基于加权子空间拟合的声源定位与跟踪方法

金乃高　殷福亮　陈　喆
(大连理工大学电子与信息工程学院　大连　116024)

摘　要：麦克风阵列声源定位可为复杂环境下的说话人空间位置估计问题提供一种有效的解决方案。该文基于粒子滤波框架，提出了一种加权子空间拟合声源定位与跟踪方法。该方法将窄带子空间拟合算法的代价函数推广至宽带情形，构建了一种适用于宽带语音信号的似然函数，并结合说话人的运动模型估计声源的位置。计算机仿真与实测结果验证了该方法的有效性。

关键词：粒子滤波；加权子空间拟合；声源定位；麦克风阵列
中图分类号：TN912.3　　**文献标识码**：A　　**文章编号**：1009-5896(2008)09-2134-04

Weighted Subspace Fitting Sound Source Localization Method Based on Particle Filtering

Jin Nai-gao　Yin Fu-liang　Chen Zhe
(*School of Electronic and Information Engineering*, *Dalian University of Technology*, *Dalian* 116024, *China*)

Abstract: Sound source localization using microphone array provides an effective solution to speaker tracking problem under adverse environments. A new method based on weighted subspace fitting is presented for sound source localization and tracking within the framework of particle filtering. The cost function of weighted subspace fitting algorithm is extended to the wideband case, and the location of speaker is estimated using dynamical model of speaker motion and the likelihood function model based on the wideband cost function. The results of both simulations with synthetic data and experiments with real-world data show that the proposed method has good performance.

Key words: Particle filter; Weighted subspace fitting; Sound source localization; Microphone array

1　引言

说话人定位与跟踪是人机交互研究中的一个重要课题，它在多媒体系统、视频会议系统、移动机器人等领域有着广泛的应用。例如，在视频会议系统中，说话人跟踪可为摄像机转向控制与语音拾取提供位置信息。另外，在移动机器人中，也需要根据说话人的空间位置进行路径规划。与人脸检测与跟踪方法[1]相比，基于麦克风阵列的声源定位方法具有全向定位能力，且运算量较低，易于构建实时系统，已成为估计说话人空间位置的又一可行方案[2]。

现有的声源定位方法主要分为基于时延估计的两步定位方法、波束形成方法和高分辨率空间谱估计方法。基于时延估计的方法[3]先进行时延估计，即计算声源到达两个麦克风的时间差，然后根据时延和麦克风阵列的几何结构估计出声源的位置。该类方法计算量小，易于实时实现，已在单声源定位系统中得到广泛应用。波束形成方法[4]不需要显式计算时延，而是通过优化目标函数直接实现声源定位。但在实际环境中，由于目标函数往往存在多个极值点，复杂峰值的搜索过程对优化方法提出了较高要求。基于空间谱估计的声源定位方法，如宽带MUSIC方法[5]和最大似然方法[6]，具有较高的空间分辨率，且可以同时定位多个声源，因此受到广泛关注。

在空间谱估计方法中，加权子空间拟合方法具有与最大似然方法相近的渐进最优估计性能，可在高信噪比下逼近最大似然方法的Cramer-Rao下界。考虑到虚假声源的位置变化具有明显的不确定性，而实际声源的运动通常具有一定的规律性，文献[7]将说话人运动模型引入声源定位问题，通过粒子滤波跟踪说话人的空间位置，可在一定程度上降低房间混响的影响。

本文在研究宽带加权子空间拟合方法基础上，提出了一种基于粒子滤波的麦克风阵列声源定位方法。该方法将窄带加权子空间拟合算法推广至宽带情形，在贝叶斯估计理论框架下构造出一个适用于宽带语音信号的似然函数，并采用粒子滤波跟踪声源的位置。计算机仿真与实测结果验证了本文方法的有效性。

2　麦克风阵列信号模型

在由 M 个全向麦克风组成的均匀线阵中，假设量测噪声为高斯白噪声，各麦克风之间的噪声不相关，信号与噪声相互独立。设空间中有 P 个声源，第 p 个声源信号为 $s_p(t)$，第 p 个声源至第 m 个麦克风的传播时延为 τ_{mp}，第 m 个麦克风的加性噪声为 $n_m(t)$。基于远场平面波假设，第 m 个麦克

2007-02-09收到，2008-01-10改回
国家自然科学基金(60772161，60372082)资助课题

风接收到的语音信号 $x_m(t)$ 可表示为

$$x_m(t)=\sum_{p=1}^{P}s_p(t-\tau_{mp})+n_m(t) \tag{1}$$

将观测数据 $x_m(t)$ 分成 K 个子段(K 为快拍数)，每个子段中的语音信号可视为平稳信号，$x_m(t)$、$s_p(t)$ 与 $n_m(t)$ 的离散傅里叶变换分别为 $X_m(f_j)$，$S_p(f_j)$，$N_m(f_j)$，则有

$$\boldsymbol{X}(f_j)=\boldsymbol{A}_\theta(f_j)\boldsymbol{S}(f_j)+\boldsymbol{N}(f_j),\quad j=1,\cdots,J \tag{2}$$

其中 $\boldsymbol{X}(f_j)=[X_1(f_j)\ X_2(f_j)\cdots X_M(f_j)]^{\mathrm{T}}$；$\boldsymbol{N}(f_j)=[N_1(f_j)\ N_2(f_j)\ \cdots\ N_M(f_j)]^{\mathrm{T}}$；$\boldsymbol{S}(f_j)=[S_1(f_j)\ S_2(f_j)\cdots S_P(f_j)]^{\mathrm{T}}$。方向矩阵 $\boldsymbol{A}_\theta(f_j)$ 由 P 个指向矢量(Steering Vector) $\boldsymbol{a}_{\theta,p}$ 组成，即

$$\boldsymbol{A}_\theta(f_j)=[\boldsymbol{a}_{\theta,1},\boldsymbol{a}_{\theta,2},\cdots,\boldsymbol{a}_{\theta,P}] \tag{3}$$

其中

$$\boldsymbol{a}_{\theta,p}=[1,\exp(-j2\pi f_j\tau_1),\cdots,\exp(-j2\pi f_j\tau_{M-1})]^{\mathrm{T}},p=1,\cdots,P \tag{4}$$

3　基于粒子滤波的加权子空间拟合声源定位方法

3.1 宽带加权子空间拟合方法

在频率 f_j 处，麦克风阵列频域采样数据 $X(f_j)$ 的协方差矩阵 $\boldsymbol{R}_{XX}(f_j)$ 为

$$\boldsymbol{R}_{XX}(f_j)=\boldsymbol{A}_\theta(f_j)\boldsymbol{R}_{SS}(f_j)\boldsymbol{A}_\theta^{\mathrm{H}}(f_j)+\boldsymbol{R}_{NN}(f_j) \tag{5}$$

其中 $\boldsymbol{R}_{SS}(f_j)$ 为频率 f_i 处的窄带信号 $S(f_j)$ 的协方差矩阵；$\boldsymbol{R}_{NN}(f_j)$ 为窄带噪声的协方差矩阵。$\boldsymbol{R}_{XX}(f_j)$ 的信号子空间 $\boldsymbol{U}_S(f_j)$ 与噪声子空间 $\boldsymbol{U}_N(f_j)$ 正交，且信号子空间 $\boldsymbol{U}_S(f_j)$ 与阵列流形 $\boldsymbol{A}_\theta(f_j)$ 张成同一空间。

若每个子段的持续时间大于源的相关时间，且各个窄带频率成分之间没有频谱混叠，则宽带子空间拟合方法的目标函数 $L(\theta)$ 可以表述为 J 个窄带目标函数 $L_i(\theta)$ 之和[8]，即

$$L(\theta)=\frac{1}{J}\sum_{j=1}^{J}L_j(\theta)=\frac{1}{J}\sum_{j=1}^{J}\left\|\boldsymbol{A}_\theta(f_j)\boldsymbol{T}_j-\boldsymbol{U}_S(f_j)\boldsymbol{W}^{1/2}(f_j)\right\|_F^2 \tag{6}$$

其中 $||\cdot||$ 表示矩阵的Frobenius范数，$\boldsymbol{W}(f_j)$ 为加权矩阵，$\boldsymbol{T}_j$ 为与 $\boldsymbol{W}(f_j)$ 有关的满秩变换矩阵。考虑阵列流形误差的影响，$\boldsymbol{T}_j$ 的最小二乘解 $\widehat{\boldsymbol{T}}_j$ 为

$$\widehat{\boldsymbol{T}}_j=\boldsymbol{A}_\theta^{\dagger}(f_j)\boldsymbol{U}_S(f_j)\boldsymbol{W}^{1/2}(f_j) \tag{7}$$

这里“$\dagger$”表示矩阵的Moore-Penrose伪逆，即 $\boldsymbol{A}_\theta^{\dagger}=(\boldsymbol{A}_\theta^{\mathrm{H}}\boldsymbol{A}_\theta)^{-1}\cdot\boldsymbol{A}_\theta^{\mathrm{H}}$。定义投影矩阵 $\boldsymbol{P}_A^{\perp}=\boldsymbol{I}-\boldsymbol{P}_A$，其中 $\boldsymbol{P}_A=\boldsymbol{A}_\theta\boldsymbol{A}_\theta^{\dagger}$。于是，宽带子空间拟合方法可以表示为以 $L(\theta)$ 为目标函数的优化问题，即

$$\begin{aligned}\hat{\theta}_{\mathrm{WB-WSF}}&=\arg\min_\theta L(\theta)\\&=\arg\min_\theta\left\{\frac{1}{J}\sum_{j=1}^{J}\left\|\boldsymbol{P}_A^{\perp}(f_j)\boldsymbol{U}_S(f_j)\boldsymbol{W}^{1/2}(f_j)\right\|_F^2\right\}\\&=\arg\max_\theta\left\{\frac{1}{J}\sum_{j=1}^{J}\mathrm{tr}\{\boldsymbol{P}_A(f_j)\boldsymbol{U}_S(f_j)\boldsymbol{W}(f_j)\boldsymbol{U}_S^{\mathrm{H}}(f_j)\}\right\}\end{aligned} \tag{8}$$

令信号子空间对应的对角阵为 $\boldsymbol{\Sigma}_S$，噪声协方差阵 $\boldsymbol{R}_{NN}(f_j)=\sigma_n^2\boldsymbol{I}_M$，则最优加权对角阵 $\boldsymbol{W}_{\mathrm{opt}}(f_j)$ 为

$$\boldsymbol{W}_{\mathrm{opt}}(f_j)=(\boldsymbol{\Sigma}_S(f_j)-\sigma_n^2\boldsymbol{I}_M)^2\boldsymbol{\Sigma}_S^{-1}(f_j) \tag{9}$$

在声源定位问题中，归一化似然函数 $\bar{L}(\theta)$ 符合概率密度函数的要求，方位参数 θ 的估计最终归结为一个多维积分问题。本文采用蒙特卡罗方法将多维积分问题转化为计算参数 θ 的数学期望，即从 $\bar{L}(\theta)$ 中抽取样本，并将样本的均值作为 θ 的估计值。由于 $\bar{L}(\theta)$ 是多维非线性函数，很难从中直接产生样本，因此本文采用基于贯序重要性抽样的粒子滤波算法来解决这个问题。

3.2 粒子滤波在麦克风阵列声源定位与跟踪中的应用

粒子滤波将贝叶斯理论与蒙特卡罗方法相结合，使用非参数化的序贯蒙特卡罗方法实现递推贝叶斯滤波[9]。本文采用粒子滤波对非线性代价函数 $\bar{L}(\theta)$ 进行优化，以确定声源的位置。

在笛卡尔坐标系下，设 k 时刻说话人的三维坐标为 $\boldsymbol{\chi}_k=(x_k,y_k,z_k)^{\mathrm{T}}$，相应的速度为 $\dot{\boldsymbol{\chi}}_k=(\dot{x}_k,\dot{y}_k,\dot{z}_k)^{\mathrm{T}}$。对于固定声源的定位问题，运动模型可用高斯噪声 $\boldsymbol{\omega}_k$ 驱动的随机游走模型加以描述，对于运动声源跟踪问题，本文使用Langevin 过程建立声源的运动模型[7]。假设说话人在 x 轴与 y 轴方向上的运动相互独立，且说话人声源的高度固定。在 x 轴方向上，说话人声源运动模型的状态方程可以描述为

$$\begin{bmatrix}x_k\\\dot{x}_k\end{bmatrix}=\begin{bmatrix}1&\Delta T\\0&a_\chi\end{bmatrix}\begin{bmatrix}x_{k-1}\\\dot{x}_{k-1}\end{bmatrix}+\begin{bmatrix}0\\b_x\end{bmatrix}u_x \tag{10}$$

其中 $\Delta T=L/f_s$，这里 L 为语音帧长度，f_s 是语音采样率；$a_x=e^{-\beta_x\Delta T}$，这里 β_x 为常数；$b_x=\upsilon_x\sqrt{1-a_x^2}$，$\upsilon_x$ 为稳态均方根速度；u_x 为单位方差的高斯白噪声。

粒子滤波的核心思想是利用一系列随机样本 $\boldsymbol{\chi}_k^{(i)}$ 及其对应权值 $w_k^{(i)}$ 来表示后验概率密度或滤波概率密度，即

$$p(\boldsymbol{\chi}_k\mid\boldsymbol{Y}_k)=\sum_{i=1}^{N}w_k^{(i)}\delta(\boldsymbol{\chi}_k-\boldsymbol{\chi}_k^{(i)}) \tag{11}$$

$$w_k^{(i)}\propto w_{k-1}^{(i)}\frac{p(\boldsymbol{Y}_k\mid\boldsymbol{\chi}_k^{(i)})p(\boldsymbol{\chi}_k^{(i)}\mid\boldsymbol{\chi}_{k-1})}{\pi(\boldsymbol{\chi}_k^{(i)}\mid\boldsymbol{\chi}_{1:k-1},\boldsymbol{Y}_k)} \tag{12}$$

本文采用系统状态转移函数 $p(\boldsymbol{\chi}_k\mid\boldsymbol{\chi}_{k-1})$ 作为重要性概率密度函数 $\pi(\cdot)$，根据系统的状态方程生成随机采样粒子 $\{\boldsymbol{\chi}_k^{(i)}\}_1^N$；并将归一化的目标函数 $\bar{L}(\theta)$ 作为似然函数 $p(\boldsymbol{Y}_k\mid\boldsymbol{\chi}_k^{(i)})$ 来计算粒子权值 $w_k^{(i)}$。于是，说话人位置 $\boldsymbol{\chi}_k$ 的最小均方误差估计为

$$\widehat{\boldsymbol{\chi}}_k=\int\boldsymbol{\chi}_k p(\boldsymbol{\chi}_k\mid\boldsymbol{Y}_k)\,\mathrm{d}\boldsymbol{\chi}_k=\frac{1}{N}\sum_{i=1}^{N}w_k^{(i)}\boldsymbol{\chi}_k^{(i)} \tag{13}$$

综上所述，现将基于粒子滤波的加权子空间拟合声源定位方法的具体步骤归纳如下：

(1)粒子集初始化：令所有粒子权值为$1/N$，N为粒子数；For $k=1,2,\cdots$

(2)对各路语音信号进行傅里叶变换，在每一个频率段构造空间协方差矩阵 $\boldsymbol{R}_{XX}(f_j)$；

(3)对协方差矩阵 $\boldsymbol{R}_{XX}(f_j)$ 进行特征值分解，估计信号子空间 $\widehat{U}_S(f_j)$ 与噪声方差 $\hat{\sigma}_n^2$；

(4)根据式(8)构造宽带信号的目标函数；

(5)采用粒子滤波算法对目标函数进行非线性全局优化：

(a)对粒子集进行重采样，并根据式(10)建立的声源运动模型产生新的采样粒子；

(b)将粒子状态转换为对应的时间延迟，以构造方向矩阵 $A_\theta(f_j)$；

(c)根据归一化的目标函数 $\bar{L}(\theta)$ 计算粒子权值，采用最小均方误差准则估计声源位置。

4 实验结果与分析

为了验证本文方法的有效性，在不同的信噪比与混响环境下进行了一系列仿真实验。仿真实验模拟了中小型会议室的声学环境，其中房间大小为 5m×4m×3m，房间的冲激响应函数由 IMAGE 模型产生。语音信号与高斯白噪声均以 44.1kHz 的采样率进行采样，将纯净语音与高斯白噪声按比例线性相加，生成不同信噪比的带噪语音。FFT 变换长度为 1024，相邻两帧重叠 50%，选用汉明(Hamming)窗函数。

麦克风阵列的摆放如图 1 所示，其中水平麦克风之间的距离为 D，垂直麦克风之间的距离为 H，声源方位角与俯仰角分别为 θ 与 ϕ，水平麦克风之间时间延迟为 τ_{01}，垂直麦克风之间时间延迟为 τ_{23}。在图 1 中，4 个麦克风的位置分别为 m_0=[0.50, 2.00, 0.88]，m_1=[0.50, 2.00, 1.12]，m_2=[0.50, −1.88, 1.00]，m_3=[0.50, 2.12, 1.00]，水平麦克风之间的距离与垂直麦克风之间的距离均为 0.24m，测试语音长度为 12s。

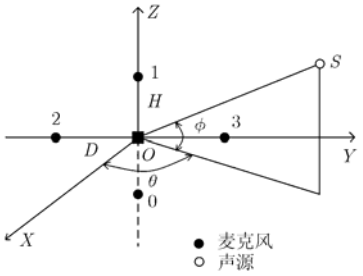

图 1 麦克风摆放示意图

实验1 声源定位仿真实验结果 在仿真实验中，声源的起始位置为 X_0=[4.0,0.0,1.5]，沿着y轴正向每0.5m进行一次估计。在粒子滤波算法中，粒子数为500，采用残差重采样(sidual resampling)降低退化现象的影响。在SNR=25dB, 混响时间为50ms与SNR=5dB，混响时间为200ms两种典型声学环境下，本文分别用文献[4]的SRP-PHAT方法、文献[5]的宽带MUSIC方法和本文方法对一段含有噪声与混响的说话人语音进行20次仿真实验，定位结果的均方根误差(RMSE)分别如表1和表2所示。

从表 1 可以看出，在信噪比较高、混响影响较弱的环境下，3 种方法都具有较好的定位能力。从表 2 可以看出，由于噪声干扰和混响的影响较严重(强噪声、强混响)，3 种方法都出现了较大的跟踪偏差。与其他两种方法相比，本文方法在背景噪声和房间混响均较强的情况下，跟踪偏差相对较小，这表明本文方法具有良好的抗噪声、抗混响能力。

实验 2 实际环境中的声源跟踪结果 本实验进一步比较两种方法在实际环境中的跟踪能力。说话人跟踪系统的设备包括 4 个全指向麦克风、信号放大器、多通道 A/D 转换器与多通道声卡。实验房间大小为 7.4m × 4.0m × 3.3m 。在本实验中，假设说话人在 x 轴与轴 y 方向上的运动相互独立，且说话人声源的高度固定为 z = 1.40m 。在 Langevin 模型中，$\beta_x = \beta_y = 10$，$\upsilon_x = \upsilon_y = 1$，粒子数为 100，采用残差重采样方法以降低退化现象对估计性能的影响，大约每 0.5s 给出一次跟踪结果。

3种方法在y方向上的跟踪结果如图2所示，其中横轴为时间，纵轴为y方向上说话人的位置，虚线为说话人的实际

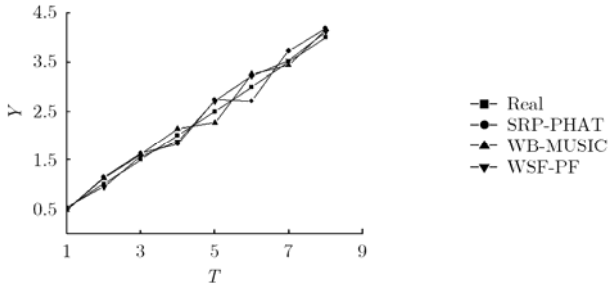

图 2 3 种方法的跟踪结果比较

表 1 SNR=20dB，混响时间 50ms 环境下的 RMSE 测试结果

实际位置(m)	0.00	0.50	1.00	1.50	2.00	2.50	3.00	3.50	4.00
SRP-PHAT 方法	0.091	0.086	0.073	0.061	0.044	0.065	0.071	0.083	0.093
宽带 MUSIC 方法	0.089	0.078	0.068	0.063	0.042	0.049	0.059	0.069	0.079
本文方法	0.071	0.062	0.052	0.048	0.039	0.046	0.054	0.061	0.069

表 2 SNR=5dB，混响时间 250ms 环境下的 RMSE 测试结果

实际位置(m)	0.00	0.50	1.00	1.50	2.00	2.50	3.00	3.50	4.00
SRP-PHAT 方法	0.464	0.447	0.413	0.404	0.389	0.411	0.424	0.451	0.476
宽带 MUSIC 方法	0.256	0.245	0.241	0.227	0.218	0.225	0.231	0.242	0.257
本文方法	0.249	0.227	0.218	0.204	0.201	0.212	0.219	0.231	0.246

运动轨迹，实线为3种方法的跟踪结果。实验结果表明，宽带MUSIC方法与SRP-PHAT方法在声源跟踪过程中有时会产生较大的跟踪误差，而本文方法将说话人的运动模型引入到声源跟踪问题中，可以有效抑制虚声源对跟踪性能的影响，其跟踪结果能较好地吻合说话人真实的运动轨迹，这表明本文方法在实际环境中仍具有较好的跟踪能力。

5　结束语

本文将传统的窄带加权子空间拟合方法推广至宽带情形，结合语音信号的宽带特性提出了一种基于粒子滤波的子空间拟合声源定位方法。该方法采用粒子滤波对宽带语音信号的非线性目标函数进行全局优化，从而有效地确定声源的空间位置。计算机仿真与实测结果表明，基于粒子滤波的子空间拟合声源定位方法具有良好的抗噪声与抗混响能力，可以为实际环境中的声源定位问题提供一种可行的解决方案。考虑到说话人运动模型在声源跟踪系统中的重要作用，采用交互多模型方法来进一步改进系统的定位与跟踪性能则是我们下一步将要开展的工作。

参 考 文 献

[1] Verma R C, Schmid C, and Mikolajczyk K. Face detection and tracking in a video by propagating detection probabilities[J]. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 2003, 25(10): 1215–1228.

[2] Mungamuru B and Aarabi P. Enhanced sound localization [J]. *IEEE Trans. on Systems, Man and Cybernetics*, 2004, 34(3): 1526–1540.

[3] Omologo M and Svaizer P. Use of the crosspower-spectrum phase in acoustic event location [J]. *IEEE Trans. on Speech and Audio Processing*, 1997, 5(3): 288–292.

[4] DiBiase J. A high-accuracy, low-latency technique for talker localization in reverberant environments [D]. Brown University, Providence RI, USA, 2000.

[5] 居太亮, 彭启琮, 邵怀宗等. 基于任意麦克风阵列的声源二维DOA 估计算法研究[J]. 通信学报, 2005, 26(8): 129–133.
Ju Tai-liang, Peng Qi-cong, and Shao Huai-zong, *et al.*. Speech source 2D DOA estimation algorithm based on random microphone array. *Journal on Communications*, 2005, 26(8): 129–133.

[6] Chen J C, Yao K, and Hudson R E. Acoustic source localization and beamforming: theory and practice [J]. *EURASIP Journal on Applied Signal Processing*, 2003, 2003(4): 359–370.

[7] Ward D B, Williamson R C, and Lehmann E A. Particle filtering algorithms for tracking an acoustic source in a reverberant environment [J]. *IEEE Trans. on Speech and Audio Processing*, 2003, 11(6): 826–836.

[8] Di Claudio E and Parisi R. Multi-Source Localization Strategies. Microphone Arrays: Signal Processing Techniques and Applications [M]. M. Brandstein and D. Ward, Eds., Boston, London: Springer, 2001: 181–201.

[9] Arulampalam M, Maskell S, and Gordon N, *et al.*. A tutorial on particle filters for on-line nonlinear/non-Gaussian Bayesian tracking [J]. *IEEE Trans. on Signal Processing*, 2002, 50(2): 174–188.

金乃高：　男，1977 年生，博士生，研究方向为语音处理、信息融合等.

殷福亮：　男，1962 年生，教授，博士生导师，主要研究方向为语音处理、宽带无线通信技术等.

陈　喆：　男，1975 年生，副教授，主要研究方向为语音处理、宽带无线通信技术等.