基于粒子滤波的交互式多模型说话人跟踪方法

侯代文[1,2],殷福亮[1]
(1.大连理工大学电子与信息工程学院,辽宁大连 116023; 2.海军试验基地,辽宁大连 116041)

摘 要: 本文提出一种基于采样交互的多模型粒子滤波方法,实现了对随意运动说话人的有效跟踪.该方法根据说话人跟踪问题的特点,用马尔可夫跳变系统描述说话人的动态特性,用粒子滤波方法估计说话人的位置.在说话人跟踪过程中,通过调整滤波粒子的采样区域,完成交互式多模型方法中系统状态的交互过程,这不仅实现了各子滤波器中粒子数目的任意设定,避免了模型转换过程中的性能退化现象,而且取消了对模型后验概率密度函数的高斯分布假定,增强了说话人跟踪系统的鲁棒性.计算机仿真实验结果验证了本文方法的有效性.

关键词: 说话人跟踪; 交互式多模型方法; 马尔可夫跳变系统; 粒子滤波; 状态估计
中图分类号: TP272 **文献标识码:** A **文章编号:** 0372-2112 (2010) 04-0835-07

An IMM Particle Filtering Method for Speaker Tracking

HOU Dai-wen[1,2], YIN Fu-liang[1]
(1. *School of Electronic and Information Engineering, Dalian University of Technology, Dalian, Liaoning* 116023, *China*;
2. *Naval Test Base, Dalian, Liaoning* 116041, *China*)

Abstract: A new interacting multiple model(IMM) algorithm based on particle filter is proposed to track a randomly moving speaker. Based on the characteristic of speaker tracking problem, the proposed method represents the dynamic model with Markov jump system and filtering the system state with particle filter. The interacting process is accomplished by properly selecting the sampling region. Thus, not only the number of particles in each mode can be controlled so that the degeneracy problem around mode transition is avoided, but also the Gaussian assumption of posteriori probability density function of the state is cancelled. Simulation results show the validity of the proposed method.

Key words: speaker tracking; interacting multiple model; jump Markov system; particle filter; state estimation

1 引言

说话人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、计算机人机界面以及机器人导航等场合[1].

Bangs等[2]和Brandstein[3]最先开展了说话人定位问题的研究,他们分别使用波束形成方法和时延估计方法,根据当前时刻麦克风阵列接收到的语音信息,确定说话人的位置.这些方法在自由声场条件下,能够实现对说话人的准确定位,但在混响和噪声较强时,往往会由于虚声源的产生而导致对说话人位置的错误估计.针对这一问题,Sturim等[4]提出利用状态空间方法跟踪说话人位置,该方法通过建立跟踪系统的动态方程,滤除观测序列中具有明显误差的观测信息,从而在一定程度上解决了说话人跟踪中的虚声源问题.在基于状态空间方法的说话人跟踪系统中,由于说话人运动往往具有机动性,普通的匀速直线运动模型[5]难以充分描述说话人的运动状态.为此,Vermaak等[6]采用随机游走模型描述说话人的运动,能够基本适应说话人运动状态的机动性,但该方法舍弃了状态方程提供的有用信息,且各运动参数需要根据实际的使用条件训练获取,因此适应性较差.交互式多模型方法(Interacting Multiple Model, IMM)[7~10]利用多个运动模型描述说话人的运动方式,每个运动模型对应于一个滤波器,最终的估计结果是各滤波器输出的加权组合.由于该方法能够适应各种复杂的运动方式,因而成为语音跟踪最具吸引力的方法之一.

在IMM方法中,各模型对应的子滤波器通常采用卡尔曼滤波方法或扩展卡尔曼滤波方法.但在系统模型非线性程度较高或噪声不满足高斯分布时,该方法精度较低,容易产生发散现象[11].近年来发展起来的粒子滤

收稿日期:2008-12-19;修回日期:2009-03-22
基金项目:国家自然科学基金(No.60772161,60372082);高等学校博士学科点专项科研基金(No.200801410015)

波方法,由于能够较好地处理非线性、非高斯问题,已广泛应用于说话人跟踪问题[6,12].McGinnity 等[7]最早将 IMM 方法与粒子滤波方法相结合,提出了基于维数扩张的多模型粒子滤波方法,提高了跟踪性能.但是,在该方法中,各子滤波器采样粒子的数目不能直接控制,而必须由对应模型的概率决定,这在模型转换过程中经常会导致粒子退化现象[13].Boers 等[14]对 IMM 方法进行了改进,实现了对各子滤波器中粒子数目的直接控制,但该方法仅适用于模型的条件后验概率密度函数为高斯分布的情形,因而限制了它的适用范围.基于此,本文针对说话人的运动特点,提出了一种基于采样交互的多模型粒子滤波方法.该方法在说话人跟踪过程中,通过调整粒子的采样区域,完成多模型方法中滤波器输入的交互过程,这不仅能够直接控制各子滤波器中采样粒子的数目,而且摒弃了对各模型后验概率密度函数的高斯假定,使得算法能适应任意的概率分布形式,增强了说话人跟踪系统的鲁棒性.计算机仿真实验结果,验证了本文所提方法对说话人跟踪的有效性.

2 基于时间延迟的说话人定位系统

假定在某一多径环境中,有 M 个任意放置的麦克风,当只存在一个声源信号 $s(k)$时,第 m 个麦克风接收到的信号为

$$y_m(k) = h_m(k) \otimes s(k) + n_m(k) \tag{1}$$

其中 $h_m(k)$是声源到第 m 个麦克风的冲激响应,"$\otimes$"表示卷积运算,$n_m(k)$是加性噪声,一般假定与语音信号 $s(k)$以及其它麦克风的接收噪声 $n_l(k)$ $(l \neq m)$互不相关.语音到达第 m 个麦克风的时间延迟为 $\tau_m = c^{-1} \cdot |r_s - r_m|$,其中 $r_s = (x_k, y_k)$为声源位置,$r_m = (x_m, y_m)$为第 m 个麦克风的位置,c 为声音在空气中的传播速度.

传统的定位方法利用当前时刻获取的语音信息确定声源的当前位置.在这些方法中,每两个麦克风作为一组分别组对,通过自适应特征值分解(AEDA)[15]或者广义互相关方法(GCC)[16]估计时间延迟,本文主要采用 GCC 方法.

设 k 时刻第 p 个麦克风对接收到的信号分别为 $y_{p,1}(k)$和 $y_{p,2}(k)$,则它们之间的互功率谱为 $S_{y_{p,1}y_{p,2}}(f) = E[Y_{p,1}(f)Y^*_{p,2}(f)]$,广义互相关函数可表示为

$$\begin{aligned}\Psi_{y_{p,1}y_{p,2}}(\tau) &= \int_{-\infty}^{+\infty} \varphi(f) S_{y_{p,1}y_{p,2}}(f) e^{j2\pi f\tau} df \\ &= \int_{-\infty}^{+\infty} \boldsymbol{\Psi}_{y_{p,1}y_{p,2}}(f) e^{j2\pi f\tau} df\end{aligned} \tag{2}$$

其中 $\varphi(f)$为加权函数,$\boldsymbol{\Psi}_{y_{p,1}y_{p,2}}(f) = \varphi(f) S_{y_{p,1}y_{p,2}}(f)$为广义互功率谱.这样,两麦克风之间的声波到达时间差 $\tau_p = \tau_{p,1} - \tau_{p,2}$对应于广义相关函数的极值点位置,即

$$\hat{\tau}_p = \arg\max_{\tau} \Psi_{y_{p,1}y_{p,2}}(\tau) \tag{3}$$

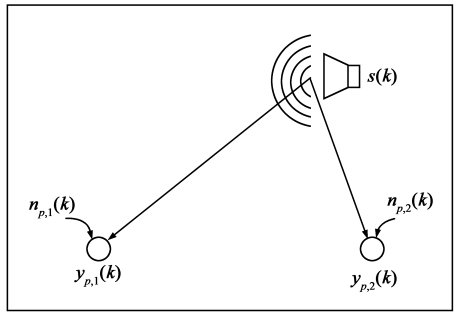

图1 用于时延估计的麦克风阵列摆放示意图

在式(2)中,选取不同的加权函数 $\varphi(f)$,可以得到不同的时延估计方法.在相位变换(Phase Transform,PHAT)方法中[16],加权函数取为 $\varphi(f) = 1/|S_{y_{p,1}y_{p,2}}(f)|$.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,能锐化函数 $\Psi_{y_{p,1}y_{p,2}}(\tau)$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.

对每一个麦克风对及时间延迟估计 $\hat{\tau}_p$ $(p = 1, 2, \cdots)$,其对应的声源轨迹在二维空间上是一双曲线(或三维空间上的双曲面).能够匹配所有麦克风对所对应的双曲线(面)的点,我们就认为是声源所在位置[17].在实际应用中,这种声源定位方法有两个主要缺点:一是在时延估计过程中,当多径现象存在时,实际的时间延迟未必对应于广义相关函数的全局最大值,这将导致出现虚声源;二是存在噪声时,各麦克风对所对应的双曲线(面)不能相交于一点,而只能寻求次优的解决方案.为此,人们考虑采用状态空间方法来确定声源位置.

3 基于采样交互的多模型粒子滤波方法

3.1 说话人跟踪系统描述

状态空间方法通过动态方程和观测方程来描述说话人跟踪系统.由于说话人运动一般具有机动性,其运动模型随时间会发生变化,用单一的运动模型难以描述实际的说话人运动状态,因而需要用一组模型集合才能充分描述说话人在不同时刻的运动特性.一般地,这一类系统可以用马尔可夫跳变系统来描述[13].

设说话人状态为 $x_k = [x_k, v_{x_k}, y_k, v_{y_k}]^{\mathrm{T}}$,其中 x_k, y_k 表示说话人的位置,v_{x_k}, v_{y_k}表示说话人的运动速度.假定说话人不同的运动模式对应于不同的运动状态模型 θ_k,各模型之间的演变服从马尔可夫转移概率矩阵 $\boldsymbol{\Pi} = \{\pi_{ij}\}$,即

$$P\{\theta_k = i \mid \theta_{k-1} = j\} = \pi_{ij} \tag{4}$$

则描述说话人运动的非线性马尔可夫跳变系统可表示为

$$x_k = \boldsymbol{A}(\theta_k)x_{k-1} + \boldsymbol{F}(\theta_k)u_k + \boldsymbol{B}(\theta_k)v_k \quad (5)$$

$$\tau_k = h(x_k, \theta_k) + \boldsymbol{G}(\theta_k)n_k \quad (6)$$

其中,$\boldsymbol{A}(\theta_k)$为状态转移矩阵,$\boldsymbol{F}(\theta_k)$为输入矩阵,$\theta_k \in \{1,2,\cdots,I\}$是系统模型,$I$ 是全体模型种类的个数,u_k 为外部输入激励,$h(x_k,\theta_k)$是系统观测函数,v_k 与 n_k 是相互独立的噪声序列,$\boldsymbol{B}(\theta_k)$和 $\boldsymbol{G}(\theta_k)$为噪声矩阵.

假定 $p(x_0,\theta_0)$已知,设 $\tau_{1:k}=\{\tau_1,\cdots,\tau_k\}$为 k 时刻及以前时刻的时延估计集合,非线性马尔可夫跳变系统(5)、(6)的状态估计问题,就是根据各时刻的观测量 τ_k,递推地估计出后验概率密度函数 $p(x_k|\tau_{1:k})$,从而有效估计说话人状态 x_k.

3.2 粒子滤波方法

对于说话人跟踪问题所构成的非线性系统,必须采用非线性滤波方法处理.扩展卡尔曼滤波方法是使用最广泛的非线性滤波方法,但在系统模型非线性程度较高或噪声不满足高斯分布时,该方法精度较低,容易产生发散现象.由于粒子滤波方法能够较好地处理非线性、非高斯问题,该方法现已广泛应用于说话人跟踪系统.

粒子滤波方法,也称序贯蒙特卡罗方法,它通过蒙特卡罗模拟实现状态的贝叶斯递推估计[18].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,蒙特卡罗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.

对于某一确定的模型 $\theta_k = i \in \{1,2,\cdots,I\}$,令 $\{x^i_{0:k}, i=1,\cdots,N_s\}$表示一支撑点集,对应的权值集合为$\{w^i_{0:k}, i=1,\cdots,N_s\}$,其中 $x_{0:k}=\{x_i, i=0,\cdots,k\}$表示从0到 k 时刻的说话人状态集合,加权值 $w^i_{0:k}$满足归一化条件 $\sum_i w^i_{0:k}=1$.用$\{x^i_{0:k}\}_{i=1}^{N_s}$表示描述后验概率密度函数 $p(x_{0:k}|\tau_{1:k})$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

$$p(x_{0:k}|\theta_k=i,\tau_{1:k}) \approx \sum_{i=1}^{N_s} w^i_k \delta(x_{0:k}-x^i_{0:k}) \quad (7)$$

其中 $\delta(\cdot)$为 Dirac delta 函数.当 $N_s\to\infty$时,式(7)对后验概率密度函数的估计接近于其真实值.

如果支撑点集$\{x^i_{0:k}\}_{i=1}^{N_s}$由重要性概率密度函数 $q(x)$抽样得到,其相应的权值为

$$w^i_k = \frac{p(x^i_{0:k}|\theta_k=i,\tau_{1:k})}{q(x^i_{0:k}|\theta_k=i,\tau_{1:k})} \quad (8)$$

对应的递推估计形式为

$$w^i_k \propto w^i_{k-1}\frac{p(\tau_k|\theta_k=i,x^i_k)p(x^i_k|\theta_k=i,x^i_{k-1})}{q(x^i_k|\theta_k=i,x^i_{k-1},\tau_{1:k})} \quad (9)$$

3.3 交互式多模型方法

对非线性马尔可夫跳变系统,Bar-Shalom 和 Blom 在广义伪贝叶斯方法的基础上,提出了一种具有马尔可夫跳变系数的 IMM 滤波方法[10].该方法利用多个模型并行工作,模型间以概率矩阵 $\boldsymbol{\Pi}$ 进行转移,状态估计值是多个并行处理的滤波器估计结果的综合.IMM 方法在滤波过程中,通过对各种模型假设的删减与合并,实现了滤波性能和计算复杂性之间的较好折衷,成为使用较多的状态估计方法.利用 IMM 方法实现说话人跟踪的具体过程如图 2 所示,主要包括以下四个步骤:

(1)交互过程,实现模型概率和系统状态后验概率密度函数的预测:

$$P\{\theta_{k-1}=j|\tau_{1:k-1}\}\to P\{\theta_k=i|\tau_{1:k-1}\} \quad (10)$$

$$p(x_{k-1}|\theta_{k-1}=j,\tau_{1:k-1})\to p(x_{k-1}|\theta_k=i,\tau_{1:k-1}) \quad (11)$$

(2)滤波过程,利用观测信息,实现各模型下状态条件后验概率密度函数的更新:

$$p(x_{k-1}|\theta_k=i,\tau_{1:k-1})\to p(x_k|\theta_k=i,\tau_{1:k-1}) \quad (12)$$

$$p(x_k|\theta_k=i,\tau_{1:k-1})\to p(x_k|\theta_k=i,\tau_{1:k}) \quad (13)$$

(3)模型概率修正,利用观测信息及模型预测概率,求得模型后验概率:

$$P\{\theta_k=i|\tau_{1:k-1}\}\to P\{\theta_k=i|\tau_{1:k}\} \quad (14)$$

(4)组合过程,根据状态的条件后验概率密度函数,求得状态后验概率密度函数,并计算说话人的状态值:

$$p(x_k|\theta_k=i,\tau_{1:k})\to p(x_k|\tau_{1:k}) \quad (15)$$

IMM 方法不仅保持了滤波过程中的递推结构,而且在每次迭代中计算量基本不变.其优点是能够对具有多种行为模式的动态系统作状态估计,在滤波过程中,通过模型概率的调整,实现滤波结构的自适应变化.另外,IMM 方法通过实时地增减或变更模型,能增强滤波结构的适应能力,尤其适用于说话人跟踪.

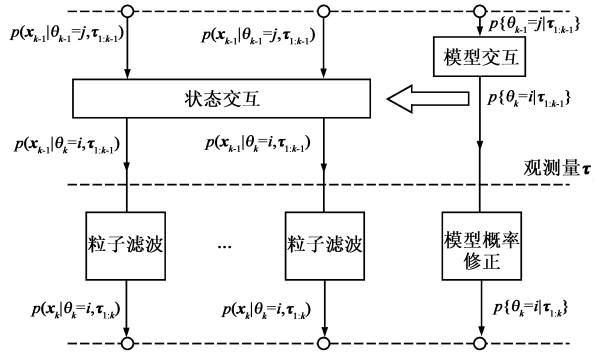

图2　交互式多模型滤波方法

3.4 基于采样交互的多模型粒子滤波方法

在交互式多模型方法中,状态交互(11)是其关键部分,一般通过以下步骤完成:

首先根据全概率公式

$$p(x_{k-1}|\theta_k=i,\tau_{1:k-1})=\sum_{j=1}^{I}p(x_{k-1}|\theta_{k-1}=j,\theta_k=i,\tau_{1:k-1})\cdot P\{\theta_{k-1}=j|\theta_k=i,\tau_{1:k-1}\} \quad (16)$$

由于 θ_{k-1}已知时,x_{k-1}独立于 θ_k,因此

$$p(x_{k-1}|\theta_{k-1}=j,\theta_k=i,\tau_{1:k-1})=p(x_{k-1}|\theta_{k-1}=j,\tau_{1:k-1}) \quad (17)$$

又因为

$$P\{\theta_{k-1}=j|\theta_k=i,\tau_{1:k-1}\}=\sum_{j=1}^{I}\pi_{ij}P\{\theta_{k-1}=j|\tau_{1:k-1}\}/P\{\theta_k=i|\tau_{1:k-1}\} \quad (18)$$

所以

$$p(x_{k-1}|\theta_k=i,\tau_{1:k-1})=\sum_{j=1}^{I}\pi_{ij}p(x_{k-1}|\theta_{k-1}=j,\tau_{1:k-1})\cdot P\{\theta_{k-1}=j|\tau_{1:k-1}\}/P\{\theta_k=i|\tau_{1:k-1}\} \quad (19)$$

基本的 IMM 方法为了进行状态交互,在式(11)中对 $p(x_{k-1}|\theta_{k-1}=j,\tau_{1:k-1})$作了高斯分布假定,这样 $p(x_{k-1}|\theta_k=i,\tau_{1:k-1})$也满足高斯分布,因而可以利用均值和方差的交互,求得交互后的概率密度函数 $p(x_{k-1}|\theta_k=i,\tau_{1:k-1})$.但该方法在滤波过程中不能直接控制各滤波器采样粒子的数目,必须由对应模型的概率决定,这在模型转换过程中经常会导致粒子退化现象.此外,$p(x_{k-1}|\theta_{k-1}=j,\tau_{1:k-1})$在非线性条件下也未必服从高斯分布.为了在非高斯条件下使用基于粒子滤波的 IMM 方法,并避免模型转换过程中的粒子枯竭现象,本文提出基于采样交互的多模型粒子滤波方法.该方法在状态交互过程中,没有直接构造建议分布函数,而是根据概率的相对频率定义,用事件发生的频率表示它可能发生的概率,即根据粒子滤波方法的特点,用某一模型中抽取粒子数目占总粒子数的比例表示该模型出现的概率.在状态交互过程中,从上一时刻各模型的后验概率密度函数 $p(x_{k-1}|\theta_{k-1}=j,\tau_{1:k-1}),(j=1,\cdots,I)$中分别采样,且从模型 j 的后验概率密度函数 $p(x_{k-1}|\theta_{k-1}=j,\tau_{1:k-1})$中采样粒子的数目,与模型转移后验概率 $P\{\theta_{k-1}=j|\theta_k=i,\tau_{1:k-1}\}$的大小成正比,全部采样粒子共同构造模型 i 下的条件概率密度函数 $p(x_{k-1}|\theta_k=i,\tau_{1:k-1})$.在此基础上,进行粒子滤波,并对各滤波器输出加权组合,得到系统状态的统计估计.

基于采样交互的多模型说话人跟踪过程主要包括以下步骤:

(1)计算模型预测概率

在 k 时刻,根据 $k-1$ 时刻的模型后验概率以及马尔可夫模型转移矩阵,预测各模型的概率

$$P\{\theta_k=i|\tau_{1:k-1}\}=\sum_j P\{\theta_k=i|\theta_{k-1}=j\}\cdot P\{\theta_{k-1}=j|\tau_{1:k-1}\} \quad (20)$$

(2)计算模型转移的后验概率

$$P\{\theta_{k-1}=j|\theta_k=i,\tau_{1:k-1}\}=\frac{P\{\theta_k=i|\theta_{k-1}=j\}\cdot P\{\theta_{k-1}=j|\tau_{1:k-1}\}}{P\{\theta_k=i|\tau_{1:k-1}\}} \quad (21)$$

(3)选定采样区域

预先设定或者自适应地选取模型 i 下滤波器所使用的粒子个数 $M_i(i=1,2,\cdots,I)$,根据模型转移后验概率,确定在模型 i 下从第 j 个后验概率密度函数 $p(x_{k-1}|\theta_{k-1}=j,\tau_{1:k-1})$中抽取粒子的个数

$$M_{i,j}=M_i\cdot P\{\theta_{k-1}=j|\theta_k=i,\tau_{1:k-1}\} \quad (22)$$

(4)抽取粒子

由函数 $p(x_{k-1}|\theta_{k-1}=j,\tau_{1:k-1})$中抽取 $M_{i,j}$个粒子 $x_{k-1}^{j,l}$,并对每一个粒子赋以权值 $\tilde{w}_{k-1}^{j,l}=p(x_{k-1}^{j,l}|\theta_{k-1}=j,\tau_{1:k-1})$.

(5)权值归一化

$$w_{k-1}^{j,l}=\tilde{w}_{k-1}^{j,l}\Big/\sum_{j=1}^{I}\sum_{l=1}^{M_{i,j}}\tilde{w}_{k-1}^{j,l} \quad (23)$$

(6)输入交互

在 k 时刻,模型 i 下滤波器的状态输入,是上一时刻全部滤波器状态输出交互的结果.经输入交互后,模型 i 下的条件概率密度函数可以表示为

$$p(x_{k-1}|\theta_k=i,\tau_{1:k-1})=\sum_{j=1}^{I}\sum_{l=1}^{M_{i,j}}w_{k-1}^{j,l}\cdot x_{k-1}^{j,l} \quad (24)$$

(7)粒子滤波

将$\{x_{k-1}^{j,l},w_{k-1}^{j,l}\}_{l=1:M_{i,j}}^{j=1:I}$代入基于第 i 个模型的粒子滤波器,利用 k 时刻的时间延迟观测量,实现各模型匹配下的时间更新和观测更新,从而得到模型 i 下预测状态的概率密度函数 $p(x_k|\theta_k=i,\tau_{1:k-1})$和后验概率密度函数 $p(x_k|\theta_k=i,\tau_{1:k})$.

(8)模型概率修正及归一化

$$\tilde{P}\{\theta_k=i|\tau_{1:k}\}=P\{\theta_k=i|\tau_{1:k-1}\}\cdot p(\tau_k|\theta_k=i) \quad (25)$$

$$P\{\theta_k=i|\tau_{1:k}\}=\tilde{P}\{\theta_k=i|\tau_{1:k}\}\Big/\sum_{j=1}^{I}\tilde{P}\{\theta_k=j|\tau_{1:k}\} \quad (26)$$

(9)加权输出

根据全概率公式,有

$$p(x_k|\tau_{1:k})=\sum_{i=1}^{I}p(x_k|\theta_k=i,\tau_{1:k})\cdot P\{\theta_k=i|\tau_{1:k}\} \quad (27)$$

于是,k 时刻的说话人状态估计值为

$$\hat{x}_k=\int x_k p(x_k|\tau_{1:k})\mathrm{d}x_k \quad (28)$$

在步骤(7)的粒子采样过程中,我们使用了正则粒子滤波方法,并利用 Epanechikov 核函数[19]使采样函数连续化,避免了由于多样性丧失而引起的粒子枯竭现象.在滤波过程中,本方法不同于文献[14]中仅传递状态均值 $E[x_k]$和协方差阵 $P(x_k)$,而是传递整个概率密度函数 $p(x_k|\theta_k=i,\tau_{1:k})$,从而消除了状态必须为高斯分布的限定条件.另外,由于该方法没有在模型空间采样,各子滤波器中采样粒子的数目可以预先确定或者自适应地设定,与模型概率无关,因而能够避免模型转换过程中的粒子枯竭现象.然而,由于每一次迭代过程都需要根据模型转移后验概率计算采样粒子的个数 $M_{i,j}$,本文方法虽然计算量基本保持不变,但滤波结构的复杂性有所增加.

4　计算机仿真与实验结果

为了检验基于采样交互的 IMM 方法对说话人的跟踪效果,我们利用两个麦克风阵列对二维空间上随意运动的说话人进行模拟跟踪,并将跟踪结果与文献[14]中基于维数扩张的多模型粒子滤波方法的跟踪结果进行比较.

4.1　计算机仿真设计

如图 3 所示,在一个 3m × 5m 的房间内,在 X、Y 两个方向上,分别放置两组麦克风,第一组麦克风的位置为(0,2)和(0,3),第二组麦克风的位置为(1.5,0)和(2,0).两组麦克风接收到的说话人语音信息用 IMAGE 模型[20]仿真获得,信噪比为 30dB,利用 PHAT 方法分别计算出每组麦克风之间的时间延迟 τ_1 和 τ_2.利用时间延迟,递推估计说话人位置.

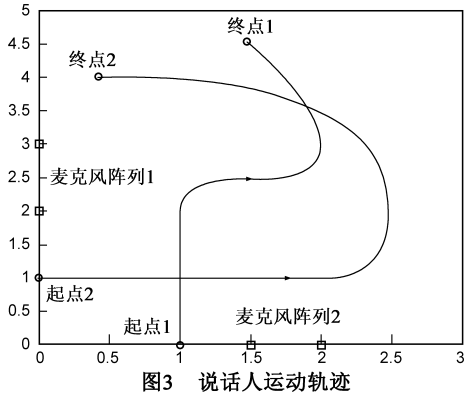

图3　说话人运动轨迹

在跟踪系统中,说话人运动状态随时间的变化服从式(5)和(6)描述的马尔可夫跳变模型,式(5)和(6)中的各参数设定[21]如下:

$$\boldsymbol{A}=\begin{bmatrix}1 & T & 0 & 0\\ 0 & 1 & 0 & 0\\ 0 & 0 & 1 & T\\ 0 & 0 & 0 & 1\end{bmatrix},\boldsymbol{B}=\begin{bmatrix}0.1T\\ 0\\ 0.1T\\ 0\end{bmatrix},\boldsymbol{G}=\begin{bmatrix}(0.02/c)^2\\ (0.02/c)^2\end{bmatrix}$$

这里 $T=0.05$s 为采样周期,$c=343$m/s 为声音在空气中的传播速度.式(6)中的 $\tau_k=(\tau_{k_1},\tau_{k_2})^{\mathrm{T}}$,系统观测方程的具体形式为

$$\tau_{k_i}=c^{-1}\cdot[\sqrt{(x_k-x_{i,1})^2-(y_k-y_{i,1})^2}-\sqrt{(x_k-x_{i,2})^2-(y_k-y_{i,2})^2}],\ i=1,2 \tag{29}$$

其中$(x_{i,j},y_{i,j})$为第 i 个麦克风对中第j 个麦克风的位置.

通常,系统模型的改变是由外部输入控制 $F(\theta_k)u_k$ 的变化引起的,因此不同的输入控制会对应于不同的系统模型 θ_k.本文共设计了两种说话人运动方案,说话人运动轨迹如图 3 所示.

方案一:说话人由点(1,0)开始运动,三种运动模型分别对应于外部输入控制

$F(1)u_k=[0\ 000]^{\mathrm{T}}$,

$F(2)u_k=[0.0025\quad 0.05\quad -0.0025\quad -0.05]^{\mathrm{T}}$,

$F(3)u_k=[-0.0025\quad -0.05\quad 0.0025\quad 0.05]^{\mathrm{T}}$

在仿真实验中,模型的初始概率全部设定为 $P\{\theta_k=i\}=1/3$,它们之间的转移概率为 $\pi_{ii}=0.90$,$\pi_{ij}=0.05(i\neq j)$,即认为说话人保持某一运动模型的可能性远大于作模型转换的概率.初始状态设定为$[0.8\quad 0\quad 0.2\quad 1]^{\mathrm{T}}$,真实值为$[1\quad 0\quad 0\quad 1]^{\mathrm{T}}$,各模型下采样粒子都由先验分布 $p\{x_k|x_{k-1}\}$产生.在本文方法中,每一种模型采样粒子的个数都选定为 $N=500$.在标准的粒子滤波方法中,总采样粒子个数为 $N=2000$.

方案二:说话人由点(0,1)开始运动,对应的三种外部输入控制为

$F(1)u_k=[0\ 000]^{\mathrm{T}}$,

$F(2)u_k=[-0.0025\quad -0.05\quad 0.005\quad 0.1]^{\mathrm{T}}$,

$F(3)u_k=[-0.0025\quad -0.05\quad -0.0025\quad -0.05]^{\mathrm{T}}$

在状态估计中,初始状态设定为$[0\quad 1\quad 0.8\quad 0]^{\mathrm{T}}$,真实值为$[0\quad 1\quad 1\quad 0]^{\mathrm{T}}$.各滤波器中采样粒子的个数与方案一相同.

4.2　实验结果分析

按照上面的仿真设计方案,本文对两种跟踪方案分别进行了 50 次蒙特卡罗仿真实验,实验结果分别如图 4 和图 5 所示.在方案一情况下,文献[14]方法和本文方法的跟踪效果如图 4(a)所示,图 4(b)和图 4(c)分别对应于两种跟踪方法在 X 轴和 Y 轴上的跟踪误差.在方案二情况下,两种多模型方法的跟踪效果如图 5(a)所示,图 5(b)和图 5(c)分别对应于各自在 X 轴和 Y 轴上的跟踪误差.图 6 是滤波器对说话人所处运动模型的实时估计概率.

由图 4(a)和图 5(a)可以看出,使用粒子滤波技术,两种多模型方法都能够实现说话人跟踪,但本文所

提出的方法在滤波精度和收敛速度上均明显优于文献[14]的方法.本文方法在每一种模型上使用 500 个粒子,明显少于文献[14]方法所使用的粒子数,因而能够提高计算速度.此外,本文方法克服了文献[14]方法中由于维数扩张而引起的采样粒子方差增大的问题.由图 5 可以看出,在方案二情况下,文献[14]方法在模型转换过程中的跟踪误差很大,这是因为该方法中各子滤波器采样粒子的数目与模型概率成正比,在模型转换过程中出现了粒子枯竭现象,而本文方法能够控制各子滤波器中采样粒子的数目且不受该模型概率的影响,因而在模型转换过程中仍然能较好地实现说话人跟踪.由图 6 可以看出,本文所提方法对运动所处模型概率的估计与实际的运动轨迹相一致,即前 2 秒,说话人沿直线运动,尔后 1 秒向右转向,然后向左转向,而文献[14]方法对状态模型概率的估计不如本文所提方法准确,这也是它滤波精度较低的原因之一.

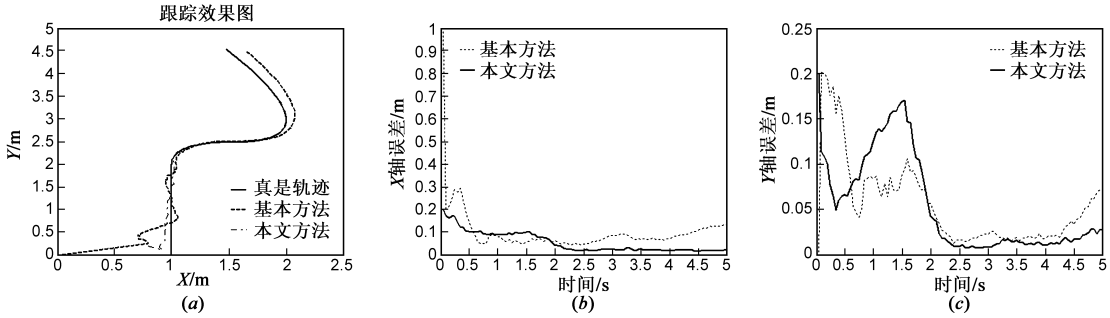

图4 方案一情况下两种滤波方法的说话人跟踪结果比较

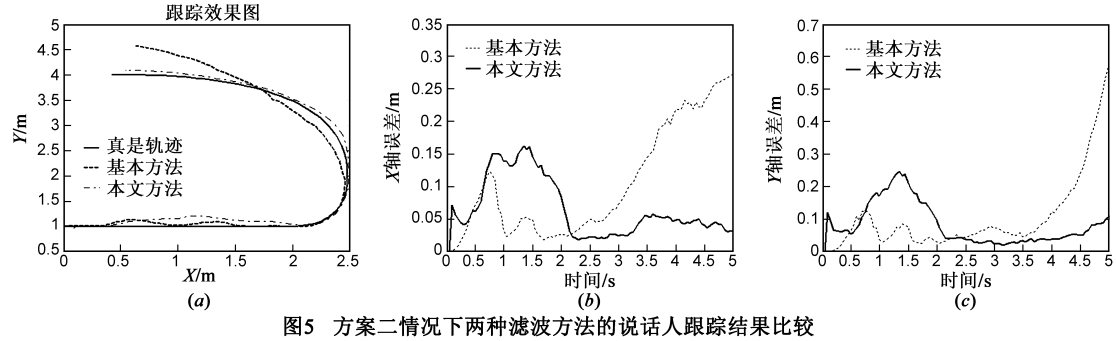

图5 方案二情况下两种滤波方法的说话人跟踪结果比较

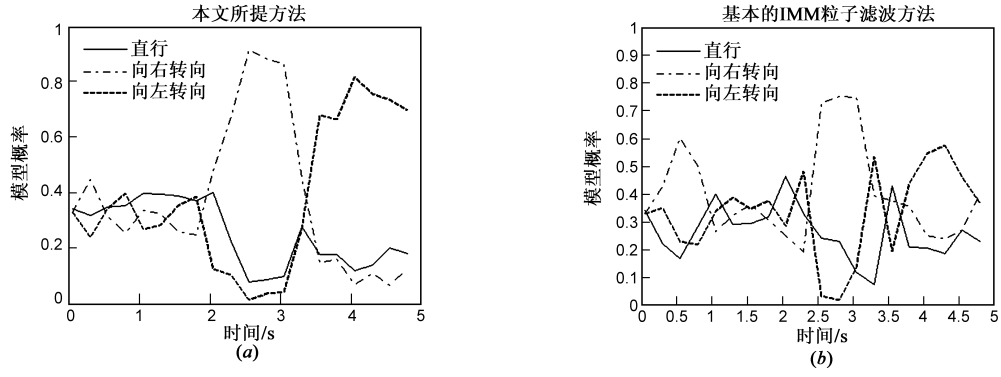

图6 方案一情况下两种方法对说话人运动所处模型概率的估计结果

为了对滤波算法的估计精度作定量比较,定义单次实验的均方误差[18]为

$$RMSE = \frac{1}{M}\sum_{m=1}^{M}\left[\frac{1}{N}\sum_{k=1}^{N}(\hat{x}_{k,m} - x_k)^2\right]^{1/2} \quad (30)$$

其中 N 为总的滤波时间,M 是蒙特卡罗仿真次数,$\hat{x}_{k,m}$ 是第 m 次仿真实验中 k 时刻状态 x 的滤波估计值.

经过计算,两种跟踪方法在 X 轴和 Y 轴上跟踪误差的均方估计误差如表 1 所示.从表 1 可以看出,本文方法总是优于文献[14]方法.

表 1 两种跟踪方法估计误差比较(单位:m)

状态估计误差 / 滤波方法	实验方案一		实验方案二	
	X 轴	Y 轴	X 轴	Y 轴
文献[14]方法	0.1024	0.0752	0.1088	0.1071
本文方法	0.0694	0.0606	0.0705	0.0891

5 结束语

说话人跟踪问题是一种具有多种运动模式的非线

性滤波问题.本文提出的基于采样交互的多模型方法,实现了对说话人位置的准确估计.该方法根据粒子滤波的特点,通过调整采样区域,实现滤波过程中系统状态的交互,不仅实现了对各个模型中所使用粒子数目的任意设定,避免了模型转换过程中的粒子退化现象,而且对各概率密度函数不需再作高斯假定,增强了说话人跟踪系统的鲁棒性.仿真实验结果验证了本文方法的有效性.

参考文献:

[1] I Potamitis, H Chen, G Tremoulis. Tracking of multiple moving speakers with multiple microphone arrays[J]. IEEE Transactions on Speech and Audio Processing, 2004, 12(5): 520 – 529.

[2] W Bangs, P Schultheis. Space-time processing for optimal parameter estimation[A]. Signal Processing[C]. J Grriffiths, P Stocklin, and C Van Schooneveld, Eds., New York: Academic Press, 1973. 577 – 590.

[3] M A Brandstein. Framework for Speech Source Localization Using Sensor Arrays[D]. Ph. D. Thesis, Brown University, Providence, RI, U.S.A, 1995.

[4] D Sturim, M Brandstein, H Silverman. Tracking multiple talkers using microphone array measurements[A]. Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing[C]. Munich, Germany, 1997. 371 – 374.

[5] D Bechler, M Grimm, K Kroschel. Speaker tracking with a microphone array using Kalman filtering[J]. Advances in Radio Science, 2003, 1: 113 – 117

[6] J Vermaak, A Blake. Nonlinear filtering for speaker tracking in noisy and reverberant environments[A]. Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing[C]. Salt Lake City, USA: IEEE, 2001. 3021 – 3024.

[7] S Mcginnity, G Irwin. Multiple model bootstrap filter for maneuvering target tracking[J]. IEEE Transactions on Aerospace and Electronic Systems, 2000, 36(3): 1006 – 1012.

[8] W Farrell. Interacting multiple model filter for tactical ballistic missile tracking[J]. IEEE Transactions on Aerospace and Electronic Systems, 2008, 44(2): 418 – 426.

[9] E Mazor, A Averbuch, Y Bar-Shalom, et al. Interacting multiple model methods in target tracking: a survey[J]. IEEE Transactions on Aerospace and Electronic Systems, 1998, 34(1): 103 – 123.

[10] H A P Blom, Y Bar-Shalom. The interacting multiple model algorithm for systems with Markovian switching coefficients[J]. IEEE Transactions on Automatic Control, 1988, 33(8): 780 – 783.

[11] N Gordon, D Salmond, A F M Smith. Novel approach to nonlinear and non-Gaussian Bayesian state estimation[J]. IEE Proceedings-F, 1993, 140(2): 107 – 113.

[12] D Ward, E Lehmann, R Williamson. 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.

[13] H Driessen, Y Boers. Efficient particle filter for jump Markov nonlinear systems[J], IEE Proceedings on Radar, Sonar and Navigation, 2005, 152(5): 323 – 326.

[14] Y Boers, J Driessen. Interacting multiple model particle filter[J]. IEE Proceedings on Radar, Sonar and Navigation, 2003, 150(5): 344 – 349.

[15] J Benesty. Adaptive eigenvalue decomposition algorithm for passive acoustic source localization[J]. Journal of Acoustical Society of American, 2000, 107(1): 384 – 391.

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

[17] M S Brandstein, J E Adcock, H F Silverman. A closed-form location estimator for use with room environment microphone arrays[J]. IEEE Transactions on Speech and Audio processing, 1997, 5(1): 45 – 50.

[18] M S Arulampalam, S Maskell, et al. A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking[J]. IEEE Transactions on Signal Processing, 2002, 50(2): 174 – 188.

[19] B W Silverman. Density Estimation for Statistics and Data Analysis[M]. London: Chapman & Hall, 1986. 75 – 88.

[20] J B Allen, D A Berkley. Image method for efficiently simulating small-room acoustics[J]. Journal of the Acoustical Society of America. 1979, 65(4): 943 – 950.

[21] X R Li, V P Jilkov. Survey of maneuvering target tracking. part I: dynamic models[J]. IEEE Transactions on Aerospace and Electronic Systems, 2003, 39(4): 1333 – 1364.

作者简介:

侯代文　男,1972 年生于山东嘉祥,博士.主要研究方向为语音信号处理、阵列信号处理、跟踪与定位技术等.

E-mail: hodevin@gmail.com

殷福亮　男,1962 年生于辽宁抚顺,博士生导师,教授.主要研究方向为语音处理、图像处理、阵列信号处理、宽带无线通信技术等.

E-mail: flyin@dlut.cn