基于 FVQ/ HMM 的无教师说话人自适应

赵 力,邹采荣,吴镇扬
(东南大学无线电工程系,江苏南京 210096)

摘 要: 本文提出了一种新的语音识别方法,它综合了 VQ、HMM 和无教师说话人自适应算法的优点,在每个状态通过用矢量量化误差值取代传统 HMM 的输出概率值来建立 FVQ/ HMM,同时采用基于模糊矢量量化的无教师自适应算法,来改变 FVQ/ HMM 的各状态的码字,从而实现对未知说话人的码本适应. 本文通过非特定人汉语数码(孤立和连续数码)语音识别实验,把该新的组合方法同基于 CHMM 的自适应和识别方法进行了比较,实验结果表明该方法的自适应和识别效果优于基于 CHMM 的方法.

关键词: 语音识别; 模糊集; VQ; HMM; 无教师说话人自适应
中图分类号: TP391.42 **文献标识码**: A **文章编号**: 0372-2112 (2002) 07-0967-03

Unsupervised Speaker Adaptation Based On FVQ/ HMM

ZHAO Li, ZOU Cai-rong, WU Zhen-yang
(Department of Radio Engineering, Southeast University, Nanjing, Jiangsu 210096, China)

Abstract: We proposed a new speech recognition method by the integration of the VQ, HMM and an unsupervised speaker adaptation algorithm, it comply a VQ-distortion measure at each state instead of a output probability used by a traditional HMM, and uses an fuzzy adaptive VQ algorithm to alter the codewords for speaker adaptation. In this paper, the new combined method is compared with CHMM by the task of speaker-independent Chinese spoken digit (isolated/ connected) recognition, and the results confirmed that the good performance of the new method and superior to traditional CHMM.

Key words: speech recognition; fuzzy; VQ; HMM; unsupervised speaker adaptation

1 引言

几乎所有成功的语音识别方法都是基于统计的、概率的或信息理论的方法. 其中较具代表性的方法有矢量量化法(VQ)和隐马尔可夫模型法(HMM). VQ 法是由 Shore 和 Burton 首先提出,并应用于特定人数码识别[1],其主要优点是无需时间规正或进行动态时间伸缩. 但该方法对于由话者差别引起的语音特征的变化却无能为力. HMM 方法则适合于非特定人语音识别系统,因为它作为统计模型能够吸收由不同说话人引起的的语音特征的变化[2]. 然而完全去除话者差别是非常困难的,所以在实际应用中通常采用话者自适应的方法使未知说话人的语音去适应已知标准说话人的语音模型. 因此,对于非特定人语音识别系统来说,利用少量训练数据的说话人自适应技术是非常重要的. 尽管关于这方面的研究已有多种方法被提出[3~5],但是说话人自适应技术仍然是一个需要进一步研究的重要课题. 说话人自适应在方式上分为有教师和无教师两种,前者较后者实现起来容易,然而后者更具实用价值. 本文提出了一种基于模糊集矢量量化(Fuzzy VQ: FVQ)技术和 HMM 的无教师说话人自适应方法. 它综合了 FVQ、HMM 和无教师说话人自适应算法的优点,首先在每个状态通过用 VQ 误差尺度取代传统 HMM 的输出概率函数或输出概率矩阵来建立 FVQ/ HMM,同时采用基于 FVQ 的的模型参数估计和无教师自适应算法,来改变 FVQ/ HMM 的各状态的码字,从而实现对未知说话人的码本自适应. 由于汉语数码语音识别是语音识别的一个重要应用领域,因此本文用汉语数码语音的识别实验来检验这种新的综合方法的性能. 通过和基于连续隐马尔可夫模型(CHMM)的自适应方法的比较,证实了新方法的自适应和识别效果优于 CHMM 的方法.

收稿日期:2000-04-03;修回日期:2001-10-12
基金项目:国家自然科学基金;高校博士点基金

2 利用分段模糊聚类算法的 FVQ/ HMM 参数估计

在 FVQ/ HMM 中,对于一个给定的输入时间序列,识别系统将分别针对各个类别的模型,逐帧计算该序列的量化误差值. 得到最小累积量化误差值的模型所对应的类别即为识别结果.

代替标准 HMM 在每个状态的输出概率函数或矩阵,FVQ/ HMM 的模型参数由状态转移概率矩阵和每个状态的码本组成. 对于输入语音时间序列,FVQ/ HMM 利用各状态的码本逐帧计算输入序列的量化误差值,并按式(1)计算所有输入帧的累积误差值.

$$D = \min_{x} \sum_{i=1}^{T} \{ d(y_i, C_{x_i}) + d(x_{i-1}, x_i) \} \qquad (1)$$

这里 $y_1, y_2, \cdots, y_T$ 表示输入时间序列; $x_0, x_1, \cdots, x_T$ 表示状态序列(设共有 S 个状态); C_{x_i} 表示与状态 x_i 相对应的码本; $d(x_{i-1}, x_i)$ 表示从 x_{i-1} 状态转移到 x_i 状态的代价函数; $d(y_{i-1}, C_{x_i})$ 表示 y_i 和 C_{x_i} 间的距离. 该距离定义如下:

$$d(y_i, C_{x_i}) = \min_j d(y_i, c_j),\ c_j \in C_{x_i} \tag{2}$$

上述式(2)是采用最近邻准则(NN)计算误差的. 也可以采用其它误差准则,如 k 最近邻准则(KNN),即, $d(y_i, C_{x_i}) = \sum_{i=1}^{k} d_i$,式中 d_i 表示 y_i 和 C_{x_i} 中所有码字的第 i 个最小距离. 也可以和标准 HMM 类似采用概率计算形式,即, $-\log P(y|i,j) = -\log(\sum_{i=1}^{K} e^{-d_i})$. 最小累积误差距离通过如下式(3)所示的 Viterbi 算法求取:

$$\begin{cases} \text{令 } t = 1,2,\cdots,T;\ j = 1,2,\cdots,S \\ g(j,t) = \min_i\{ g(i,t-1) + d(y_i, C_j) + d(i,j) \} \\ D = g(S,T) \end{cases} \tag{3}$$

此处 $g(j,t)$ 表示 $y_1, y_2, \cdots, y_t$ 和 $x_1, x_2, \cdots, x_t\ (x_t = j)$ 间的最小累积距离,即:

$$g(j,t) = \min_{x, x_t = j} \sum_{i=1}^{t} \{ d(y_i, C_{x_i}) + d(x_{i-1}, x_i) \} \tag{4}$$

在标准离散 HMM(DHMM)中,Viterbi 得分值是由下式给出的:

$$\log P = \max_x \sum_{i=1}^{T} \{ \log P(y_i|x_{i-1}, x_i) + \log P(x_i|x_{i-1}) \}$$

$$= -\left[\min_x \sum_{i=1}^{T} \{ -\log P(y_i|x_{i-1}, x_i) - \log P(x_i|x_{i-1}) \} \right] \tag{5}$$

所以,通过对比 DHMM 可知,FVQ/HMM 中的 D、$d(y_i, C_{x_i})$ 和 $d(x_{i-1}, x_i)$ 分别相应于 DHMM 的 $-\log P$, $-\log P(y_i|x_{i-1}, x_i)$ 和 $-\log P(x_i|x_{i-1})$. FVQ/HMM 模型参数估计迭代过程可以描述如下:

1. 初始化,建立初始 FVQ/HMM.

2. 用 Viterbi 算法和回溯法(backtracking)解码实现观察序列相应于 FVQ/HMM 各状态的分段.

3. 收集每个状态的数据帧,并根据这些帧的数据重估新的码本.

4. 由解码时状态间转移次数的比率重估状态转移概率.

5. 重复(2)~(4)直至模型参数收敛.

模型参数重估采用对分段后对应各状态的观察序列进行模糊 C 均值(FCM)聚类算法实现. FCM 聚类是在引入模糊 c 划分后,对传统 k 均值聚类算法的模糊推广,它通过隶属度函数引入不确定性思想,实现对硬聚类算法的有效扩展,在实际应用中取得了非常优良的效果[6]. 首先定义 FCM 聚类算法目标函数为如下式(6)所示:

$$J_{FCM}(y, u, a) = \sum_{i=1}^{n} \sum_{k=1}^{M} u_k^m(y_i)\, d(y_i, a_k) \tag{6}$$

其中 $y = \{y_1, y_2, \cdots, y_n\}$ 是对应于某一状态的观察矢量序列; $a = \{a_1, a_2, \cdots, a_M\}$ 是类聚中心; $d(y_i, a_k)$ 表示距离; $u = \{u_1, u_2, \cdots, u_M\}$ 是 FCM 隶属度函数,它满足 $0 \le u_k(y) \le 1 (1 \le k \le M, \sum_{k=1}^{M} u_k(y) = 1)$; $m \in [1, \infty)$ 代表模糊度. 根据目标函数的 FCM 类聚算式如下:

$$\begin{cases} a_k = \dfrac{\sum_{i=1}^{n} u_k^m(y_i) \cdot y_i}{\sum_{i=1}^{n} u_k^m(y_i)}, & 1 \le k \le M \\ u_k(y_i) = \left(\sum_{j=1}^{M} \dfrac{d(y_i, a_k)^{2/(m-1)}}{d(y_i, a_j)^{2/(m-1)}} \right)^{-1}, & 1 \le k \le M, 1 \le i \le n \end{cases} \tag{7}$$

FCM 算法的收敛性在文献[7]中给出了证明. 在迭代计算聚类中心 a_k 及隶属度函数 u_k 直到收敛后,由新的聚类中心组成重估后的新码本.

3 基于 FVQ 的无教师说话人自适应算法

我们根据文献[8]介绍的一种基于 FVQ 的码本自适应方法,提出了 FVQ/HMM 的无教师说话人自适应算法. 定义 $\{a_i\}$ 为标准说话人码本的码字集合, $\{y_j\}$ 为未知说话人的训练数据序列,则提出的无教师说话人自适应算法如下:

(1)设定模糊度 $m \in [1, \infty)$ 的初始值.

(2)利用标准说话人码本的码字集合对未知说话人的训练数据序列进行矢量量化. 求落入相同码字空间的训练数据的质心 V_i.

(3)在各码字空间求该码字和该码字空间的 V_i 的误差矢量 Δ_i.

$$\Delta_i = V_i - a_i \tag{8}$$

(4)对于标准说话人码字集合中的每一个码字 a_i,利用该码字以外的其他码字对其进行模糊矢量量化. 隶属度 $u_j(a_i)$ 由式(9)求得.

$$u_j(a_i) = \frac{1}{\sum_k (d_{ij}/d_{jk})^{1/(m-1)}} \tag{9}$$

其中, d_{AB} 表示矢量 A 和 B 的欧氏距离.

(5)根据 Δ_i,用下式将标准说话人码本 $\{a_i\}$ 更新成被适应说话人的码本 $\{a_i'\}$,即:

$$a_i' = \alpha \sum_j (u_{ij}^m \Delta_j) \Big/ \sum_j u_{ij}^m + (1-\alpha)\Delta_i + a_i \tag{10}$$

(6)改变模糊度 m 的值,重复上面过程,直到 FVQ 收敛且 m 接近于 1.0 为止.

在以上的自适应算法中,我们将全部类别的每个 HMM 的每个状态的所有码字形成的一个码字集合作为 $\{a_i\}$,其码本尺寸为所有码字的数目.

4 实验结果

4.1 语音数据和分析方法

本文用 2 个实验来测试上述基于 FVQ/HMM 的无教师说话人自适应算法,实验中采用的标准说话人语音模型是一个多说话人模式 HMM. 第 1 个实验是孤立字汉语数码语音识别. 标准说话人模型的训练数据包括 50 个男性说话人对每个数码的 4 次发音(共 2000 个语音数据). 用于自适应测试的未知说话人是 3 个男性和 2 个女性:M1、M2、M3、F1、F2. 自适应训练数据集包括这 5 个说话人对每个数码的 20 次发音(共 1000 个数据). 测试数据集包括这 5 个说话人对每个数码的另 10 次发音(共 500 个数据). 第 2 个实验是连续汉语数码语音

识别.数码字符串长为 4,共有 35 种不同的字符串种类.连续汉语数码测试数据集由上文所提及的 5 个说话人对每种字符串的 3 次发音组成(共计 485 个数码字符串).

语音信号经 12kHz 采样,$1 - 0.98 Z^{-1}$ 的预加重,窗长 21.33ms(256 点),窗移 10ms 的汉明窗后,求出 10 维的 MFCC (Mel Frequency Cepstrum Coefficient)和 10 维的线性回归一阶 MFCC (△MFCC)以及 10 维的线性回归二阶 MFCC (△△MFCC)作为语音特征参数.

4.2　自适应实验和结果

为了比较,我们使用 CHMM 和 FVQ/ HMM 进行了比较实验.CHMM 中每个状态使用了高斯分布函数(均值矢量和全协方差矩阵).我们仅根据以上的无教师适应算法对 CHMM 的均值矢量进行了自适应.在自适应训练过程中,我们将全部类别的每个 CHMM 模型的所有状态的均值矢量的集合作为一个码本$\{a_i\}$,所有均值矢量的个数作为码本尺寸.

表 1 和表 2 给出了孤立和连续数码自适应识别实验结果.两表中 FVQ/ HMM 和 CHMM 都使用了 6 个状态,同时 FVQ/ HMM 每个状态对应码本的尺寸为 32,这是因为这样设定的实验结果最好.表中分别给出了用初始标准多说话人模式 FVQ/ HMM 和 CHMM 进行的识别实验结果、用 5 个未知说话人对每个数码的 10 次发音对初始模型自适应训练以后的识别实验结果以及 20 次发音对初始模型自适应训练以后的识别实验结果.

从表 1 可以看出,在用初始模型识别时,CHMM 的平均识别率比 FVQ/ HMM 分别低:3 男 0.9 %、2 女 4.3 %、总平 2.3 %.当分别用 10 次发音自适应训练后,CHMM 的平均识别率分别是:3 男 97.9 %、2 女 91.5 %、总平 95.3 %,而 FVQ/ HMM 分别是:3 男 99.4 %、2 女 98.2 %、总平 98.8 %.用 20 次发音自适应训练后,CHMM 的平均识别率提高了:3 男 1.2 %、2 女 4.2 %、总平 2.4 %,而 FVQ/ HMM 只提高了:3 男 0.2 %、2 女 0.4 %、总平 0.3 %.所以,对于孤立汉语数码语音识别来说,FVQ/ HMM 初始模型以及自适应模型的性能在一定程度上都优于 CHMM.并且,FVQ/ HMM 的自适应效果优于 CHMM.

表 1　孤立数码语音自适应识别实验结果(%)

(状态数:6,码本尺寸:32)

方法	数据个数	M1	M2	M3	平均	F1	F2	平均	总平均值
CHMM	——	95.4	97.9	98.5	97.3	84.0	87.3	85.7	92.6
适应 CHMM	10	95.9	98.7	99.0	97.9	90.4	92.6	91.5	95.3
	20	97.7	99.9	99.7	99.1	94.1	97.2	95.7	97.7
FVQ/ HMM	——	97.1	98.7	98.9	98.2	88.9	91.1	90.0	94.9
适应 FVQ/ HMM	10	98.9	99.9	99.5	99.4	98.2	98.1	98.2	98.8
	20	99.5	99.9	99.5	99.6	98.1	98.7	98.6	99.1

在连续数码语音识别实验中,我们采用一次通过 DP(One Pass DP)算法,用孤立数码 HMM 的连接来实现数码字符串的识别.从表 2 列出的试验结果可知,与孤立汉语数码语音识别类似,FVQ/ HMM 对连续汉语数码语音识别的性能同样优于 CHMM.而经过少量的 10 次发音数据自适应性训练后,FVQ/ HMM 的平均识别率显著提高,增长幅度达 2.3 %,高于 CHMM,而 CHMM 在经过 20 次发音数据自适应性训练后,才达到较高识别率.因此,该实验再次证实了基于 FVQ/ HMM 自适应算法对汉语数码语音识别的有效性.

表 2　连续数码语音自适应识别实验结果(%)

(状态数:6,码本尺寸:32)

方法	数据个数	M1	M2	M3	平均	F1	F2	平均	总平均值
CHMM	——	77.1	79.1	80.5	78.9	64.0	71.3	67.7	74.4
适应 CHMM	10	79.4	81.2	81.8	80.8	68.2	72.9	70.6	76.7
	20	80.9	82.2	83.9	82.3	75.1	77.4	76.3	79.9
FVQ/ HMM	——	78.9	80.8	80.9	80.2	68.9	74.5	71.7	76.8
适应 FVQ/ HMM	10	80.9	82.6	83.9	82.5	75.7	77.8	76.8	80.2
	20	80.9	82.9	84.3	82.8	75.9	77.8	76.9	80.4

5　结论

本文提出并评价了基于 FVQ/ HMM 的无教师说话人自适应方法.FVQ/ HMM 作为 HMM 的特殊形式,其模型参数数量较传统 HMM 少,模型学习对训练数据量要求不高;具有学习收敛速度快,适合于实时自适应学习;识别速度快,适合于实时大词汇量连续语音识别等特点.另外,和传统分段 VQ 识别方法相比,FVQ/ HMM 可以得到最佳的分段效果,且通过 FCM 聚类分析减少了码本的量化误差.在我们的实验中,FVQ/ HMM 取得了比标准 HMM 有效的识别性能.而且,利用 FVQ/ HMM 可以实现有效的无教师说话人自适应,并且,这种方法鲁棒性好,所需计算量较少,自适应效果好.我们提出的模型类似于具有 32 个混合高斯密度函数且每个状态共享方差的 CHMM,因此 2 种模型的性能比较有待研究.同时,和对角协方差矩阵的 CHMM 的性能比较也有待研究.

参考文献:

[1] J E Shore, D K Burton. Discrete utterance speech recognition without time alignment [J]. IEEE Trans, 1983, IT-29(4): 473 - 49.

[2] L Zhao, H Suzuki, S Nakagawa. A Comparison study of probability functions in HMMs through spoken digit recognition [J]. TRANS. INF and SYST, 1995, E78-D(6): 669 - 675.

[3] Yoshimitsu Hirata, Seiichi Nakagawa. Speaker adaptation of continuous parameter HMM [A]. In: IEICE, ed. International Conference Spoken Language Processing '90 [C], Kobe Japan: IEICE, 1990. 67 - 70.

[4] Gauvian J L, Lee C H. Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains [J]. IEEE Trans. on Speech and Audio Processing, 1994, 2(2): 291 - 298.

[5] Huo Qiang, Lee Chin-Hui. On-line adaptive learning of the continuous density hidden Markov model based on approximate recursive Bayes estimate [J]. IEEE Trans. on Speech and Audio Processing, 1997, 5(2): 161 - 172.

[6] 马小辉. 基于分段模糊 c-均值的连续密度 HMM 语音识别模型参数估计 [J]. 声学学报, 1997, 22(6): 550 - 554.

[7] Bezdek J C. A convergence theorem for the fuzzy ISODATA clustering algorithms [J]. IEEE Trans. 1990, PAMI(2): 1 - 8.

[8] 中村哲. Unsupervised speaker adaptation using fuzzy clustering [A]. 日本音响学会讲演论文集 '93 [C], yaokohama Japan: 日本音响学会, 1993. 43 - 44.