第25卷第3期　　湘潭大学自然科学学报　　Vol.25 No.3
2003年9月　　Natural Science Journal of Xiangtan University　　Sept 2003

# 训练方法对汉语数字串识别率影响的研究*

汤　森[1]，　蔡莲红[2]

（1.江门教育学院计算机系，广东 江门 529000；2.清华大学计算机系，北京 100084）

[摘要]　在语音识别系统中，都是通过提取特征向量来计算待识语音与模型之间的概率或距离，然后根据最大概率或最小距离判断待识语音的类别.对大量实验数据的观察发现：特征向量的各维对语音的表达能力是不一样的，同时特征向量在语音的时间轴上表达能力也不一样.根据这种特性，提出了三种训练算法：在训练中计算出加权矩阵，以此来加强易混淆数字间的本质区分特征，减弱随机特征.在汉语数字串识别实验中，得到了比较理想的实验结果，错误下降40.8%，系统识别率上升到94.08%.

关　键　词：语音识别；汉语数字串；训练算法；加权矩阵

中图分类号：TP391.42　　文献标识码：A　　文章编号：1000-5900(2003)03-0016-05

# Research work on the influence of training method in mandarin connected digit speech recognition

*TANG Lin*[1]，　*CAI Lian-hong*[2]

(1. Department of Computer Science and Technology, Jiang Men Education College, Guang Dong, Jiangmen 529000 China;
2. Department of Computer Science and Technology, Tsinghua University, Beijing 100084 China)

**[Abstract]**　In speech recognition system, the probability or distance between the speech to be recognized and the model is calculated by extracted feature vectors, the speech is recognized as the same kind that who has the maximum probability or minimum distance. By lots of experiments, it is observed that feature vectors is different in describing different phoneme in different characters and in different frames. By this, this paper provide three training methods using weight matrix to emphasize the inherent difference and decline the random difference. In mandarin connected speech recognition system, the recognition rate has been increased to 94.08% by using those methods, and the recognition error rate has been decreased by 40.8%.

**Key words:**　speech recognition; mandarin connected digit; training method; weight matrix

汉语口呼数字串的识别一直是语音识别的重要研究课题，虽然只有10个数字，但连呼数字识别所遇到的问题几乎不亚于连续语音识别，特别是数字串的识别没有语法知识可利用，从而缺少语言模型对识别系统的支持.但这也为研究语音参数的表达能力、鲁棒性，识别方法的有效性等提供一个良好的机会.同时，连续数字串应用极为广泛，如电话拨号、密码验证、身份验证等.因此，连呼数字串的识别被众多的研究者所关注.

在已实现的连续数字识别系统中[1~3]，系统的识别率都比较高，最高的识别率达到98.5%.在这些系统中，数字的误识主要集中在2与8、0与6等上面，该类误识占整个误识的91%.这类问题严重影响识别系统性能的提高.解决这个问题有两个途径，一是使用新的参数，如在参数中加入基频[4]、加入低频能量比、频谱质心等[5].另一种，是研究参数对识别的贡献而改进训练过程.

不少系统的训练方法都是以类内最大概率或类内最小距离作为训练准则，而识别系统的目标却是正识率，也就是各类样本的区分度.这种系统只利用了训练样本分类后各类类内信息而没有利用所有样本的信息.怎样在训练过程中充分应用所有训练样本的先验知识来提高识别系统的区分能力从而提高识别率是一个值得研究的课题.在传统的利用HMM技术的语音识别系统中，基于Baum—Welch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，因此无法保证对于由Forward—backward重估公式得到的HMM模型参数识别器具有最高识别率.有人从信息论的角度构造目标函数，训练的目标是

* 收稿日期：2003-06-20
作者简介：汤　森(1962-)，男，湖南人，硕士，讲师.

使模型间的区分能力最大，提出了“最大互信息(MMI)准则”[6]和“最小区别信息(MDI)准则”[7]．也有人提出了“矫正训练(CT)算法”[8]及改进[9,10]．这些系统在训练过程中都充分利用了所有训练样本的信息，而不是只用同类样本来训练该类模型，这样提高了模型的区分能力．

通常，识别准则里各特征对识别结果的影响是等同的，而在实验观察中发现：语音特征在时间轴上和特征轴上对语音的表达能力是不一样的．在此观察结果的基础上，本文提出了通过增加加权矩阵来提高系统对易混淆语音对的识别性能的训练方法．下面，先讨论识别准则和特征的表达能力，再分别叙述各训练算法，并给出识别结果．

## 1　识别准则和特征的区分性能分析

### 1.1　识别准则

在 DTW 算法中[1]，识别的准则为：

$$\underset{n}{\operatorname{argmin}}\{\sum_{i}\sum_{j=1}^{M}[x(i,j)-c(k,j,n)]^2\} \tag{1}$$

$x(i,j)$是待识样本第 $i$ 帧的第 $j$ 维特征，$c(k,j,n)$ 是第 $n$ 个模板在 DTW 路径中与待识样本第 $i$ 帧相匹配的第 $k$ 帧的第 $j$ 维特征，$M$ 为特征的维数．从公式中可以看到，每维特征对欧氏距离的贡献是一样的．在实验中，发现每维特征对不同音素的表达能力是不一样的，特别是对最小对来说．特征表达能力的不同表现在特征维上和时间轴上．本文提出了三种利用这种能力的方法：第一种就是在时间轴上利用特征时间轴上不同的表达能力来调整模板参数，第二种是在特征轴上对每维特征加上不同的权重来提高系统性能，另一种是用加权矩阵同时在时间轴和特征轴上对每维特征和每帧加上不同的权重来强调特征对不同音素的区分性能．也就是将(1)式改为：

$$\underset{n}{\operatorname{argmin}}\{\sum_{i}\sum_{j=1}^{M}[(x(i,j)-c(k,j,n))^2\times p(k,j,n)]\} \tag{2}$$

加权矩阵是通过训练过程得到的．如果对所有样本都训练一个加权矩阵将增大几倍的存储量同时增加大量的计算时间．在本文实现的系统中，采用两级识别算法，第一级用 Level Building 识别出初始数字，第二级识别才采用以上方法对各数字精确识别．第二级识别的关键在于由训练得到易混淆的数字之间的加权矩阵．

### 1.2　特征的区分性能

目前，大多数语音识别系统都采用 MFCC 参数加上短时能量及差分能量作语音特征参数，在对这些特征的观察中发现：对于易混淆的语音对，由于其他非区分特征的随机噪声经常淹没该语音对的本质区分特征，造成误识．如口呼数字识别中，从直观上考虑，2 与 8 可用差分能量来区分，而事实上，不同情况下的语音 2 之间的特征差别已大于 2 和 8 之间的差别，造成 2 与 8 的混淆．如图(1)所示．该图是对 4 位男声的 5 遍孤立数字发音的 DTW 平均而作．其中细实线为数字 2 内部的平均距离，细虚线为数字 8 内部的平均距离，粗实线为数字 2 与 8 之间的平均距离．从图中可以发现第 8—12 帧的差分能量还是有一定的区分能力．

在加权矩阵中，将 1－7 帧参数的权值变小、8－12 帧参数的权值增大，这将增加该特征对 2－8 的区分能力．由此可见，利用所有训练数据来生成各易混淆的数字对之间的加权矩阵来增加特征的区分能力是可行的．

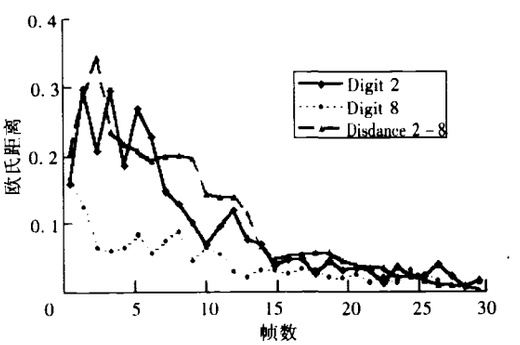

图 1　特征区分图

## 2　识别引导下的训练

### 2.1　Bayes 训练

借用 Bayes 分类器的方法，先将分好类的样本用聚类法求出每类样本的平均模板(模板数可以事先确

定,也可以根据类内距离小于某域值自动确定模板数),然后,对所有训练样本进行识别实验,当样本与本类模板最小距离同该样本与其他类模板最小距离的差大于某域值时,认为分类正确,反之,则调整该两个模板使该样本与这两个模板的距离差朝区分加大的方向移动.对所有样本实验多遍,直至识别率不再提高.具体算法如下:

设训练集为 $T=\{A^1,A^2,\cdots,A^N\}$,$N$ 为训练样本数,$A^i=(a_1^i,a_2^i,\cdots,a_{n_i}^i\}$, $1\leqslant i\leqslant N$,$n_i$ 为第 $i$ 个样本的帧数,其中 $a_j^i=(a_1^i,a_2^i,\cdots,a_p^i$,$p$ 为特征参数的维数.$\arg C(A^i)$ 为 $A^i$ 的类别号.

ⅰ 用聚类法形成初始模板,$M^h=\{M_1^h,M_2^h,\cdots,M_{w_h}^h\}$,$h$ 为类别号,$w_h$ 为 $h$ 类的模板数.

ⅱ 设定循环次数和参数 $\delta$、$\eta$($\delta$ 为模板是否可调的距离比值,$\eta$ 是模板特征参数每次调整的比例),根据经验取值,一般循环 5 次,$\delta$ 取 0.3 和 $\eta$ 取 0.1.

ⅲ 用所有样本进行识别实验:

a) $i=1$;

b) 计算 $\min\limits_{\forall k\ \arg C(M^k)\neq\arg C(A^i)}(D(A^i,M^k))$和 $\min\limits_{\forall j\ \arg C(M^j)\neq\arg C(A^i)}(D(A^i,M^j))$

c) 当 $\min\limits_{\forall k\ \arg C(M^k)\neq\arg C(A^i)}(D(A^i,M^k))-\min\limits_{\forall j\ \arg C(M^j)=\arg C(A^i)}(D(A^i,M^j))>\delta\times\min\limits_{\forall j\ \arg C(M^j)=\arg C(A^i)}(D(A^i,M^j))$,不做任何修改,跳到 $d$)步;

当 $\min\limits_{\forall k\ \arg C(M^k)\neq\arg C(A^i)}(D(A^i,M^k))-\min\limits_{\forall j\ \arg C(M^j)=\arg C(A^i)}(D(A^i,M^j))>\delta\times\min\limits_{\forall j\ \arg C(M^j)=\arg C(A^i)}(D(A^i,M^j))$.调整模板 $M^k$ 和 $M^j$,这两模板分别为 $A^i$ 与本类模板和非本类模板中距离最小者.调整公式为:

按 DTW 路径,对于 $M^k$ 的第 $v$ 帧对应 $A^i$ 的第 $u$ 帧作如下调整:

$$m_v^k(s)=m_v^k(s)+\eta\times(m_v^k(s)-a_u^i(s))\qquad\forall 1\leqslant s\leqslant p$$

从 $M^k$ 的第 1 帧直到最后一帧.

同样,按 DTW 路径,对于 $M^j$ 的第 $v$ 帧对应 $A^i$ 的第 $u$ 帧作如下调整:

$$m_v^j(s)=m_v^j(s)+\eta\times(m_v^j(s)-a_u^i(s))\qquad\forall 1\leqslant s\leqslant p$$

从 $M^j$ 的第 1 帧直到最后一帧.

跳到 d)步;

d) $i=i+1$;

e) 如果 $i$ 小于 $N$,跳到 $a$)继续执行.否则执行ⅳ.

ⅳ 小于控制循环次数同时误识别次数在减少,跳到ⅲ继续执行,否则结束训练.

### 2.2 参数加权算法

MFCC 参数对语音的表达能力比较强,但对一些易混淆语音对还是较难区分.根据特征向量的各维对不同语音有不同的表征能力的特点,在训练过程中,对易混淆语音对求出针对该语音对的特征向量各维的加权系数.在初识别后,针对易混淆语音对加上加权系数再识别一次,这样能有效提高对易混淆语音的区分能力.

为求易混淆语音对的加权系数,先构造一目标函数:

$a$. Right__Number 为训练集中易混淆语音对之间正确距离与错误距离比小于 0.8 时的样本数.

$b$. Right__Rate 为训练集中易混淆语音对之间所有正确距离之和与所有错误距离之和的比值.

调整各维特征参数的系数,使 Right__Number 最大,然后使 Right__Rate 最大.由于特征维数相当大,如果用穷举法计算计算量非常大,本文用分组穷举来解决这问题,只得到次优解.

### 2.3 加权矩阵算法

参数加权算法只是针对特征维上的优化,没作时间轴上的优化.而时间轴上的优化对区别易混淆语音对更有效,因此,给所有易混淆语音对中的每个模板增加一个与它易混淆的语音相对应的加权矩阵 $\sigma(M,L)$,以此增加参数在特征维上和时间轴上的区分能力.其中 M 为语音特征的维数,L 是该语音模板的帧长.(如:0 与 6 和 0 与 9 易混淆,因 DTW 路径不对称,所以要生成四个加权矩阵,即:0 与 6 的矩

阵、0 与 9 的矩阵、6 与 0 的矩阵、9 与 0 的矩阵).矩阵是在训练过程中求得.矩阵元素的构成准则为:使本类样本内的距离下降,而类间距离增加.下面详细讨论该算法.

先定出识别系统的易混淆语音对,然后,用所有训练样本计算易混淆语音对的加权矩阵.

设待识样本为 a,帧长为 N.识别结果中待识样本与非本类的最小距离模板为 b,待识样本与正确类的最小距离模板为 c,设 b 对 a 的加权矩阵为 $\sigma_{bn}(M,L)$,c 对 a 的加权矩阵为 $\sigma_{cn}(M,L)$,K 为 b 的帧数,L 为 c 的帧数.X 为易混淆语音对总数.

ⅰ　将分好类的训练样本用聚类法形成各类初始模板.将易混淆语音对中所列元素的加权矩阵的所有元素置 1,设置参数 θ(根据多次实验确定,一般取 0.3),x = 1.

ⅱ　设置最大循环次数 LOOP.

ⅲ　用第 x 对易混淆语音对所对应的所有同类训练样本进行识别实验.

当:$D_{DTW}(a,b) - D_{DTW}(a,c) > \theta \times D_{DTW}(a,c)$ 时,跳到ⅳ.

当:　$|D_{DTW}(a,b) - D_{DTW}(a,c)| <= \theta \times D_{DTW}(a,c)$ 时,

也就是:

$$\left|\sum_{i=1}^{N}\sum_{j=1}^{M}[d_{DTW}(a(i,j),b(k,j))]\sigma_{bn}(k,j) - \sum_{i=1}^{N}\sum_{j=1}^{M}[d_{DTW}(a(i,j),c(u,j))]\sigma_{cn}(u,j)\right| <= \theta \times \sum_{i=1}^{N}\sum_{j=1}^{M}[d_{DTW}(a(i,j),c(u,j))]\sigma_{bn}(u,j)$$

计算:

$$D_e = \sum[d_{DTW}(a(i,j),c(u,j))\sigma_{cn}(u,j) - d_{DTW}(a(i,j),b(k,j))\sigma_{bn}(u,j)]$$

$$\forall\, d_{DTW}(a(i,j),c(u,j))\sigma_{bn}(u,j) \geqslant d_{DTW}(a(i,j),b(k,j))\sigma_{cn}(u,j)$$

$$D_r = \sum[d_{DTW}(a(i,j),b(k,j))\sigma_{bn}(k,j) - d_{DTW}(a(i,j),c(u,j))\sigma_{cn}(u,j)]$$

$$\forall\, d_{DTW}(a(i,j),c(u,j))\sigma_{bn}(u,j) \leqslant d_{DTW}(a(i,j),b(k,j))\sigma_{cn}(u,j)$$

$$E = D_e - D_r$$

当 $D_r \geqslant \eta \times D_e$ 时(η 由经验确定,一般取 0.5),认为参数可调.否则,跳到ⅳ.

设 α 为加权矩阵的调整因子,有:

$$D_r \times (1-\alpha) \leqslant D_r \times (1+\alpha) \qquad \text{即:}\alpha \geqslant \frac{|D_e - D_r|}{D_e + D_r} \qquad \text{一般取:}\alpha = \frac{|D_e - D_r|}{D_e + D_r} + 0.05$$

调整加权矩阵:

$$\left.\begin{aligned}\sigma_{bn}(k,j) &= \sigma_{bn}(k,j) \times (1+\alpha)\\ \sigma_{cn}(u,j) &= \sigma_{cn}(u,j) \times (1+\alpha)\end{aligned}\right\} \forall\, d_{DTW}(a(i,j),c(u,j))\sigma_{cn}(u,j) \leqslant d_{DTW}(a(i,j),b(k,j))\sigma_{bn}(k,j)$$

$$\left.\begin{aligned}\sigma_{bn}(k,j) &= \sigma_{bn}(k,j) \times (1-\alpha)\\ \sigma_{cn}(u,j) &= \sigma_{cn}(u,j) \times (1-\alpha)\end{aligned}\right\} \forall\, d_{DTW}(a(i,j),c(u,j))\sigma_{cn}(u,j) \geqslant d_{DTW}(a(i,j),b(k,j))\sigma_{bn}(k,j)$$

ⅳ　当所有样本计算完毕跳到ⅴ,否则继续执行ⅲ.

ⅴ　当循环数小于 LOOP,且调整后正确识别的样本数在增加或持平继续执行ⅲ.否则执行下一步.

ⅵ　当 $x \leqslant X$ 时跳到ⅱ,否则训练结束.

## 3　实验结果与结论

### 3.1　实验条件

在普通微机实验室中,通过录音平台,录制了 6 男、4 女各 80 个数字串(孤立数字 10 个、2 位数字串 30 个、3 位数字串到 6 位数字串各 10 个,共含 250 个数字),以 wav 格式保存,平均串长 3.125 个数字,每人读 5 遍.人员来自全国各地,都以普通话朗读,但带有一定的地方口音,年龄大多在 20 ~ 24 岁,最大的 49 岁.数字串中,数字和数字连音出现的次数基本相同.在录音过程中,可能包含自然的开门关门声音、喘气声、咂嘴声等,这些可通过有效的端点检测算法滤除掉.对发音速度不过分限制,最快的每秒钟平均

4.7 个数字，最慢的每秒钟只有 2.92 个数字.

### 3.2 实验方法

用 4 人(2 男 2 女)前三次的单数字(数字串长为 1)和双数字(串长为 2)作为训练集(每个数字有 84 个样板，共 840 个)，所有语音为识别集.人工判定训练集中各数字的语音的起止点，根据训练语音的音长将每个数字的训练集分别分成 4 个子集，然后对每个子集用聚类算法中的最大距离分裂与 K－mean 算法计算通过 DTW 规整的平均语音模板，因此，各数字的模板数由训练集中该数字的距离分布状况决定，分布集中的只有 4 个语音模板，分布分散的有 12 个语音模板，各语音模板的长度也因训练集中各数字语音长度的不同而不同.

本文所述系统的语音参数采用 12 阶 MFCC 系数、及其一阶差分和二阶差分再加上短时能量与差分能量共 38 维特征.

识别系统采用二级识别法，即采用传统的 DTW 识别法后，再对易混淆语音进行二次识别，由于 2－8、0－6 之间的误识占整体误识的 90% 以上，因此，本系统只对 2－8、0－6 之间采用了前面的基于识别引导下的训练及相应的二级识别算法.

实验平台为 VC＋＋6.0.

### 3.3 实验结果

其它错误率包含插入错误(一个数字误识成多个数字)、删除错误(多个数字误识成一个数字)和除 2－8、0－6 外的其它误识.因本系统只对 2－8、0－6 采用了二级识别，所以该类错误在各个实验中没有改变.

表 1 实验结果

| | | 2－8 误识率 | 0－6 误识率 | 其它错误率 | 总错误率 |
|---|---|---|---|---|---|
| 普通训练 | 训练集 | 1.19 | 0.00 | 0.00 | 1.19 |
| | 非训练集 | 5.75 | 1.46 | 2.23 | 9.44 |
| | 总体平均 | 5.44 | 1.36 | 2.08 | 8.88 |
| Bayes 法 | 训练集 | 0.83 | 0.00 | 0.00 | 0.83 |
| | 非训练集 | 5.34 | 1.37 | 2.23 | 8.94 |
| | 总体平均 | 5.04 | 1.28 | 2.08 | 8.40 |
| 参数加权算法 | 训练集 | 0.71 | 0.00 | 0.00 | 0.71 |
| | 非训练集 | 3.72 | 0.94 | 2.23 | 6.89 |
| | 总体平均 | 3.52 | 0.88 | 2.08 | 6.48 |
| 加权矩阵算法 | 训练集 | 0.24 | 0.00 | 0.00 | 0.24 |
| | 非训练集 | 3.41 | 0.69 | 2.23 | 6.33 |
| | 总体平均 | 3.20 | 0.64 | 2.08 | 5.92 |

### 3.4 结 论

从实验结果中可以看到：训练集的识别率远高于非训练集，原因在于训练集由孤立数字和双数字组成，语速较慢而且音变的影响小.同时可看到：训练方法的改进，识别率有一定的提高.Bayes 法的改进不明显，原因就在于数字在不同上下文中的特征有明显变化，靠改变样本并不能解决问题.参数加权算法和加权矩阵算法对 2－8 与 0－6 的区分有明显的作用，错误分别降低了 30% 以上，这说明各维特征参数内在的对不同音素的表现力是不同的和稳定的，区别对待有助识别率的提高.

感谢清华大学计算机系媒体所提供的实验条件，以及老师和同学协助收集数据.

## 参 考 文 献

[1] 顾良，刘润生.改进汉语数码语音识别中的语音特征提取性能[J].电路与系统学报 1997(4):.

[2] 李虎生，刘加，刘润生.高性能汉语数码串语音识别[J].电子学报，2001(5):.

[3] 许海天，吴及，王作英.汉语连续数字串语音识别系统[J].计算机工程与应用，2002(2):.

[4] 顾良，刘润生.利用声调判别提高汉语数码语音识别性能[J].清华大学学报(自然科学版)，1998(9):.

[5] 李虎生，刘加，刘润生.高性能汉语数码语音识别算法[J].清华大学学报(自然科学版)，2000(1):.

[6] Bahl L R, Brown P F, deSouza P V, et al. Maximum mutual information estimation[A]. In: IEEE Proc ICASSP'86. Tokyo

[7] Nphraim Y, Dembo A, Rabiner L. A minimum discrimination approach for Hidden Markov Modeling[A]. In: IEEE Proc ICASSP'88. New York

[8] Bahl L R, Brown P F, deSouza P V, et al. A new algorithm for the estimation of Hidden Markov Models with corrective training. [A] In: IEEE Proc ICASSP'89. Scotland

[9] 关存太，陈永彬，吴伯修.HMM 语音识别模型与一种修正训练算法[J].东南大学学报，1994 (1)

[10] 张春涛，吴善培.语音识别中基于最小误识率准则的区分训练方法[J].信号处理，1999(1):.

[11] Rabiner L., Juang B H. Fundamentals of speech recognition[M]. 北京：清华大学出版社，1999，200－240.