文章编号: 1000-3630(2000)02-68-04

音素基随机轨迹模型的理论机制和参数估计方法*

黄心晔，施 嵘，富煜清，陆佶人
(东南大学无线电工程系, 南京 210096)

摘 要: 随机轨迹模型是针对基本隐马尔可夫模型所隐含的三个不合理假设而提出的统计建模方法。本文介绍了随机轨迹模型相对于基本隐马尔可夫模型的优越性, 综述了以音素为语音识别基元的随机轨迹模型的理论机制和参数估计方法。
关键词: 音素; 轨迹元; 参数估计
中图分类号: O423 文献标识码: A

The theory and parameter estimation method to stochastic trajectory models based on phoneme

HUANG Xin-ye, SHI Rong, FU Yu-qing, LU Ji-ren
(Department of Radio Engineering, Southeast University., Nanjing 210096, China)

Abstract: Stochastic trajectory models can effectively overcome the three unreasonable assumptions caused by basic hidden Markov models. This paper introduces the advantage of stochastic trajectory models, and makes a review on the theory and the parameter estimation method of stochastic trajectory models based on phoneme.
Key words: phoneme; component trajectory; parameter estimation

1 引 言

语音识别的最大困难之一是如何对语音的发音速率及声学变化建模。随着隐马尔可夫模型(Hidden Markov Models, 简称HMM)[1]的引入, 这一难题得到了较圆满的解决。它通过状态转移概率对基元的发音速率及声学变化建模, 通过倚赖状态的观察概率密度对基元发音的声学变化建模。但是, 基本隐马尔可夫模型为了保证其计算的有效性和训练的可实现性, 其前向算法隐含了以下三个假设: (1) 状态转移概率与观察序列无关, 且时不变; (2) 状态观察概率密度函数与过去状态无关; (3) 状态观察概率密度函数与过去观察无关。

由于语音是由发音系统连续变化所产生的, 以上基本HMM假设无疑是不合理的。随着发音系统的连续变化, 语音信号可以表示为一个在参数空间(例如倒谱空间)中移动着的点, 称此移动点的序列为语音的轨迹[2]。基本HMM的三个假设, 尤其是假设2, 导致HMM不能保存语音的轨迹信息, 容易产生轨迹折叠现象[2], 使HMM在复杂的上下文环境中鉴别能力不高。为了克服这些现象, 在基本HMM的基础上, 引入了许多扩展, 例如将观察特征矢量扩展为动态特征[3]、以bigram约束对观察特征矢量间相关性建模[4]以及引入依赖上下文的语音基元模型[5]等, 但这些都导致了计算量及训练量的大大增加, 且没有从根本上解决基本HMM假设的不合理性, 因而对其识别性能的改进也必然是有限的。

随机轨迹模型(Stochastic Trajectory

* 收稿日期: 99-08-30; 修订日期: 2000-01-02
国家自然科学基金资助项目(69672010)
作者简介: 黄心晔(1972-), 男, 博士研究生

Models, 简称 STM)[2] 定义轨迹为产生对应于语音基元的语音段的随机过程, 它摈弃了基本 HMM 的三个不合理假设, 从而在理论上更加完善。基于 STM 的连续语音识别系统在实践中也取得了突出的效果[2,6]。本文将详细论述以音素为语音识别基元的随机轨迹模型的理论机制和参数估计方法。

2 音素的概率

令 $\mathbf{X}_n$ 是一个以 n 为中心的 Q 帧特征矢量(参数空间中的点)的序列:

$$\mathbf{X}_n = \mathbf{x}_{n-(Q/2)}, \mathbf{x}_{n-(Q/2)+1}, \cdots, \mathbf{x}_n, \cdots, \mathbf{x}_{n+(Q/2)-1} \quad (1)$$

令每一个音素 s 与一个随机轨迹模型 T_s 相对应。假设 $\mathbf{X}_n$ 是通过对 d 帧矢量序列重采样获得的:

$$[0,d) \xrightarrow{f(i,d,Q)} [0,Q) \quad (2)$$

则: $p(\mathbf{X}_n, d, s) = p(\mathbf{X}_n \ d, s) P(d \ s) P(s)$ (3)

这里, $p(\mathbf{X}_n, d, s)$ 表示出现观察序列 $\mathbf{X}_n$, 持续时间 d 和音素 s 的联合概率密度; $p(\mathbf{X}_n \ d, s)$ 表示给定 d 和 s 观察到 $\mathbf{X}_n$ 的条件概率密度; $P(d \ s)$ 表示给定音素 s 的持续时间 d 的条件概率; $P(s)$ 表示音素 s 的先验概率。以下符号表示的含义类似。

令轨迹模型 T_s 表示为轨迹元 t_k 的混合, 即将其对应的音素表示为参数空间中轨迹聚类的混合, 每一个轨迹元对应一个轨迹聚类, 则:

$$p(\mathbf{X}_n \ d, s) = \sum_{t_k \in T_s} p(\mathbf{X}_n \ t_k, d, s) P(t_k \ d, s) \quad (4)$$

假定对于一个给定的音素, 轨迹元出现的概率并不依赖于它的持续时间, 则有:

$$P(t_k \ d, s) = P(t_k \ s) \quad (5)$$

那么:

$$p(X_n \ d, s) = \sum_{t_k \in T_s} p(X_n \ t_k, d, s) P(t_k \ s) = \sum_{t_k \in T_s} \xi_k^s p(X_n \ t_k, d, s) \quad (6)$$

其中 $\xi_k^s = P(t_k \ s)$ 为对应于某一音素 s 出现第 k 个轨迹元的概率, 有:

$$\sum_{k \in T_s} \xi_k^s = 1 \quad (7)$$

由贝叶斯公式, 在给定观察和持续时间的条件下, 出现音素的概率为:

$$P(s \ \mathbf{X}_n, d) = \frac{p(\mathbf{X}_n, d, s)}{p(\mathbf{X}_n, d)} = \frac{p(\mathbf{X}_n, d, s)}{\sum_{all\ s} p(\mathbf{X}_n, d, s)} = \frac{p(\mathbf{X}_n \ d, s) P(d \ s) P(s)}{\sum_{all\ s} p(\mathbf{X}_n \ d, s) P(d \ s) P(s)} \quad (8)$$

3 轨迹的时间抽取

同一个音素的轨迹 $y_0, y_1, \cdots, y_{d-1}$ 的时间持续长度 d 是不同的。为了满足等持续时间长度的约束, 我们在这里对轨迹进行时间抽取, 即将一段以 y_n 为中心, 长度为 d 的语音映射到一个长度为 Q 的轨迹 X_n。用公式表示为 $[0,d) \xrightarrow{f(i,d,Q)} [0,Q)$

如果是线性映射, 则有:

$$f(i,d,Q) \triangleq \frac{d-1}{Q-1}\left(\frac{Q}{2} - i\right), \quad 0 \le i \le Q \quad (9)$$

$\triangleq$ 符号的含义: 定义为……

很显然, 如图 1 所示, 在时间抽取时会导致信息的丢失, 但是语音是一种时变的短时平稳信号, 在很小的时间段上语音的特性近似不变, 因此这种信息丢失是可以忽略不计的。这种线性时间抽取的方法已经在文献[7]得到了成功的运用。如果使用非线性采样同时使观察概率达到最大, 可以轻微地提高一些识别率, 但是却需要付出巨大运算量的代价。

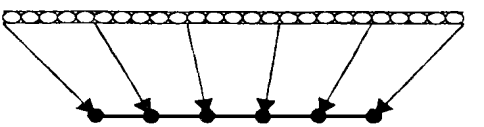

图 1 轨迹的线性时间抽取

4 轨迹概率的计算

在 STM 中计算音素概率的关键就是对 $p(X_n \ t_k, d, s)$ 的计算(参见公式(6))。定义每一个轨迹元 t_k 为一 Q 个状态点的随机状态

序列, 为了训练和计算的有效性, 我们假设, 它的Q 个点之间是统计独立的。那么就有:

$$p(\mathbf{X}_n \ t_k, d, s) = \prod_{i=0}^{Q-1} p(\mathbf{x}_{n-f(i,d,Q)} \ t_k, d, s)$$

$$= \prod_{i=0}^{Q-1} N(\mathbf{x}_{n-f(i,d,Q)}; m_{k,i}^{s}, \Sigma_{k,i}^{s}) \quad (10)$$

其中.^{s} 表示该参数是音素 s 的参数; $\Sigma_{k,i}^{s}$ 表示协方差阵; ω_i^s 是加权系数, 用来控制轨迹上状态的概率分布以对状态方差进行补偿。

在公式(10) 中, 在一个轨迹元上 Q 个采样点中的每个点都可以由一个均值矢量为 $m_{k,i}^{s}$, 协方差矩阵为 $\Sigma_{k,i}^{s}$ 的多元高斯分布来表征:

$N(\mathbf{x}; m_{k,i}^{s}, \Sigma_{k,i}^{s})$, $0 \le k \le card(T_s)$, $0 \le i \le Q$

其中

$$N(\mathbf{x}; m_{k,i}^{s}, \Sigma_{k,i}^{s}) \overset{\Delta}{=} \frac{1}{(2\pi)^{D} \det \Sigma_{k,i}^{s}}$$

$$\exp\left[-\frac{1}{2} (\mathbf{x} - m_{k,i}^{s})^{T} \Sigma_{k,i}^{-1} (\mathbf{x} - m_{k,i}^{s}) \right] \quad (11)$$

$card(T_s)$ 是音素 s 的轨迹元个数。

假设随机矢量 x 是统计独立的, 那么 $\Sigma_{k,i}^{s}$ 成为对角阵。我们就有:

$$N(\mathbf{x}; m_{k,i}^{s}, \Sigma_{k,i}^{s}) \overset{\Delta}{=} \prod_{j=1}^{D} \frac{1}{\sqrt{2\pi\sigma_{k,i,j}^{2}}} \exp\left[-\frac{(xj - mk,i,j)^{2}}{2\sigma_{k,i,j}^{2}} \right] \quad (12)$$

5 参数估计

STM 的参数集为{$P(d \ s)$, $P(s)$, $P(t_k \ s)$, $m_{k,i}^{s} \Sigma_{k,i}^{s}$}。对于每一个音素 s, 随机轨迹模型参数可以通过对训练数据训练得到。

(1) $P(d \ s)$, 给定音素的情况下, 持续时间的概率。我们采用 Γ 函数来对该概率分布建模。文献[8] 证明这是有效的。

$$g\Gamma(d; p, \alpha) = \frac{\alpha^{p} d^{p-1} \exp(-\alpha d)}{\Gamma(p)}$$

$$d \ge 0, p > 0, \alpha > 0 \quad (13)$$

其中

$$\Gamma(p) = \begin{cases} 1 & p = 1 \\ (p-1)\Gamma(p-1) & p > 1 \end{cases}$$

p 和 α 满足 $m_s = \frac{p}{\alpha}$ 和 $\sigma_s^2 = \frac{p}{\alpha^2}$

m_s 和 σ_s^2 是音素 s 的持续时间的均值和方差。由此可见, Γ 函数音素持续时间建模对于每个音素仅仅需要两个额外的参数, 非常方便。

(2) $P(s)$, 音素 s 的概率。

$$P(s) = \frac{\text{训练集中 } s \text{ 出现的个数}}{\text{训练集中所有音素总的个数}} \quad (14)$$

(3) $P(t_k \ s)$, $t_k \in T_s$, 给定音素 s 出现第 k 个轨迹元的概率。

$$P(t_k \ s) = \frac{\text{音素 } s \text{ 的训练集中属于第 } k \text{ 个轨迹元的样本个数}}{\text{音素 } s \text{ 的训练集中总的样本个数}} \quad (15)$$

$$\forall s, \sum_{k \in T_s} P(t_k \ s) = 1 \quad (16)$$

(4) $m_{k,i}^{s}$ 和 $\Sigma_{k,i}$ 是第 k 个轨迹元的第 i 个抽样点的均值矢量和方差矩阵。一个数据样本集(音素训练集), 如果对其以高斯分布建模, 则其分布的均值和协方差的最大似然估计等于该样本集中数据样本的均值和协方差。根据这一思想, 我们就可以在轨迹聚类的基础上, 对其进行高斯分布建模, 从而得出随机轨迹模型中各轨迹元高斯分布均值和方差的最大似然估计。聚类算法采用分段 K 均值算法或 LBG 算法。在聚类过程中, 对音素训练集内所有样本, 我们使用公式(10) 作为距离测度。可以注意到, 该距离测度是基于矢量序列的, 而不是基于单个矢量。在通过聚类算法得到了训练样本集的轨迹聚类后, 以一个聚类对应于一个随机轨迹模型的混合分量, 就可以估计出模型的各个参数了。

6 实验验证[6]

文献[6] 中, 本课题组马小辉博士等进行了基于音素基随机轨迹模型的汉语连续语音识别系统测试。实验结果表明该方法在汉语连续语音识别中可以取得很好的结果。

参 考 文 献

[1] L·R·Rabiner· A tutorial on hidden Markov models and selected applications in speech recognition [J]. Proc. of the IEEE, 1989, 77 (2): 257-285.

(下转第 74 页)

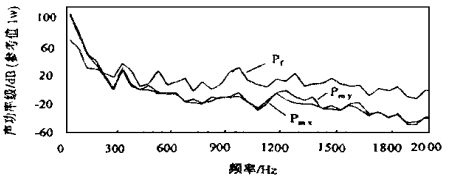

图 7　机器的平动和转动结构声功率流的比较

图 8 所示的是使用大、小水泥地板力导纳数据所预测出的两条平动结构声功率流曲线。其中 P_{f1} 和 P_{f2} 实际上分别代表安装地板为大水泥板和小水泥板时机器的结构声功率流。此结果生动的说明了安装条件对结构声功率流的影响。机器安装在大水泥地板之上，其结构声功率流会大大低于安装在小水泥地板之上的情况，两者可以相差 30dB～40dB 之多。这个结果是可以理解的，但相差如此之大，却是出乎意料的。

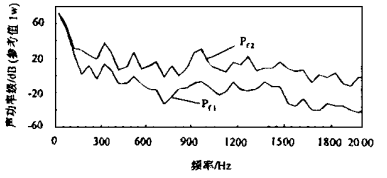

图 8　不同地板对结构声功率流的影响

本文是在机器弹性安装条件下，对结构声功率流进行了预测。但是，工程师也可以根据机器的结构声功率流的限值标准，对特定厂房的地板和所选用的机器反过来估算出所要求的 K_f 和 K_m 值，从而保证机器在安装后在结构声功率流方面获得满意的结果。因此，本文的预测方法具有广泛的应用前景。

本文仅考虑弹性支架的静态劲度系数。有关研究表明，劲度系数随着频率增大而有所增加[3]。因此，考虑这个因素，高频的结构声功率流将比图 7 和图 8 所示的大些。在实际预测中，应考虑采用动态劲度系数数据或对预测结果作适当的修正。

参考文献：

[1] 孙广荣，吴启学. 环境声学[M]. 南京大学出版社，1995：153-157.

[2] Wolde T ten and Gadeflet G R. Development of standard measurement methods for structure-borne sound emission [J]. Noise Control Engineering Journal, 1987, 28(1): 5-14.

[3] Moorhouse A T. Structure-borne emission of installed machinery in building[J]. University of Liverpool, U.K. 1989.

[4] 罗标，邱树业. 机器的自由源速度的测量[J]. 环境工程，1999，17(1)：46-48.

[5] 邱树业. 混凝土地板力导纳的研究[J]. 汕头大学学报，1991，6(1)：32-36.

[6] 邱树业，邱桂明等. 利用力-电类比方法研究力矩激励器[J]. 声学学报，1996，21(6)：941-945.

(上接第 70 页)

[2] Y. Gong. Stochastic trajectory modeling and sentence searching for continuous speech recognition [J]. IEEE Trans. SAP, 1997, 5(1): 33-44.

[3] S. Furui. Speaker-independent isolated work recognition using dynamic features of speech spectrum [J]. IEEE Trans. ASSP, 1986, 34(1): 52-59.

[4] K. K. Paliwal. Use of temporal correlation between successive frames in a hidden Markov model based speech recognizer[J]. In: Proc. ICASSP, 1993, 2: 215-218.

[5] K. F. Lee. Large-vocabulary speaker-independent continuous speech recognition: the SPHINX system [J]. [Ph. D. dissertation]. Pittsburgh: Carnegie Mellon University, 1988.

[6] 马小辉，龚一凡，富煜清等. 基于随机轨迹模型的汉语连续语音识别方法的研究[J]. 声学学报，1997，22(2)：176-181.

[7] M. Ostendorf, S. Roucos. A stochastic segment model for phoneme-based continuous speech recognition. IEEE Trans. ASSP, 1989, 37(12): 1857-1869.

[8] A. Ljolje, S. E. Levinson. Development of an acoustic-phonetic hidden Markov model for continuous speech recognition [J]. IEEE Trans. SP, 1991, 39(1): 29-39.