电子与信息工程、管理工程

文章编号：1000-8608(2009)04-0580-07

基于分层采样粒子滤波的说话人跟踪方法

侯代文[1,2]， 殷福亮*[1]， 陈 喆[1]

（1.大连理工大学 电子与信息工程学院，辽宁 大连 116024；
2.海军试验基地，辽宁 大连 116041）

摘要：利用分层采样方法，融合波达方向和时间延迟两种信息，实现了对说话人的定位与跟踪.分层采样方法考虑波达方向和时间延迟这两种不同观测信息对说话人位置估计精度的差异，将基于波达方向滤波得到的状态后验概率密度函数作为基于时间延迟滤波的重要性采样函数，增强了重要性概率密度函数与后验概率密度函数的相似程度，从而改善了重要性概率密度函数的质量，减小了采样粒子权值的方差，提高了对说话人位置的估计精度.仿真实验验证了该方法的有效性.

关键词：说话人跟踪；粒子滤波；波达方向估计；时间延迟估计；分层采样
中图分类号：TN713 **文献标志码**：A

0 引 言

说话人语音定位与跟踪问题是语音信号处理领域的重要课题之一，它可以广泛应用于电视电话会议系统、视频监控系统中的摄像头自动导引、远距离说话人语音识别、计算机人机接口以及机器人导航等场合[1].

说话人定位与跟踪是根据麦克风阵列接收到的说话人语音信息以及说话人的运动规律，实时估计说话人位置的技术.经常使用的定位方法主要有波束形成方法和时延估计方法[2,3].波束形成方法通过改变麦克风阵列的指向模式，将各麦克风接收到的信号“导向”某一方向，然后在信号空间内搜索能够使期望信号输出功率最大的方向，就认为是说话人所在方向.时延估计方法首先确定一组麦克风对之间的时间延迟，再通过求解一组非线性双曲面方程，得到说话人位置.上述两种方法在自由声场条件下，都能够实现对说话人的准确定位.然而在实际应用中，由于房间混响、噪声干扰等因素的影响，有可能产生虚声源，此时采用上述方法，会导致对说话人位置的错误估计.Sturim 等[4]提出利用状态空间方法解决这一问题，该方法通过建立动态方程，在估计说话人当前位置时，不仅利用当前观测信息，而且利用当前时刻之前的全部信息，因此能够滤除观测序列中具有明显误差的观测量，从而在一定程度上解决了说话人跟踪中的虚声源问题.以此为基础，Dvorkind等[5]利用卡尔曼滤波器跟踪说话人位置，解决了虚声源问题.然而，在非高斯观测噪声条件下，使用卡尔曼滤波方法所得估计结果偏差较大.考虑到粒子滤波方法具有较强的处理非线性、非高斯问题的能力，Vermaak 等[6]和 Ward 等[7]采用粒子滤波器[8,9]进行说话人跟踪，改善了说话人跟踪的效果.但这两种方法均选用先验分布作为重要性概率密度函数，由于粒子采样效率低，常常会出现粒子匮乏现象，导致状态估计精度降低.

为了实现对说话人的准确定位与跟踪，本文提出基于波达方向和时间延迟的说话人联合跟踪粒子滤波方法，并利用仿真实验对本文方法的有效性进行验证.

1 粒子滤波方法

考虑说话人跟踪问题，系统状态方程和观测

收稿日期：2007-07-02； **修回日期**：2009-05-13.
基金项目：国家自然科学基金资助项目(60772161,60372082)；高等学校博士学科点专项科研基金资助项目(200801410015).
作者简介：侯代文(1972-)，男，博士，E-mail:hodevin@gmail.com；殷福亮*(1962-)，男，教授，博士生导师，E-mail:flyin@dlut.edu.cn.

方程可以描述为

$$\boldsymbol{x}_k = f(\boldsymbol{x}_{k-1}, \boldsymbol{v}_{k-1}) \tag{1}$$

$$\boldsymbol{z}_k = h(\boldsymbol{x}_k, \boldsymbol{n}_k) \tag{2}$$

式中：$\boldsymbol{x}_k = (x_k \quad y_k \quad v_{x_k} \quad v_{y_k})^{\mathrm{T}}$，表示说话人状态，它包含说话人的位置 x_k、y_k 及速度 v_{x_k}、v_{y_k}；f 是状态转移函数；$\{\boldsymbol{v}_{k-1}, k \in \mathbf{N}\}$ 是独立同分布的过程噪声序列；$\boldsymbol{z}_k$ 为观测向量，它可以是时间延迟，也可以是波达方向；h 是观测函数；$\{\boldsymbol{n}_k, k \in \mathbf{N}\}$ 是独立同分布的观测噪声序列. 本文的目标是通过观测序列 $\boldsymbol{z}_{1:k} = \{\boldsymbol{z}_i, i = 1, \cdots, k\}$，得到说话人状态 $\boldsymbol{x}_k$ 的最优估计，从而实现说话人跟踪.

贝叶斯状态估计方法通过求解后验概率密度函数 $p(\boldsymbol{x}_k \mid \boldsymbol{z}_{1:k})$，根据方程

$$E(\boldsymbol{x}_k \mid \boldsymbol{z}_{1:k}) = \int \boldsymbol{x}_k p(\boldsymbol{x}_k \mid \boldsymbol{z}_{1:k}) \mathrm{d}\boldsymbol{x}_k \tag{3}$$

求得状态向量 $\boldsymbol{x}_k$ 的最优估计.

粒子滤波方法，也称序贯 Monte Carlo 方法，它通过 Monte Carlo 模拟，实现状态的贝叶斯递推估计[8]. 其核心思想是：用一组随机采样点及其对应的权值表示所需的后验概率密度函数，从而计算状态估计值. 当采样点个数趋于无穷大时，Monte Carlo 模拟的概率密度函数等价于后验概率密度函数，相应的状态估计值接近于最优的贝叶斯估计.

令 $\{\boldsymbol{x}_k^i, i = 1, \cdots, N_s\}$ 表示一支撑点集，对应的权值为 $\{w_k^i, i = 1, \cdots, N_s\}$，其中权值 w_k^i 满足归一化条件 $\sum_i w_k^i = 1$. 用 $\{\boldsymbol{x}_k^i\}_{i=1}^{N_s}$ 表示描述后验概率密度函数 $p(\boldsymbol{x}_k \mid \boldsymbol{z}_{1:k})$ 的随机采样点集合，则 k 时刻的后验概率密度函数可以表示为

$$p(\boldsymbol{x}_k \mid \boldsymbol{z}_{1:k}) \approx \sum_{i=1}^{N_s} w_k^i \delta(\boldsymbol{x}_k - \boldsymbol{x}_k^i) \tag{4}$$

其中 $\delta(\cdot)$ 为单位冲激响应函数.

如果支撑点集 $\{\boldsymbol{x}_k^i, i = 1, \cdots, N_s\}$ 由重要性概率密度函数 $q(\boldsymbol{x})$ 抽样得到，其相应的权值由以下公式计算：

$$w_k^i = \frac{p(\boldsymbol{x}_k^i \mid \boldsymbol{z}_{1:k})}{q(\boldsymbol{x}_k^i \mid \boldsymbol{z}_{1:k})} \tag{5}$$

根据贝叶斯原理，可以得到粒子权值的递推估计形式为

$$w_k^i \propto w_{k-1}^i \frac{p(\boldsymbol{z}_k \mid \boldsymbol{x}_k^i) p(\boldsymbol{x}_k^i \mid \boldsymbol{x}_{k-1}^i)}{q(\boldsymbol{x}_k^i \mid \boldsymbol{x}_{k-1}^i, \boldsymbol{z}_k)} \tag{6}$$

在粒子滤波方法中，重要性概率密度函数 $q(\boldsymbol{x})$ 的选取具有重要意义. 重要性概率密度函数与后验概率密度函数的逼近程度，直接决定重要性采样的效率，从而决定状态估计的精度. Doucet 等[10] 已经证明，最优的重要性概率密度函数为 $q(\boldsymbol{x}_k \mid \boldsymbol{x}_{k-1}, \boldsymbol{z}_{1:k}) = p(\boldsymbol{x}_k \mid \boldsymbol{x}_{k-1}, \boldsymbol{z}_k)$. 但在实际应用中，$p(\boldsymbol{x}_k \mid \boldsymbol{x}_{k-1}, \boldsymbol{z}_k)$ 的获取与直接从后验概率密度函数中抽取样本同样困难，因而只能寻求次优的重要性概率密度函数.

当多个观测信息可以利用，而且各信息对状态估计的贡献存在程度差异时，分层采样方法提供了改善重要性概率密度函数的途径. 分层采样方法[11] 利用重要性重采样 (importance resampling) 的思想，复制代表先验概率分布 $p_0(\boldsymbol{x})$ 的粒子并重新计算权值. 当重要性概率密度函数 $f_1, f_2, \cdots, f_M$ 存在，而且任意一个 f_{m-1} 近似于 f_m 但不如 f_m 准确时，可以考虑使用分层采样方法. 通常情况下，分层采样方法相对于普通的粒子滤波方法并没有优越性，但在多个观测量 $\boldsymbol{z}^1$，$\boldsymbol{z}^2, \cdots, \boldsymbol{z}^M$ 可以利用，而且不同的观测量对状态估计所提供的信息量不同时，该方法可以将上一层的最新观测信息融入下一层滤波过程的重要性概率密度函数中，得到比状态转移概率密度函数更接近后验概率密度函数的采样函数，因而能提高粒子的采样效率. 此时，按照各个观测量所估计状态精度的不同排序，分层采样方法在状态空间高效搜索，下一层的状态估计结果将优于上一层的状态估计结果. 能够融合 M 个观测量的分层采样粒子滤波算法如下.

设 $\{\boldsymbol{x}_{k-1}^i, w_{k-1}^i\}_{i=1}^{N_s}$ 为 $k-1$ 时刻的粒子及相应的权值集合，在时刻 k

(1) 初始化

$$\{\boldsymbol{x}^{0,i}, w^{0,i}\}_{i=1}^{N_s} = \{\boldsymbol{x}_{k-1}^i, w_{k-1}^i\}_{i=1}^{N_s}$$

(2) 循环

For $m = 1:M$

① 生成粒子　$\boldsymbol{x}^{m,i} \sim q_m(\boldsymbol{x}^m \mid \boldsymbol{x}^{m-1,i}, \boldsymbol{z}_k^m)$

② 权值更新

$$w^{m,i} \propto w^{m-1,i} \frac{p(\boldsymbol{z}_k^m \mid \boldsymbol{x}^{m,i}) p_m(\boldsymbol{x}^{m,i} \mid \boldsymbol{x}^{m-1,i})}{q_m(\boldsymbol{x}^{m,i} \mid \boldsymbol{x}^{m-1,i}, \boldsymbol{z}_k^m)}$$

③ 重采样　生成 $a_i \sim \{w^{m,i}\}_{i=1}^{N_s}$，替换 $\{\boldsymbol{x}^{m,i}, w^{m,i}\}_{i=1}^{N_s} \leftarrow \{\boldsymbol{x}^{m,a_i}, 1/N_s\}$

(3) 结束

$$\{\boldsymbol{x}_k^i, w_k^i\}_{i=1}^{N_s} = \{\boldsymbol{x}_k^{M,i}, w_k^{M,i}\}_{i=1}^{N_s}$$

2　基于粒子滤波的麦克风阵列说话人跟踪方法

麦克风阵列由处于不同空间位置并按一定几

何结构排列的若干个麦克风构成，如图 1 所示. 根据各个麦克风获取的语音信号，可以得到语音相对于阵列所在方向信息以及语音到达不同麦克风的时间延迟信息.

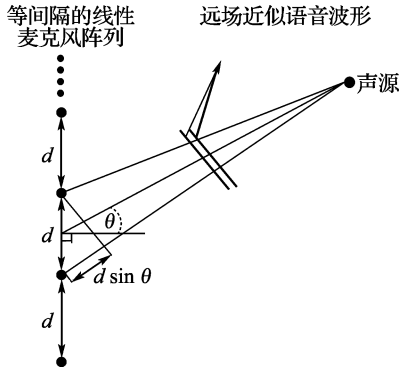

图 1 麦克风阵列说话人定位示意图

Fig. 1 Schematic diagram of microphone array speaker localization

2.1 声源波达方向估计

麦克风阵列波束形成方法通过对各个麦克风的输出信号进行加权求和，调整阵列响应函数，控制阵列响应的波束主瓣指向某一方向并计算输出功率，输出功率最大的方向就认为是说话人所在方向[12].

对于由 L 个麦克风组成的阵列，设 k 时刻第 l 个麦克风接收到的信号为 $x_{k,l}$，各麦克风的加权系数为 w_l，则阵列输出为

$$y_k = \sum_{l=1}^{L} w_l^* x_{k,l} = \boldsymbol{w}^{\mathrm{H}} \boldsymbol{x}_k \tag{7}$$

其中 $\boldsymbol{x}_k = (x_{k,1} \quad x_{k,2} \quad \cdots \quad x_{k,L})^{\mathrm{T}}$，$\boldsymbol{w} = (w_1 \quad w_2 \quad \cdots \quad w_L)^{\mathrm{T}}$. 于是，输出功率可以表示为

$$P(\boldsymbol{w}) = \frac{1}{N}\sum_{k=1}^{N} |\ y_k\ |^2 = \frac{1}{N}\sum_{k=1}^{N} \boldsymbol{w}^{\mathrm{H}} \boldsymbol{x}_k \boldsymbol{x}_k^{\mathrm{H}} \boldsymbol{w} = \boldsymbol{w}^{\mathrm{H}} \hat{\boldsymbol{R}} \boldsymbol{w} \tag{8}$$

其中 $\hat{\boldsymbol{R}} = \dfrac{1}{N}\sum_{k=1}^{N} \boldsymbol{x}_k \boldsymbol{x}_k^{\mathrm{H}}$.

对于基本的波束形成方法，加权矢量可表示为 $\boldsymbol{w} = \boldsymbol{a}(\theta) = (1 \quad \mathrm{e}^{\mathrm{j}\phi} \quad \cdots \quad \mathrm{e}^{\mathrm{j}(L-1)\phi})^{\mathrm{T}}$. 对等距线性阵列，参数 $\phi = -2\pi d \cos\theta/\lambda$，其中 λ 为语音信号波长，d 为相邻麦克风之间的距离，θ 为声波波达方向. 对加权矢量 $\boldsymbol{w}$ 作归一化处理后，得到空间谱分布函数为

$$P_{\mathrm{bf}}(\theta) = \frac{\boldsymbol{a}^{\mathrm{H}}(\theta)\hat{\boldsymbol{R}}\boldsymbol{a}(\theta)}{\boldsymbol{a}^{\mathrm{H}}(\theta)\boldsymbol{a}(\theta)} \tag{9}$$

由空间谱分布函数，可以求得说话人方向为

$$\hat{\theta} = \arg\max_{\theta} P_{\mathrm{bf}}(\theta) \tag{10}$$

求解出说话人方向以后，就可以实现对说话人的实时跟踪. 设麦克风阵列中心位置为 $(x_{\mathrm{a}}, y_{\mathrm{a}})$，在任一时刻 k，说话人位置为 (x_k, y_k)，说话人方向为 θ_k. 在如图 1 中，观测方程(2)的具体形式可表示为

$$\theta_k = \arctan\frac{y_k - y_{\mathrm{a}}}{x_k - x_{\mathrm{a}}} \tag{11}$$

方程(1)和(11)构成说话人跟踪系统的状态方程和观测方程.

2.2 声源时间延迟估计

在说话人定位系统中，由于各麦克风所处的位置不同，语音到达各麦克风的时间存在差异，本文称这种差异为两个麦克风之间的时间延迟. 广义互相关法(GCC)是广泛使用的时延估计方法，该方法通过求解两个麦克风接收信号互相关函数的极值点来估计时间延迟.

设声源信号为 $s(k)$，两个麦克风接收到的信号分别为

$$y_1(k) = s(k-\tau_1) + m_1(k) \tag{12}$$

$$y_2(k) = \alpha s(k-\tau_2) + m_2(k) \tag{13}$$

其中 α 是声源信号的衰减系数，τ_1 和 τ_2 是声源到两麦克风的传播时间，信号 $s(k)$ 与噪声 $m_1(k)$ 和 $m_2(k)$ 均不相关.

设信号 $y_1(k)$ 和 $y_2(k)$ 的互功率谱为 $\boldsymbol{S}_{y_1y_2}(f) = E[\boldsymbol{Y}_1(f)\boldsymbol{Y}_2^*(f)]$，加权函数为 $\boldsymbol{\varphi}(f)$，广义互功率谱为 $\boldsymbol{\Psi}_{y_1y_2}(f) = \boldsymbol{\varphi}(f)\boldsymbol{S}_{y_1y_2}(f)$，则广义互相关函数可表示为

$$\boldsymbol{\Psi}_{y_1y_2}(\tau) = \int_{-\infty}^{+\infty} \boldsymbol{\varphi}(f)\boldsymbol{S}_{y_1y_2}(f)\mathrm{e}^{\mathrm{j}2\pi f\tau}\mathrm{d}f = \int_{-\infty}^{+\infty} \boldsymbol{\Psi}_{y_1y_2}(f)\mathrm{e}^{\mathrm{j}2\pi f\tau}\mathrm{d}f \tag{14}$$

这样，两麦克风之间的声波到达时间差 $\tau = \tau_1 - \tau_2$ 对应于广义相关函数的极值点位置，即

$$\hat{\tau} = \arg\max_{\tau} \boldsymbol{\Psi}_{y_1y_2}(\tau) \tag{15}$$

在式(14)中，选取不同的加权函数 $\boldsymbol{\varphi}(f)$，可以得到不同的时延估计方法. 在相位变换(phase transform)方法中[13]，加权函数取为 $\boldsymbol{\varphi}(f) = 1/|\ \boldsymbol{S}_{y_1y_2}(f)\ |$. 该方法通过对互功率谱预白化，去除了与时延无关的互功率谱幅度信息而仅保留相位特性，能够锐化函数 $\boldsymbol{\Psi}_{y_1y_2}(\tau)$ 的极值，从而较好地抑制噪声和混响对时延估计带来的影响.

得到时间延迟估计后，就可以对说话人进行实时跟踪. 设在某一时刻 k 得到的时延估计为 $\tau_{k,i}$ $(i=1,2)$，声音在空气中的传播速度为 c，相应的两个麦克风位置分别为 $\boldsymbol{r}_{i,1}=(x_{i,1}\quad y_{i,1})$ 和 $\boldsymbol{r}_{i,2}=(x_{i,2}\quad y_{i,2})$，说话人位置为 $\boldsymbol{r}_k=(x_k\quad y_k)$，则时延 τ_k 与说话人位置之间的关系为

$$\tau_{k,i}=c^{-1}[\,|\,\boldsymbol{r}_{i,1}-\boldsymbol{r}_k\,|-|\,\boldsymbol{r}_{i,2}-\boldsymbol{r}_k\,|\,];\ i=1,2 \tag{16}$$

即

$$\tau_{k,i}=c^{-1}[\sqrt{(x_{i,1}-x_k)^2+(y_{i,1}-y_k)^2}-\sqrt{(x_{i,2}-x_k)^2+(y_{i,2}-y_k)^2}\,] \tag{17}$$

方程(1)和(17)构成说话人跟踪系统的状态方程和观测方程.

2.3　分层采样融合算法

利用说话人波达方向或者时间延迟作为观测量，通过粒子滤波方法，均可实现对说话人的实时跟踪，但是跟踪效果并不理想. 利用波束形成方法确定说话人方位时，由于方位角必须从一系列离散的角度值中选取，精度较低. 这样，以方位角为观测量的滤波算法难以对说话人准确跟踪. 以时间延迟作为观测量的跟踪算法虽然精度有所提高，但它选取系统状态的转移概率分布作为重要性概率密度函数，没有考虑系统状态的最新观测，所产生的样本方差较大，难以准确描述系统状态的条件概率密度函数. 对于说话人跟踪系统，由于说话人运动具有随机性，先验分布近似为均匀分布，该问题更加突出.

在粒子滤波中，选取合适的重要性概率密度函数，使之产生的预测样本接近于真实的后验分布产生的样本，能提高样本的采样效率. 为了提高说话人的跟踪精度，本文充分利用所得到的观测信息，对得到的每一帧语音信号，均同时估计波达方向和时间延迟两种观测信息，然后采用分层采样方法，融合两种观测信息，即首先利用波达方向粗略估计说话人位置，然后利用该估计结果，为基于时间延迟的跟踪方法提供重要性概率密度函数. 因为所采用的重要性概率密度函数融合了最新的波达方向信息，比状态转移概率密度函数更接近真实的后验概率密度函数，因而该方法能提高采样效率，减少预测样本重要性权值的方差，提高跟踪精度.

在任意时刻，同时利用得到的波达方向与时间延迟两种观测量，可以将跟踪算法分两步进行. 首先，利用波达方向作状态估计，由状态转移概率 $p(\boldsymbol{x}_k\mid\boldsymbol{x}_{k-1}^i)$ 抽样得到粒子集合 $\{\boldsymbol{x}_k^{\mathrm{b},i}\}_{i=1}^{N_s}$，并利用方位角观测信息计算其对应的权值 $\{w_k^{\mathrm{b},i}\}_{i=1}^{N_s}$. 用这组粒子逼近状态分布的后验概率密度函数，作为基于时间延迟状态估计的重要性概率密度函数. 然后，利用时间延迟观测量，对状态做更准确的估计，并将其估计结果作为最终的滤波值. 具体算法如下.

设 $\{\boldsymbol{x}_{k-1}^i,w_{k-1}^i\}_{i=1}^{N_s}$ 是 $k-1$ 时刻的粒子及相应的权值集合，在时刻 k

(1) 由式(10)求得 k 时刻的方位角 θ_k，应用方位角观测量进行滤波：

①For $i=1:N_s$

采样　$\boldsymbol{x}_k^{\mathrm{b},i}\sim p(\boldsymbol{x}_k\mid\boldsymbol{x}_{k-1}^i)$

权值更新　$w_k^{\mathrm{b},i}\propto w_{k-1}^i\cdot p(\theta_k\mid\boldsymbol{x}_k^{\mathrm{b},i})$，满足 $\sum_{i=1}^{N_s}w_k^{\mathrm{b},i}=1$

End for

② 计算粒子集合 $\{\boldsymbol{x}_k^{\mathrm{b},i},w_k^{\mathrm{b},i}\}_{i=1}^{N_s}$ 的经验性协方差阵 $\boldsymbol{S}_k$，并作 Cholesky 展开，使 $\boldsymbol{D}_k\boldsymbol{D}_k^{\mathrm{T}}=\boldsymbol{S}_k$

③ 重采样

$\{\boldsymbol{x}_k^{\mathrm{b},i},w_k^{\mathrm{b},i}\}_{i=1}^{N_s}=Resample\,[\{\boldsymbol{x}_k^{\mathrm{b},i},w_k^{\mathrm{b},i}\}_{i=1}^{N_s}]$

(2) 由式(15)求得 k 时刻的时间延迟 τ_k，应用时间延迟观测量滤波：

①For $i=1:N_s$

采样　$\varepsilon_k^i\sim K_{\mathrm{Epan}},\boldsymbol{x}_k^{\mathrm{t},i}=\boldsymbol{x}_k^{\mathrm{b},i}+h_{\mathrm{opt}}\boldsymbol{D}_k\varepsilon_k^i$

权值更新　$w_k^{\mathrm{t},i}\propto w_{k-1}^{\mathrm{t},i}\times[p(\tau_k\mid\boldsymbol{x}_k^{\mathrm{t},i})\times p(\boldsymbol{x}_k^{\mathrm{t},i}\mid\boldsymbol{x}_{k-1}^{\mathrm{t},i})]/p(\varepsilon_k^i)$，满足 $\sum_{i=1}^{N_s}w_k^{\mathrm{t},i}=1$

End for

② 重采样　$\{\boldsymbol{x}_k^i,w_k^i\}_{i=1}^{N_s}=Resample[\{\boldsymbol{x}_k^{\mathrm{t},i},w_k^{\mathrm{t},i}\}_{i=1}^{N_s}]$

其中 K_{Epan} 为 Epanechikov 核函数[14]，重采样过程应用了 Carpenter 等[15]提出的系统重采样方法.

3　计算机仿真与实验结果

本文融合波达方向和时间延迟两种观测信息，进行说话人跟踪仿真实验，并与单独利用时间延迟信息作说话人跟踪[6]的仿真结果进行了比较.

这里考虑二维空间上的说话人跟踪问题，共设计了两种跟踪方案. 如图 2 所示，在一个 3 m×3 m 房间内的 X、Y 两个方向上，分别放置两组各

包含 10 个麦克风的线性阵列，相邻麦克风之间的距离为 6 cm. 在方案一中，说话人由点(0.5,1.5)开始沿半圆形轨迹运动，停止于点(2.5,1.5)；在方案二中，说话人由点(0.5,0.5)开始沿折线运动，分别在点(0.5,1.1)、(2.5,1.1)向右转 90°，终止于点(2.5,0.5). 使用的说话人语音信号如图 3 所示，麦克风接收到的信号用 IMAGE 模型[16]仿真获得，信噪比为 30 dB. 对语音信号每隔 32 ms 采集一帧数据，用波束形成方法估计波达方向，同时用 PHAT 方法估计时间延迟. 考虑到利用波达方向对说话人位置的估计，只是为基于时间延迟的跟踪过程提供重要性采样函数，对波达方向精度要求不高，同时为了减少计算量，以 1.5°为单位进行角度扫描，所得角度测量误差约为 3°.

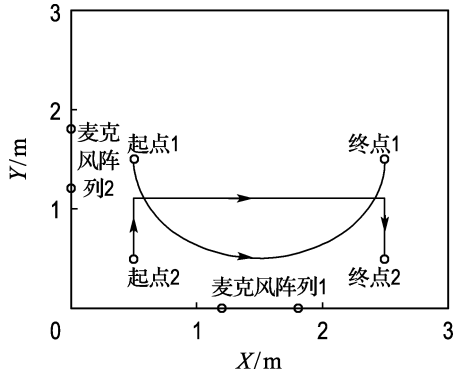

图 2 说话人跟踪中的运动轨迹设计

Fig. 2 Trajectory setup in speaker tracking

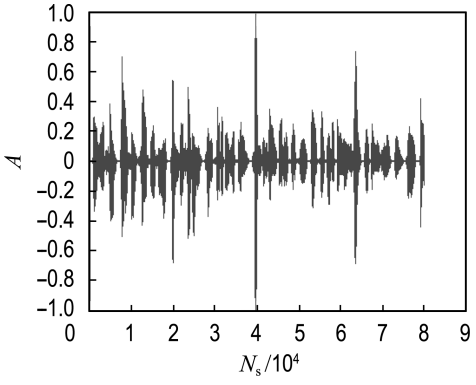

图 3 仿真实验中的语音信号

Fig. 3 Speech signal in simulation

由于说话人运动具有随意性，难以用一种简单的运动模型来描述，本文采用适应性较强的布朗运动模型，对应的动态方程为朗之万(Langevin)过程[6]，其离散形式为

$$\boldsymbol{x}_k = \boldsymbol{A}\boldsymbol{x}_{k-1} + \boldsymbol{\Gamma}v_k \tag{18}$$

式中：$\boldsymbol{x}_k = (x_k \quad y_k \quad v_{x_k} \quad v_{y_k})^{\mathrm{T}}$，表示说话人状态，包含说话人的位置 x_k、y_k 及速度 v_{x_k}、v_{y_k}；$\boldsymbol{A} = \begin{pmatrix} \boldsymbol{B} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{B} \end{pmatrix}$，为状态转移矩阵，其中 $\boldsymbol{B} = \begin{pmatrix} 1 & T \\ 0 & a \end{pmatrix}$，$T = 32$ ms；$\boldsymbol{\Gamma} = (0 \quad \beta \quad 0 \quad \beta)^{\mathrm{T}}$；$v_k \sim N(0, Q)$，$Q = 1$. α 和 β 的取值由实际的应用条件训练获得，在本文中，取 $\alpha = 0.968\,5$，$\beta = 0.249\,0$. 两种跟踪方案的 Monte Carlo 仿真次数都是 50.

仿真实验结果如图 4～7 所示，其中图 4 和 6 为说话人分别沿圆形轨迹运动和折线运动时，在 X 轴和 Y 轴上的跟踪结果，图 5 和 7 为各自对应

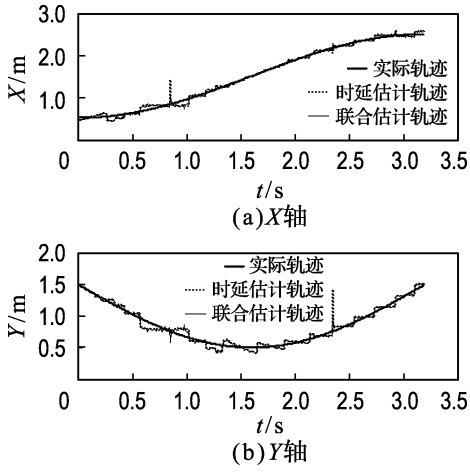

图 4 两种跟踪方法对圆形运动说话人的跟踪效果

Fig. 4 Tracking performances of the two methods with circular trajectory

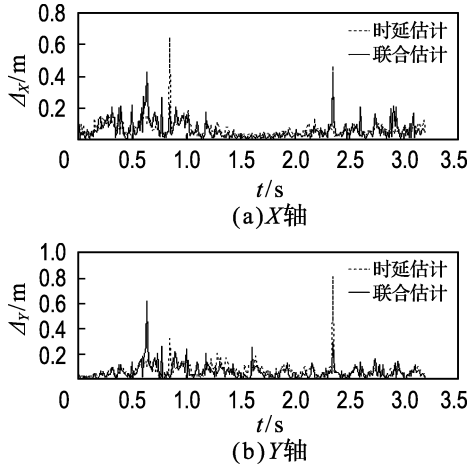

图 5 说话人作圆形运动时两种跟踪方法的误差比较

Fig. 5 Tracking errors of the two methods with circular trajectory

的跟踪误差. 结果显示，在两种运动形式下，两种跟踪方法都能快速锁定目标，并以较高的精度跟踪目标位置. 从图 4 可以看出，在初始运动阶段，跟踪效果较差，这是由此时说话人所处位置的可观测性较低所致. 在图 6 中，运动轨迹的折点处跟踪误差较大，这与此时运动形式变化剧烈，而朗之万过程不能准确描述这种快速变化有关. 在这两种跟踪方案中，本文方法的跟踪精度明显高于单独利用时间延迟一种观测信息的跟踪精度，而且收敛速度更快. 图 5 和 7 所示的跟踪误差同样表明了本文联合跟踪方法的优越性.

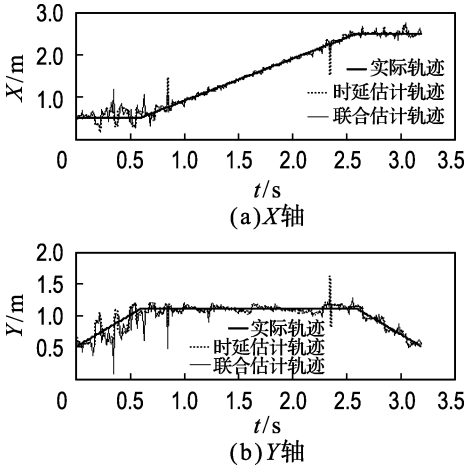

图 6　两种跟踪方法对折线运动说话人的跟踪效果

Fig. 6　Tracking performances of the two methods with zigzag trajectory

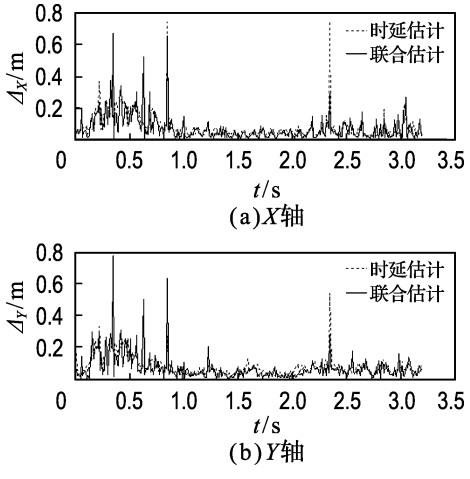

图 7　说话人作折线运动时两种跟踪方法的误差比较

Fig. 7　Tracking errors of the two methods with zigzag trajectory

为了对滤波器的估计精度作定量比较，本文定义一次独立实验的经验性标准偏差[10]为

$$R_{\mathrm{mse}}=\frac{1}{N}\sum_{k=1}^{N}\left[\frac{1}{M}\sum_{m=1}^{M}(\hat{x}_{k,m}-x_k)^2\right]^{1/2} \quad (19)$$

其中 N 为滤波过程中的迭代次数，M 是 Monte Carlo 仿真次数，$\hat{x}_{k,m}$ 是第 m 次仿真实验中状态 x 在 k 时刻的滤波估计值. 针对以上两种不同的跟踪方案，计算它们的经验性标准偏差如表 1 所示. 由表 1 可以看出，在两种运动方案中，本文的联合跟踪方法，在 X 轴和 Y 轴两个方向上对说话人的跟踪精度，均优于仅利用时间延迟估计的粒子滤波方法.

表 1　两种跟踪方案的经验性标准偏差

Tab. 1　Empirical standard deviation of the two tracking scenario　　m

跟踪滤波方法	圆形轨迹运动		折线轨迹运动	
	X 轴	Y 轴	X 轴	Y 轴
仅用时间延迟跟踪方法	0.057 3	0.057 7	0.066 1	0.067 9
联合跟踪方法	0.045 9	0.051 5	0.059 6	0.062 3

4　结　语

本文提出了基于波达方向和时间延迟的说话人联合跟踪粒子滤波方法. 该方法融合波达方向和时间延迟两种观测信息，应用分层采样方法，增强了粒子滤波方法中重要性概率密度函数与后验概率密度函数的相似度，提高了说话人的跟踪精度. 基于粒子滤波的分层采样方法，对多信息源且各信息源对状态估计的贡献存在较大差异的情况，提供了有效的数据融合方式. 特别应该指出的是，该方法对观测数据的融合不需要考虑其统计独立性，因此，在各观测信息之间统计独立性不易确定时，该方法同样适用.

参考文献：

[1] QIN J. Robust speaker tracking [D]. Pittsburgh: Carnegie Mellon University, 2007

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

[3] CHEN J, BENESTY J, HUANG Y. Time delay estimation in room acoustic environments: an overview [J]. **EURASIP Journal on Applied Signal**

Processing, 2006(1):1-19

[4] STURIM D, BRANDSTEIN M, SILVERMAN H. Tracking multiple talkers using microphone array measurements [C] // **Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing**. Germany:IEEE, 1997

[5] DVORKIND T, GANNOT S. Speaker localization exploiting spatial-temporal information [C] // **Proceedings of the IEEE International Workshop on Acoustic Echo and Noise Control**. Japan:IWAENC, 2003

[6] VERMAAK J , BLAKE A. Nonlinear filtering for speaker tracking in noisy and reverberant environments [C] // **Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing**. Piscataway:IEEE, 2001

[7] WARD D B, LEHMANN E A, WILLIAMSON R C. Particle filtering algorithms for tracking an acoustic source in a reverberant environment [J]. **IEEE Transactions on Speech and Audio Processing**, 2003, **11**(6):826-836

[8] GORDAN N, SALMOND D, SMITH A F M. Novel approach to nonlinear and non-Gaussian Bayesian state estimation [J]. **IEEE Proceedings on Radar and Signal Processing**, 1993, **140**(2):107-113

[9] CAPPE O, GODSILL S J, MOULINES E. An overview of existing method and recent advances in sequential Monte Carlo [J]. **Proceedings of the IEEE**, 2007, **95**(5):899-924

[10] DOUCET A, GODSILL S, ANDRIEU C. On sequential Monte Carlo sampling methods for Bayesian filtering [J]. **Statistical Computation**, 2000, **10**(3):197-208

[11] SULLIVAN J, BLAKE A, ISARD M, *et al*. Object localization by Bayesian correlation [C] // **Proceeding of International Conference on Computer Vision**. Greece:ICCV, 1999

[12] KRIM H, VIBERG M. Two decades of array signal processing research:the parametric approach [J]. **IEEE Signal Processing Magazine**, 1996, **13**(4):67-94

[13] 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

[14] SILVERMAN B W. **Density Estimation for Statistics and Data Analysis** [M]. London:Chapman & Hall, 1986:75-88

[15] CARPENTER J, CLIFFORD P , FEARNHEAD P. Improved particle filter for nonlinear problems [J]. **IEEE Proceeding on Radar, Sonar, Navigation**, 1999, **146**(1):2-7

[16] ALLEN J B, BERKLEY D A. Image method for efficiently simulating small-room acoustics [J]. **Journal of the Acoustical Society of America**, 1979, **65**(4):943-950

Speaker tracking method using layered sampling particle filter

HOU Dai-wen[1,2], YIN Fu-liang[*1], CHEN Zhe[1]

(1. School of Electronic and Information Engineering, Dalian University of Technology, Dalian 116024, China;
2. Naval Test Base, Dalian 116041, China)

Abstract: Utilizing layered sampling method, both direction of arrival (DOA) and time difference of arrival (TDOA) of speech source are fused to localize and track the speaker. Since the measurement modalities differ in the level of information which they provide about the state, the layered sampling method constructs an informed proposal by integrating filtering results from DOA measurement, then particles are sampled from this proposal in TDOA measurement based particle filter. As the similarity between the importance density function and the posterior density function is enhanced, the quality of the proposal function is improved and variance of sample weights is decreased, thus the speaker localization accuracy is improved. Simulation results of two scenarios show the validity of the proposed method.

Key words: speaker tracking; particle filter; DOA estimation; TDOA estimation; layered sampling