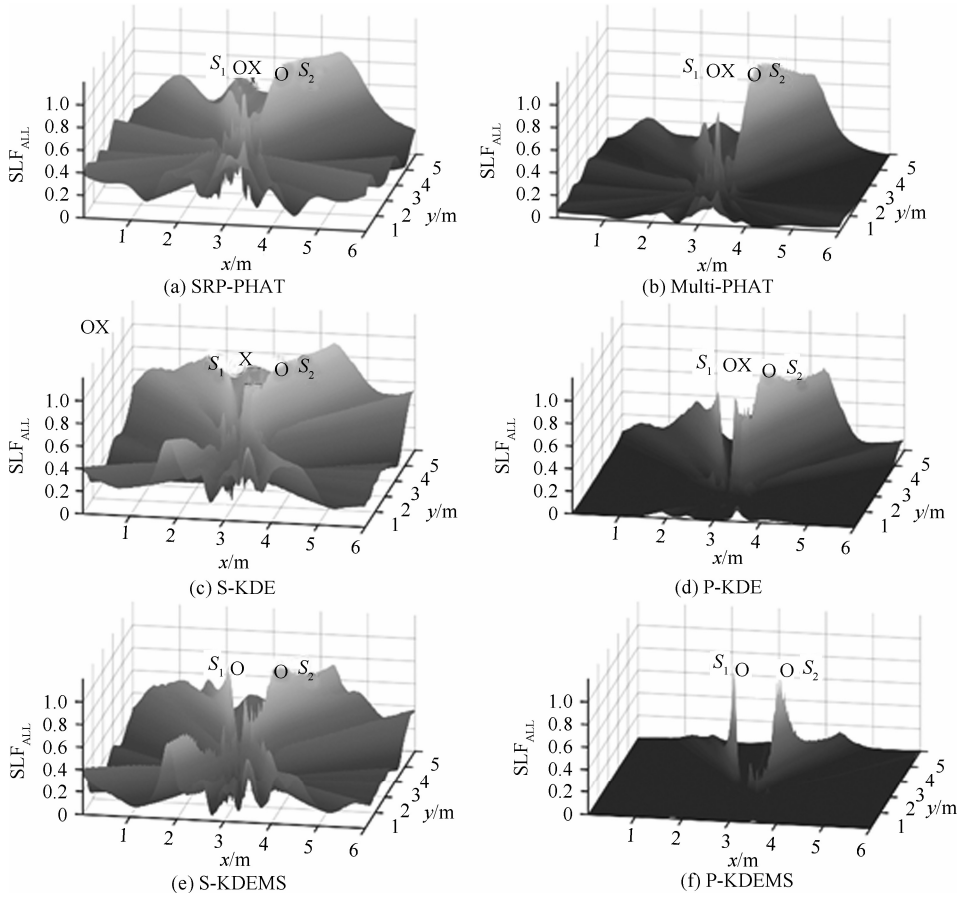

图 9 $T_{60}=450$ ms，阵元间距 $d=4\lambda_{\min}$ 时，6 种算法的定位结果

组 d 均做 500 次仿真，每次仿真按式(21)、式(22)求取 2 种指标，最后求和取平均值，所得统计结果如图 10 和图 11 所示。

图 10(a)和图 10(b)分别给出了 $T_{60}=250$ ms、450 ms 时 6 种算法的 *RMSE*。由图 10(a)可以看出，较弱混响环境下，$d=0.5\lambda_{\min}$ 时，高频混叠的影响虽不存在，但单对麦克风的时延分辨率较低(见 4.2 节)，导致估计偏差较大，最终的定位性能并不理想，由此可见，较窄间距时的低分辨率并不适用于多源定位；随着 d 的变宽，分辨率的提高使 *RMSE* 逐渐变小，在 $d=2\lambda_{\min}$ 时趋于稳定；2 种 PHAT 类定位算法在 $d\leqslant 3\lambda_{\min}$ 时与 4 种 KDE 类的变化趋势相似，但定位误差相对较高；当 $d>3\lambda_{\min}$ 时，由于 $\frac{1}{2\pi f_k}$ 加权因子对高频混叠有一定的抑制作用，4 种 KDE 类算法的定位结果均比较稳定，而 2 种 PHAT 类算法的 *RMSE* 有逐渐增大的趋势。

图 10(b)中，当 $d\leqslant 3\lambda_{\min}$ 时，6 种算法维持与弱混响时类似的变化趋势，混响增强使 *RMSE* 指标有所恶化；当 $d>3\lambda_{\min}$ 时，虽然 S-KDE、P-KDE 包含 $\frac{1}{2\pi f_k}$ 加权因子，但其抑制作用不足以克服强混响与高频混叠的综合影响，因此与无抗混叠措施的 PHAT 类算法均出现了融合谱 $\mathrm{SLF_{ALL}}$ 中伪峰幅度高于正确谱峰的情形，从而产生了较大的误差，*RMSE* 增幅急剧变大；而 S-KDEMS、P-KDEMS 的 RMSE 指标并没有出现明显变化，定位性能仅有轻微下降，可见分频带 MS 使强混响时算法仍然能够稳健的应对高频混叠的干扰，从而有效地抑制了伪峰幅度的增长，将指标值控制在一个合理的范围。

图 11(a)和图 11(b)分别给出了 T_{60} = 250 ms 和 450 ms 时 6 种算法的 *PSLF*。由图 11(a)可以看出，不同算法的 *PSLF* 差别较大，3 种 S 系算法由于不同位置麦克风对的 SLF 谱的累加，幅度高于 0.2 的谱峰数较多，其 *PSLF* 并不理想；而 3 种 P 系算法由于 SLF 谱的相乘对小伪峰（毛刺）的消除以及谱基底的抑制均比较明显，因而其 *PSLF* 指标远好于 S 系的算法，带来了较高辨识度的融合 SLF，其中，

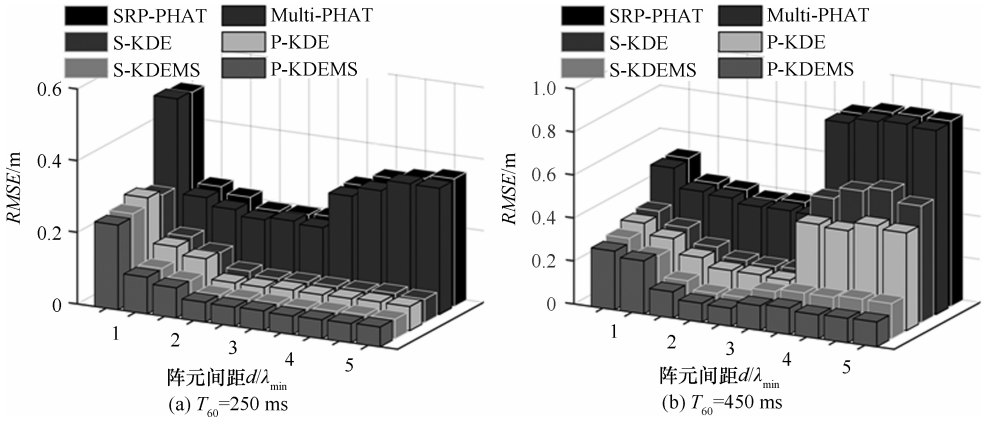

图 10　6 种算法在不同阵元间距时的 *RMSE* 统计结果

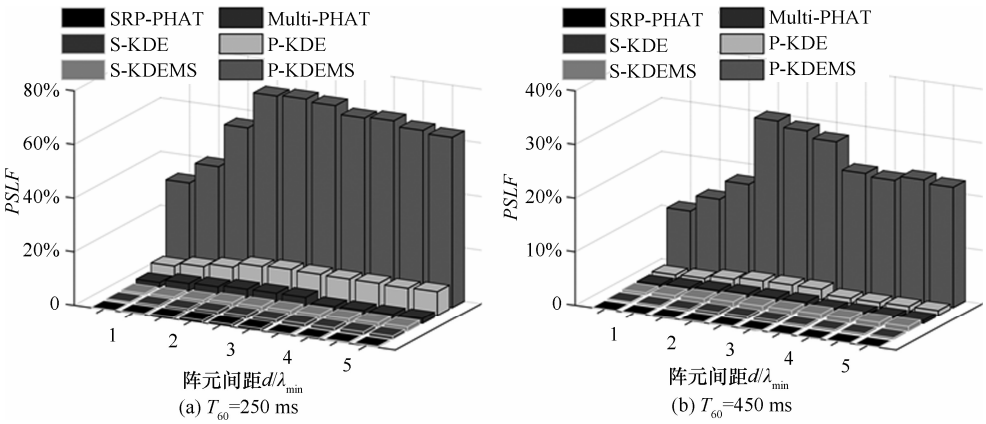

图 11　6 种算法在不同阵元间距时的 *PSLF* 统计结果

Multi-PHAT 辨识度最低，而 P-KDEMS 相比 P-KDE 多了一层 MS 处理，融合谱 SLF_{ALL} 的分布更加集中，定位的辨识度最高。

图 11(b)中，对于 P-KDE 算法，伪峰的出现使得 *PSLF* 出现了明显的恶化，但仍好于 3 种 S 系算法；由于 MS 对伪峰增长的抑制，P-KDEMS 的恶化程度没有 P-KDE 明显，其 *PSLF* 远好于其他 5 种算法。

5 结束语

针对混响环境下的近场多声源定位问题，本文运用多声源时频稀疏假设，引入相关性检测确保假设成立，同时，由 *NCS* 计算 KDE，并引入 MS 分频带处理，再由 KDE 及改进的 KDEMS 时延函数构建 SLF，进而通过 S 或 P 这 2 种基本融合方式将位于不同观察位置的麦克风对时延函数进行多维融合，从而建立了一套包含 S-KDE、P-KDE、S-KDEMS、P-KDEMS 等 4 种算法的近场多声源定位模型。通过与传统 PHAT 类算法的纵向比较以及 *RMSE* 和 *PSLF* 指标的统计分析，本文详细研究了不同算法的定位性能。研究结果表明：MS 分频带处理在较宽阵元间距时可以有效抑制高频混叠，削弱模糊伪峰对算法性能的影响；融合方式 P 相比 S 对于空域多样性的高敏感性可以消除小伪峰（毛刺）以及降低 SLF 的谱基底；集成 MS 与 P 的 P-KDEMS 其 *RMSE* 及 *PSLF* 2 个指标均优于其他几种算法，在较强混响及较宽间距环境下，仍然能够带来稳定清晰的 SLF 融合谱，因而是一种稳健性较高、辨识度较好的近场多声源定位算法。

当麦克风阵元间距过宽时，由于不模糊频带累积的能量过小，会对后续的 MS 分频带加权产生消极影响，造成声源的位置信息丢失。上述问题仍在分析解决中，同时后续研究将围绕麦克风阵元对的智能选取以及多声源的追踪展开。

参考文献：

[1] WU K, KHONG A W H. Sound source localization and tracking [M]// Context Aware Human-Robot and Human-Agent Interaction. Springer International Publishing, 2016: 55-78.

[2] KNAPP C H, CARTER G C. The generalized correlation method for estimation of time delay[J]. IEEE Transactions on Acoustics, Speech, and Signal Processing, 1976, 24(4): 320-327.

[3] BRANDSTEIN M S, SILVERMAN H F. A practical methodology for speech source localization with microphone arrays[J]. Computer Speech and Language, 1997, 11(2): 91-126.

[4] RABINKIN D V, RANOMERON R J, DAHL A, et al. A DSP implementation of source location using microphone arrays[J]. Journal of the Acoustical Society of America, 1996, 99(4): 88-99.

[5] WARD D B, WILLIAMSON R C. Particle filter beamforming for acoustic source localization in a reverberant environment[C]//2002 IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP). Orlando, USA, 2002: 1777–1780.

[6] DIBIASE J H. A high-accuracy, low-latency technique for talker localization in reverberant environments using microphone arrays[D]. Brown University, 2000: 73-105.

[7] PERTILÄ P, KORHONEN T, VISA A. Measurement combination for acoustic source localization in a room environment[J]. EURASIP Journal on Audio Speech and Music Processing, 2008, 2008: 1-14.

[8] TSIAMI A, KATSAMANIS A, MARAGOS P, et al.. Experiments in acoustic source localization using sparse arrays in adverse indoors environments[C]//2014 European Signal Processing Conference (EUSIPCO). Lisbon, Portugal, 2014: 2390-2394.

[9] XU Z Y, ZHAO Z, LIU M. Real-time unambiguous passive direction finding for multiple sound sources with widely spaced microphone array[J]. Journal of Electronics & Information Technology, 2011, 33(9): 2056-2061.

[10] NESTA F, OMOLOGO M. Generalized state coherence transform for multidimensional TDOA estimation of multiple sources[J]. IEEE Transactions on Audio, Speech, and Language Processing, 2012, 20(1): 246-260.

[11] BRUTTI A, NESTA F. Tracking of multidimensional TDOA for multiple sources with distributed microphone pairs[J]. Computer Speech and Language, 2013, 27(3): 660-682.

[12] YILMAZ O, RICKARD S. Blind separation of speech mixtures via time-frequency masking[J]. IEEE Transactions on Signal Processing, 2004, 52(7): 1830-1847.

[13] REDDY V V, KHONG W H, NG B P. Unambiguous speech DOA estimation under spatial aliasing conditions[J]. IEEE Transactions on Audio, Speech, and Language Processing, 2014, 22(12): 2133-2145.

[14] MOHAN S, LOCKWOOD M E, KRAMER M L, et al. Localization of multiple acoustic sources with small arrays using a coherence test[J]. Journal of the Acoustical Society of America, 2008, 123(4): 2136-2147.

[15] GUSTAFFSON T, RAO B D, TRIVEDI M. Source localization in reverberant environments: modeling and statistical analysis[J]. IEEE Transactions on Speech and Audio Processing, 2003, 11(6): 791-803.

[16] LEHMANN E and JOHANSSON A. Prediction of energy decay in room impulse responses simulated with an image-source model[J]. Journal of the Acoustical Society of America, 2008, 124(1): 269-277.

作者简介：

房玉琢（1987-），男，江苏南京人，南京理工大学博士生，主要研究方向为阵列信号处理、声学探测、盲信道辨识等。

许志勇（1968-），男，江苏南京人，博士，南京理工大学副教授，主要研究方向为阵列信号处理、声学探测、雷达技术等。

赵兆（1979-），男，湖北襄阳人，博士，南京理工大学副教授，主要研究方向为声探测系统与信号处理、时频分析。