# 最大互信息用于语音识别

张春涛　　吴善培
（北京邮电大学电信工程学院，北京 100876；第一作者 27 岁，男，博士生）

**摘　要**　将最大互信息理论用于语音识别，最大互信息估计作为目标函数. 在隐马尔可夫模型参数调整过程中运用了泛化概率下降方法，保证了统计意义上实现目标函数的优化. 最大互信息估计用于连接数字语音识别，识别率得到了提高.

**关键词**　语音识别；最大互信息；隐马尔可夫模型
**分类号**　TP18

在利用隐马尔可夫模型(HMM)技术的语音识别系统中，基于文献[1]的重估算法和识别中的 Viterbi 解码就其目标函数而论并不一致，且并未在模型训练中引入模型判别过程中候选模型匹配得分的竞争特征，因此无法保证对于由文献[1]重估公式得到的 HMM 模型参数，识别器具有最小误识率.

## 1　最大互信息用于语音识别

### 1.1　最大互信息原理

以下用到的 HMM 均指连续隐马尔可夫模型(CHMM).

设 $\Lambda$ 为 HMM 参数集，$\Lambda=\{\lambda^{(i)}\}_{i=1}^{M}$，$M$ 为模型数目. $\boldsymbol{X}=(\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_T)$ 为一个训练数据，其中，$\boldsymbol{x}_t=[x_{t_1}, x_{t_2}, \cdots, x_{t_D}]^{\mathrm{T}}$，$1\leqslant t\leqslant T$，$D$ 为 $\boldsymbol{x}_t$ 的维数. $\boldsymbol{X}$ 和 $\lambda^{(i)}$ 的互信息定义为

$$I[\boldsymbol{X};\lambda^{(i)}]=\log\left\{\frac{P(\boldsymbol{X},\lambda^{(i)})}{P(\boldsymbol{X})P(\lambda^{(i)})}\right\}=$$

$$\log P(\boldsymbol{X}|\lambda^{(i)})-\log\left\{\sum_{j=1}^{M}P(\boldsymbol{X}|\lambda^{(j)})P(\lambda^{(j)})\right\}$$

将上式在模型空间$\{\lambda^{(1)}, \lambda^{(2)}, \cdots, \lambda^{(M)}\}$上取统计平均得到平均互信息为

$$f(\lambda^{(1)}, \lambda^{(2)}, \cdots, \lambda^{(M)})=E\{I[\boldsymbol{X};\lambda^{(i)}]\}=\sum_{i=1}^{M}P(\lambda^{(i)})I[\boldsymbol{X};\lambda^{(i)}]$$

最大互信息(MMI)估计就是求出一组模型参数$\{\lambda^{(i)}, i=1, 2, \cdots, M\}$，使得对于给定的训练数据，$f(\cdot)$最大. MMI 估计包括的物理概念是很明显的. 因为使 $f(\cdot)$最大等价于使每一项互信息 $I[\boldsymbol{X};\lambda^{(i)}]$最大.

本文将 MMI 作为目标函数，并结合泛化概率下降算法(GPD)重估 HMM 参数. GPD 算

收稿日期：1997-08-29

法并非确保每次运算目标函数值的增加或减少，而是统计意义上实现目标函数的优化.

### 1.2 HMM 参数的调整

文献[2,3]给出了基于 GPD 算法的 HMM 参数重估方法，本文对其进行了拓展，具体给出了 HMM 参数重估的公式，即

$$\frac{\partial I[\boldsymbol{X};\lambda^{(i)}]}{\partial\lambda^{(i)}}=\frac{1}{P(\boldsymbol{X}|\lambda^{(i)})}\cdot\frac{\partial P(\boldsymbol{X}|\lambda^{(i)})}{\partial\lambda^{(i)}}-\frac{1}{\sum_{j=1}^{M}P(\boldsymbol{X}|\lambda^{(j)})P(\lambda^{(j)})}\left[\sum_{j=1}^{M}\frac{\partial P(\boldsymbol{X}|\lambda^{(j)})}{\partial\lambda^{(i)}}P(\lambda^{(j)})\right]$$

其中

$$P(\lambda^{(j)})=1/M,\ 1\leqslant j\leqslant M$$

$$P(\boldsymbol{X}|\lambda^{(i)})=\sum_{t=2}^{T}[\log a^{(i)}_{\bar{q}_{t-1},\bar{q}_t}+\log b^{(i)}_{\bar{q}_t}(\boldsymbol{x}_t)]+\log\pi^{(i)}_{\bar{q}_1}$$

其中，$\bar{q}=(\bar{q}_1,\bar{q}_2,\cdots,\bar{q}_T)$为最优状态序列；$a$ 为状态转移概率；$\pi$ 为初始状态概率；$b$ 为观察概率密度函数，$b^{(i)}_j(\boldsymbol{x}_t)=\sum_{k=1}^{K}c^{(i)}_{jk}N[\boldsymbol{x}_t,\boldsymbol{\mu}^{(i)}_{jk},\boldsymbol{R}^{(i)}_{jk}]$，这里 $K$ 表示混合度数目；$N[\cdot]$表示正态分布；$c^{(i)}_{jk}$ 为混合度系数；$\boldsymbol{\mu}^{(i)}_{jk}=\{\mu_{jkl}\}_{l=1}^{D}$为均值失量；$\boldsymbol{R}^{(i)}_{jk}$ 为协方差矩阵，实验中取对角矩阵，$\boldsymbol{R}^{(i)}_{jk}=\{\sigma^2_{jkl}\}_{l=1}^{D}$.

实验中为易于参数重估及参数值概率上的限制，首先进行参数转换. $\tilde{c}_{jk}=\log c_{jk}$，$\tilde{\mu}_{jkl}=\mu_{jkl}/\sigma_{jkl}$，$\tilde{\sigma}_{jkl}=\log\sigma_{jkl}$. 均值参数调整为

$$\tilde{\mu}^{(i)}_{jkl}(n+1)=\mu^{(i)}_{jkl}(n)-\varepsilon\left.\frac{\partial I(\boldsymbol{X};\Lambda)}{\partial\tilde{\mu}^{(i)}_{jkl}}\right|_{\Lambda=\Lambda_n}$$

其中

$$\frac{\partial I(\boldsymbol{X};\Lambda)}{\partial\tilde{\mu}^{(i)}_{jkl}}=-\sum_{t=1}^{T}\delta(\bar{q}_t-j)\frac{\partial\log b^{(i)}_j(\boldsymbol{x}_t)}{\partial\tilde{\mu}^{(i)}_{jkl}}$$

$$\frac{\partial}{\partial\tilde{\mu}^{(i)}_{jkl}}\log b^{(i)}_j(\boldsymbol{x}_t)=c^{(i)}_{jk}(2\pi)^{-D/2}|\boldsymbol{R}^{(i)}_{jk}|^{-1/2}(b^{(i)}_j(\boldsymbol{x}_t))^{-1}\left(\frac{x_{tl}}{\sigma^{(i)}_{jkl}}-\tilde{\mu}^{(i)}_{jkl}\right)\times$$

$$\exp\left\{-\frac{1}{2}\sum_{l=1}^{D}\left(\frac{x_{tl}}{\sigma^{(i)}_{jkl}}-\tilde{\mu}^{(i)}_{jkl}\right)^2\right\}$$

$$\mu^{(i)}_{jkl}(n+1)=\sigma^{(i)}_{jkl}(n)\tilde{\mu}^{(i)}_{jkl}(n+1)$$

其它参数重估依此类推.

在参数重估过程中为保证$\sum_{j=1}^{L}a_{ij}=1$ 和$\sum_{k=1}^{K}c_{jk}=1$，还应进行变换：$a_{ij}=a'_{ij}/\sum_{j=1}^{L}a'_{ij}$，$c_{jk}=c'_{jk}/\sum_{k=1}^{K}c'_{jk}$，$L$ 表示 HMM 模型的状态数目，$a'_{ij}$和 $c'_{jk}$为重估值.

实验中，只用发生误识的语音数据进行参数的调整，这样迭代次数不超过 10 次便可以收敛.

## 2 实　验

### 2.1 数据库

我们采用包括 600 人语音的数据库. 在这个数据库中，包括日常对话及连接数字发音. 选取连接数字发音作为测试，600 人的语音中，每人连续地说出 3～4 遍字长从 4～7 的连接数

字，其中 300 人的发音作为训练集，另外 300 人的发音作为测试集. 训练集中包括 1 028 个连接数字串，共 4 790 个数字，测试集中包括 1 051 个连接数字串，共 4 962 个数字.

语音数据库通过电话线采集，采样率为 8 kHz. 首先应将数据转换成线性码，再进行特征提取.

### 2.2 实验条件

预加重 $1-0.95z^{-1}$，20 ms 汉明窗，特征参数为 27 维，包括 12 维线预测倒谱参数、12 维差分倒谱参数、规一化能量参数、一阶差分能量参数和二阶差分参数[4]. 采用 CHMM，每字一个模型，每个模型有 8 个状态，混合度数目为 8.

### 2.3 实验结果

连接数字语音识别主要有两种方法，即 One Stage 算法[5]和 Level Building 算法[6,7]. 由于 One Stage 算法在识别中不容易结合字长信息，且识别率比 Level Building 算法低，所以采用 Level Building 算法作为 MMI 估计的识别算法. 表 1～表 3 分别为采用 One Stage，Level Building 及 MMI 估计的识别结果.

**表 1 One Stage 算法识别结果**

| | 字识别率/% | 串识别率/% | 替换数/个 | 插入数/个 | 删除数/个 |
|---|---|---|---|---|---|
| 训练集 | 93.401 | 84.241 | 56 | 43 | 63 |
| 测试集 | 90.333 | 80.685 | 81 | 53 | 69 |

**表 2 Level Building 算法识别结果**

| | 已知字长 | | 未知字长 | | | | |
|---|---|---|---|---|---|---|---|
| | 字识别率/% | 串识别率/% | 字识别率/% | 串识别率/% | 替换数/个 | 插入数/个 | 删除数/个 |
| 训练集 | 98.601 | 93.969 | 98.843 | 85.019 | 54 | 33 | 67 |
| 测试集 | 95.909 | 87.821 | 91.547 | 81.922 | 86 | 31 | 73 |

**表 3 MMI 识别结果**

| | 已知字长 | | 未知字长 | | | | |
|---|---|---|---|---|---|---|---|
| | 字识别率/% | 串识别率/% | 字识别率/% | 串识别率/% | 替换数/个 | 插入数/个 | 删除数/个 |
| 训练集 | 98.893 | 94.261 | 94.326 | 85.409 | 49 | 28 | 68 |
| 测试集 | 96.413 | 88.297 | 91.905 | 82.303 | 79 | 28 | 70 |

## 3 结 论

MMI 理论用于连接数字语音识别，识别率得到了提高. 相信如果增加训练数据和增加 CHMM 模型的混合度数目，并且有机地将最大似然估计和 MMI 估计结合起来训练模型，识别率还将进一步提高.

## 参 考 文 献


1 Rabiner L R, Juang B H. Fundamentals of speech recognition. New Jersey: Prentice Hall, 1993. 321～386

2 Chou W. Segmental GPD training of HMM based speech recognizer. In: Proc ICASSP. San Francisco: 1992. 473～476

3 Liu C S, Lee C H, Chou W, et al. A study on minimum error discriminative training for speaker recognition. J Acoust Soc Am, 1995, 97(1): 637～648

4 Wilpon J G, Lee C H. Rabiner L R. Improvements in connected digit recognition using higher order spectral and energy features. In: Proc ICASSP. Toronto: 1991. 349～352

5 Ney H. The use of a one-stage dynamic programming algorithm for connected word recognition. IEEE Trans Acoustics Speech and Signal Processing, 1984, 32(2): 263～271

6 Myers C S, Rabiner L R. A level building dynamic time warping algorithm for connected word recognition. IEEE Trans Acoustics Speech and Signal Processing, 1981, 29(2): 284～297

7 Rabiner L R, Wilpon J G, Soong F K. High performance connected digit recognition using hidden Markov models. IEEE Trans Acoustics Speech and Signal Processing, 1989, 37(8): 1 214～1 225


# Applying Maximum Mutual Information to Speech Recognition


Zhang Chuntao　Wu Shanpei

(School of Telecommunication Engineering, Beijing University of Posts and Telecommunications, Beijing 100876)



**Abstract** The theory of maximum mutual information is applied into speech recognition, maximum mutual information estimation is used as object function. General probability descent method is used to ensure the optimization of object function in statistics during the process of HMM parameter modification. The recognition rate is increased when maximum mutual information estimation is applied into connected digit speech recognition.



**Key words** speech recognition; maximum mutual information; HMM