改进 MCE 训练算法在说话人识别中的应用①

吕洪艳, 李 荟

(东北石油大学 计算机与信息技术学院, 大庆 163318)

摘 要: 针对实际问题中训练数据不足的特点, 在对说话人建模时采用的是高斯混合模型—通用背景模型 GMM-UBM, 针对MCE训练算法中计算量大的显著问题, 对其进行改进, 改进的MCE算法不仅能使计算量减小, 而且识别性能更佳. 实验结果表明, 在高斯混合数与说话人数不同的情况下, 改进的 MCE 比传统 MCE 算法都要节省训练时间, 且随着高斯混合数与说话人数的增长, 节省的时间越多. 针对采用 MAP、MLLR、MAP\MLLR、EigenVoice 方法作自适应得到的说话人模型, 然后应用 MCE 算法与改进的 MCE 算法, 改进的 MCE 算法比传统 MCE 方法识别率更高.

关键词: 说话人识别; 高斯混合模型; UBM; MCE; 改进 MCE

Improved MCE Training Algorithm Used in Speech Recognition

LV Hong-Yan, LI Hui

(Institute of Computer and Information Technology, Northeast Petroleum University, Daqing 163318, China)

Abstract: In practical problems, it adopts GMM - UBM as the background model when the training data is insufficient in speaker recognition system. Aiming at large amount of calculation in MCE training algorithm,it improved MCE. The improved MCE algorithm not only can reduce the amount of calculation, but also can get better recognition performance. The experimental results show that, under the different number of gaussian mixture and speakers, the improved MCE algorithm saves more training time than the traditional MCE algorithm, and as the growth of the number of gaussian mixture and speakers, the more time saving. In view of the MAP, MLLR, MAP\MLLR and EigenVoice adaptation ways which used in speaker recognition system modeling, then using MCE algorithm and the improved MCE algorithm, the improved MCE algorithm has higher recognition rate than the traditional MCE algorithm.

Key words: speaker recognition; Gaussian mixture model; UBM; MCE; improved MCE

1 引言

说话人识别指的是对语音信号进行分析与特征提取, 根据构建的模型进行判断. 说话人识别技术以其方便性、准确性、可扩展性、经济性等优点, 广泛地应用于生物特征识别领域. 具体应用在计算机远程登录、语音 E-mail、数据库访问、电话购物等方面. 目前, 在实验室环境下, 说话人识别已取得较好的效果, 但在现实应用中, 由于训练数据不足、短话音、声音模仿、背景噪音等问题使说话人识别系统的性能下降. 而在这些实际问题中, 训练数据不足致使说话人模型得不到充分的训练是影响说话人系统性能的主要问题之一[1], 因此, 如何消除由于训练语音不充分致误识率较高从而使说话人系统实用性下降, 是需要解决的关键问题, 也是此领域一直研究的重要课题.

目前, 针对训练数据不充分问题, 主要有以下两大类解决方法: 一是说话人自适应方法, 二是模型参数估计算法. 说话人自适应方法通常分为模型参数自适应、说话人归一化、说话人聚类和谱变换四类[2]. 其中说话人归一化、说话人聚类和谱变换的优势是简单易行, 但是过于粗略, 无法充分刻画说话人的个性特征. 相比之下, 模型参数自适应通过对说话人的精确建模使系统性能增强, 从而应用更广泛. 模型参数自

① 收稿时间:2014-10-12;收到修改稿时间:2014-11-28

适应方法具体可分为三类: 一是基于最大后验概率 MAP 准则的自适应方法, 二是基于 MLE 准则的自适应算法, 常用算法是基于线性变换的 MLLR 算法, 主要对高斯密度函数的均值向量进行自适应. 另外, 结合 MAP 与 MLLR 两者的优点, MAP\MLLR 算法被提出并得到广泛应用, 这种方法使用 MAP 来刻画基于音素层次的差异, 使用 MLLR 来处理所有音素共同存在的差异[1]. 三是快速说话人自适应方法, 代表方法为基于特征音 EV 模型的变换方法. 这种方法能用少量的训练数据快速的调整模型以实现自适应, 它在口语对话信息查询系统中作用显著. 在参数估计方面, 基于最小分类错误 MCE(Minimum Classification Error)的训练方法常用来弥补模型参数自适应方法的局限性, 且能在一定程度上提高说话人识别系统的性能, 但 MCE 训练算法中计算量大也成为了限制其应用的显著问题. 本文针对 MCE 训练算法中计算量大的问题, 对 MCE 算法进行改进后应用到说话人识别领域, 以期解决现实应用中实际训练数据不足的问题, 从而提高说话人识别系统的性能.

2　说话人识别基线系统

目前在文本无关的说话人识别领域, 高斯混合模型(GMM)以其能够平滑的近似任意密度分布、易与其他信道畸变的补偿技术或噪音处理技术相结合等特点成为最为成功的一种模型. 它利用多个高斯分布的加权组合描述说话人的特征空间统计分布, 模型的混合度越高, 所需的训练语音也越多, 对说话人的特征统计分布的描述越细致, 识别性能也越好, 但是在复杂背景下, GMM 识别的性能较差, 鉴于此, Douglas A Reynolds 提出了 GMM-UBM 的模型[3]. UBM 是一个混合度非常高的混合高斯模型, 通常为 1024~4096 个混合度. 它由相对很长的语音数据(至少 1 小时)用 EM(Expectation Maximization)算法训练而成. 模型中的每个高斯“隐式”所对应的声学特征都得到了较充分地描述. 由于 UBM 的训练数据来自大量不同的说话人, 因此可以认为 UBM 描述的特征分布是所有说话人特征分布的并集[4]. 与 GMM 的说话人识别系统相比, GMM-UBM 有以下主要优点:

① UBM 具有更高的混合度. 在实际情况下, 训练数据很多时候是不充分的, 由于 UBM 是由大量语音训练得到的, 相对于 GMM 模型无法取得很高的混合度, 它却能有较高的混合度.

② 克服阈值选取的困难, 提高了识别性能. GMM-UBM 模型识别过程见图 1.

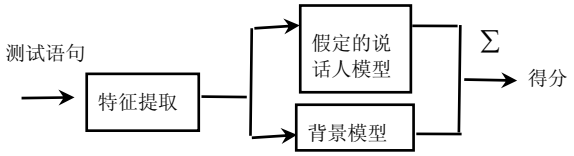

图 1　GMM-UBM 模型识别过程

GMM-UBM 模型具体识别中, 假定某一目标说话人, 通过特征处理会得到表示一段语音的特征矢量序列 $X=\{x_1,x_2,\cdots x_n,\}$, 然后计算目标说话人模型和背景模型上的似然分. 当满足公式(1)时, 系统才会接受其识别结果[4].

$$S(X)=\left[p(X|\lambda_{TAR})/p(X|\lambda_{BAK})\right]>\gamma \tag{1}$$

在公式(1)中, $p(X|\lambda_{TAR})$ 表示目标说话人模型输出特征矢量序列的似然度, $p(X|\lambda_{BAK})$ 表示伪冒者输出特征矢量序列的似然度, $S(X)$ 表示待测语音的似然比, γ 表示确认环节的判定阈值. 通常采用对数似然比来进行比较, 对数似然比能够增大不同说话人之间的可区分特性, 且能够减少对判定阈值的依赖性.

为了比较 GMM 与 GMM-UBM 的识别性能, 特作出以下实验. 数据集合由 50 个说话人的语音数据组成, 每人 30 秒左右, 测试语音为 20s, 闭集测试, GMM 的混合数为 32, GMM-UBM 混合数为 1024, 实验结果见图 2.

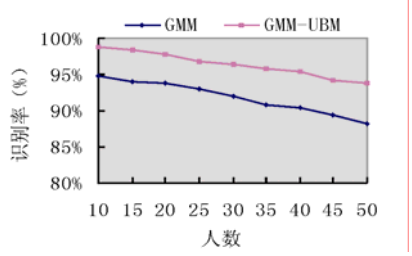

图 2　GMM 与 GMM-UBM 识别率对比

由图 2 可知, GMM-UBM 的识别率一直高于 GMM, 随着人数的增大, 两个系统的识别率都有所下降, 当人数为 50 人时, GMM 的识别率为 88.2%, 而 GMM-UBM 的识别率仍保持在 94.1%. 因此, 本文选取 GMM-UBM 为基准系统模型.

3　基于改进MCE训练的GMM-UBM参数调整

GMM-UBM 模型参数训练的目的是在给定的一

个人的所有训练数据时, 确定最符合该说话人特征的一组模型参数 λ. 目前常用的是最大似然估计 MLE 法, 但由于其依据假设的参数分布形式进行参数优化, 而假设的参数分布形式与其实际分布是不同的, 所以无法得到最佳性能. 针对这一局限性, 出现了一些分类器参数估计的准则, 其中以最大互信息估计 MMIE 和最小鉴别信息 MDI 应用最为广泛, 但它们的优化目标也不是直接和分类误差联系在一起的[5]. 鉴于此, 基于最小分类错误准则的最小分类错误算法 MCE 被提出来, 它直接将识别系统的误识率作为系统的优化标准来训练模型参数, 取得了很好的效果.

3.1 基于 MCE 的判别学习方法

假设有一个观测样本集合 $\Gamma=\{x_1,x_2,\ldots,x_M\}$, 其中每个 x_m (m=1,2, ..., M)是一个 d 维矢量, 并且属于 N 个类 C_i (i = 1, 2, ... , N)中的某一类. 对通常包含一个参数集和一个决策规则的分类器来说, 最小分类错误分类器设计的任务就是: 基于给定的样本集 Γ, 找出分类器的参数集 λ 以及相关的决策规则, 使误分类任何样本 x_m (m=1,2, ..., M)的概率最小, 一般误分类的概率用误识率来近似. 如果假设存在与误分类有关的惩罚或代价, 则这种分类器设计的目标就变为: 找出合适的分类器参数集 λ 和相关的决策规则, 使期望的代价最小[6].

第 1 步. 定义一个判别函数. 对于每一个类 C_i 有 $g_i(x;r)$ 使得 $C(x)=\arg\max_i g_i(x)$, 即样本向量 x 属于 $g_i(x;r)$ 取得最大值的 C_i 类.

第 2 步. 选择合适的错误分类量. 误分类测度不连续是不适合梯度运算的. 这里定义一种连续的误分类测度, 定义见公式(2).

$$d_k(x)=-g_k(x;\lambda)+\{[1/(N-1)]\sum_{j,j\neq k}g_j(x;\lambda)^{\eta}\}^{\frac{1}{\eta}} \tag{2}$$

上式右边第二项是所有其它竞争类似然度的几何平均值. 参数 η 可被看成为一个调整其它竞争类对整个判别函数贡献的权系数. 在搜索分类器参数 λ 的过程中, 通过变化 η 值可以找到许多潜在的分类, 一个极端的情况是当 $\eta\to\infty$ 时, 上式右边第二项中最大竞争类的判别函数将起主导作用, 即公式(3):

$$\eta\to\infty\text{ 时, }[\frac{1}{N-1}\sum_{j,j\neq k}g_j(x;\lambda)^{\eta}]^{\frac{1}{\eta}}=\max_{j,j\neq k}g_j(x,\lambda) \tag{3}$$

误分类测度变为:

$$d_k(x)=-g_k(x;\lambda)+g_i(x;\lambda) \tag{4}$$

其中 C_i 是除 C_k 外, 所有其它类中具有最大判别值所代表的类, 这是因为 $(N-1)^{1/\infty}\cong 1$. 显然, 在上面这种情况下, $d_k(x)>0$ 隐含着为误分类, $d_k(x)\le 0$ 为正确分类, 因此决策规则就变为一个标量值的判定问题.

第 3 步. 选择合适损失函数. 为了完成目标准则的定义, 将上面的误分类测度用到损失函数中, 以利于实现最小分类错误的目标, 代价函数可取 Sigmoid 形式, 见公式(5):

$$l_k(d_k(x))=\frac{1}{1+e^{-\xi(d_k(x)+\alpha)}},\quad \xi>0 \tag{5}$$

其中 ξ 是一个正常数, 它反映了接近决策边界 $d_k(x)+\alpha=0$ 时 Sigmoid 函数的陡度. 常数 α 反映了决策边界偏离原点($\alpha=0$)的情况.

当使用 0－1 之间的代价函数或任何上述平滑的 0－1 函数时, $d_k(x)>0$ 导致的惩罚将近似为一个误分类的计数值. 这样对任何未知样本 x, 分类器性能可按下式测得:

$$l(x;\lambda)=\sum_{k=1}^{N}l_k(x;\lambda)\mathrm{l}(x\in C_k) \tag{6}$$

其中 $l(\ell)$ 是一个逻辑值 ℓ 的 indicator 函数:

$$l(\ell)=\begin{cases}1 & if\ \ell\ is\ true\\ 0 & otherwise\end{cases} \tag{7}$$

第 4 步. 定义目标函数. 对一个给定的训练样本集 $\Gamma=\{x_1,x_2,...,x_M\}$, 有两种定义性能目标的方法: 一为整体平均损失, 另一为期望损失. 尽管这两种代价优化的算法间差别较小, 但对这两个不同目标的优化将导致不同收敛性能的梯度解. 这里采用整体平均代价和梯度下降算法实现最小分类错误, 具体见公式(8)和公式(9).

对上述给定的训练样本集 Γ, 其整体平均代价($L_0(\lambda)$)定义为:

$$L_0(\lambda)=\frac{1}{M}\sum_{m=1}^{M}\sum_{k=1}^{N}l_k(x_m;\lambda)\mathrm{l}(x_m\in C_k) \tag{8}$$

这一代价函数能用梯度下降法进行最小化:

$$\lambda_{u+1}=\lambda_u-\varepsilon\nabla L_0(\lambda_u) \tag{9}$$

其中 λ_u 代表第 u 次迭代时的参数集, ε 是反映收敛速度的常数, 而 $\nabla L_0(\lambda_u)$ 代表 $L_0(\lambda_u)$ 的梯度.

3.2 改进的 MCE 算法

MCE 算法在说话人识别中得到了成功的应用, 但是与传统的 HMM 算法相比, MCE 算法的最大问题就是计算量太大, 使训练时间大量增加. 由上述步骤可知, 运用这种方法首先要求通过基准的说话人训练方

法建立好模型，然后再对模型进行计算，修改参数，以做到最大化模型差异，使分类错误最小. 在进行模型参数修改的计算时，若有 N 个说话人的系统，每一类别的分类错误都需要计算 $N-1$ 类的判别函数，随着 K 的增加，计算量也大大增加，于是针对这个问题，提出了一种改进的 MCE 算法.

对每一个说话人定义一组改进的分类错误函数来代替一个函数. 其函数形式为公式(10):

$$d_{kj}(x)=-g_k(x;\lambda_k)+g_j(x;\lambda_j) \tag{10}$$

对每一个说话人，可以定义 $N-1$ 个 $d_{kj}(x)$. 按照 $d_{k,j}$ 的大小，对于每一个说话人 k 可以定义一个数值从小到大的 $d_{kj}(x)$ 序列，代表除了 k 以外的每一个类和 k 的相似度. 从而，可以定义一个参数集 $M(O,k,M)$ 来求分类函数，其中 $M<N-1$. 代表对于每一个说话人，用 M 个和他最接近的说话人的模型来生成模型参数.

此时，相应的子代价函数为:

$$l_{k,j}(d_k(x))=\frac{1}{1+e^{-\xi(d_k(x)+\alpha)}} \tag{11}$$

对应一个特征值的代价函数为:

$$l(x,\lambda)=\sum_k\sum_{j\in K(O,k,M)}l_{k,j}I_k(x\in C_k) \tag{12}$$

同 MCE 算法一样，采用梯度下降法来实现 $l(x,\lambda)$ 的最小化.

4　实验结果及分析

4.1 识别率对比实验

这里采用的语音数据分为两个部分：训练语音和测试语音. 其中 UBM 是由 30 个说话人，每人 2-3 分钟，总长大约 90 分钟的纯净语料训练得出. 说话人模型由 1024 维的 GMM-BUM 分别采取 MAP、MLLR、EigenVoice 和 MAP 与 MLLR 相结合的综合渐进方法作自适应生成，闭集测试. 测试语料选择的是 20 个说话人，每人大约 40 秒的语音. 测试部分的 20 个说话人中有一部分是训练使用的 30 个说话人的子集. 说话人测试语音分段，段长从 5 秒到 30 秒不等. 其中 EigenVoice 方法里特征音维数取 10，MCE 中的 $\eta=1,a=0.1,b=-1,\varepsilon=0.01$ ，M 取 20. 表 1 为 UBM 分别用 MAP、MLLR、MAP\MLLR、EigenVoice 作自适应(自适应时长为 5s)后，再引入 MCE 模型与改进的 MCE 对系统改进前后识别率的对比情况.

表 1　MCE 算法改进前后识别率的对比

迭代次数 (n)	MAP 作自适应		MLLR 作自适应		MAP\MLLR 作自适应		EigenVoice 作自适应	
	识别率 61.3%		识别率 75.6%		识别率 77.1%		85.2%	
	MCE (%)	改进 MCE (%)	MCE (%)	改进 MCE (%)	MCE (%)	改进 MCE (%)	MCE (%)	改进 MCE (%)
200	64.2	64.3	80.1	80.3	80.5	80.8	88.2	90.2
150	64.0	64.2	80.0	80.4	79.0	79.4	88.1	90.2
100	63.8	64.1	79.7	80.3	78.6	79.0	87.7	90.1
50	63.5	64.0	79.5	80.2	78.2	78.6	87.5	89.8

从表 1 可以看出，无论 GMM-UBM 采用 MAP、MLLR、MAP\MLLP 或 EigenVoice 方法作自适应，而后又用 MCE 与改进的 MCE 训练方法，系统的识别率都随着梯度下降法迭代次数的增多而提高，而改进的 MCE 比传统 MCE 识别率更高一些，可见此改进算法可行. 当梯度下降法迭代次数取 200 时，系统的识别率达到最优，下面比较梯度下降法迭代次数 $n=200$ ，用不同的方法做自适应时，随着自适应时长的增长，采用 MCE 算法与采用改进的 MCE 算法所导致的识别率的差异，具体见图 3、图 4、图 5、图 6.

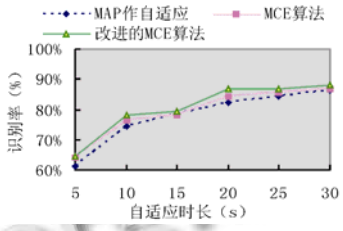

图 3　基于 MAP 应用 MCE 与改进 MCE 识别率对比

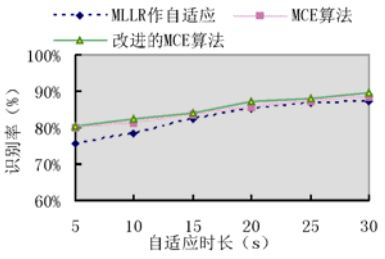

图 4　基于 MLLR 应用 MCE 与改进 MCE 识别率对比

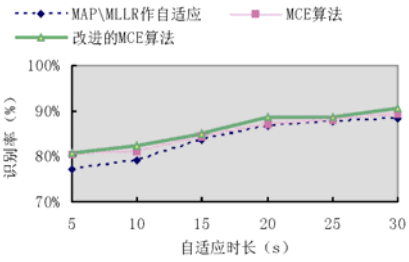

图 5 基于MAP\MLLR 应用 MCE 与改进 MCE 识别率对比

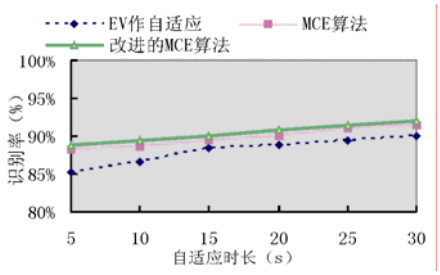

图 6 基于 EV 应用 MCE 与改进 MCE 识别率对比

从图 3、图 4、图 5、图 6 可以看出, 无论采用 MAP、MLLR、MAP\MLLR、EigenVoice 方法作自适应得到的说话人模型, 而后应用 MCE 算法与改进的 MCE 算法, 系统的识别率都随着自适应时长的增长有所改进, 同时, 从图中我们也可以看出, 改进的 MCE 算法比传统 MCE 方法识别率更高一些. 当采用 EigenVoice 作自适应, 自适应时长为 30s, 且 M 的值取 20 时, 系统识别率最高, 达到了 92%.

4.2 训练时长对比实验

UBM 用 EigenVoice 作自适应(自适应时长为 5s), 首先假定说话人数为 20, 在高斯混合数不同的情况下, 分别引入 MCE 模型与改进 MCE 对系统改进, 训练时长对比见表 2. 然后假定高斯混合数 N 为 1024, 在说话人数不同的情况下, 分别引入 MCE 模型与改进 MCE 对系统改进, 训练时长对比见表 3.

表 2 不同高斯混合数下, MCE 与改进 MCE 训练时间比较

高斯混合数(N)	训练时长(s)	
	MCE 算法	改进的 MCE 算法
N=256	186	159
N=512	224	189
N=1024	311	265
N=2048	632	530
N=4096	1336	1105

表 3 不同说话人数下, MCE 与改进 MCE 训练时间比较

说话人数(K)	训练时长(s)	
	MCE 算法	改进的 MCE 算法
K=10	204	189
K=20	311	265
K=30	472	396
K=40	635	516
K=50	845	650
K=60	1183	910

从表 2 可以看出, 在高斯混合数不同的情况下, 改进的 MCE 比传统 MCE 算法要节省训练时间, 平均节省时间约 19%, 随着高斯混合数的增长, 节省的时间越多, 当 N 取 1024 时, 节省约 21%的时间. 从表 3 可以看出, 在说话人数不同的情况下, 改进的 MCE 比传统 MCE 算法节省训练时间更多, 平均节省时间约 25%, 而且随着说话人数的增多, 节省的时间越多, 当说话人数为 60 时, 节省的时间最多, 约节省 30%的时间.

5 结语

针对训练数据不充分问题, 本文选取 GMM-UBM 为基准系统模型. 模型参数训练的目的是确定最符合该说话人特征的一组模型参数, 基于最小分类错误准则的 MCE 方法用来训练模型参数效果较好, 但缺点是计算量大. 针对传统 MCE 判别学习方法的缺点, 本文提出了改进的 MCE 方法并应用到说话人识别模型中, 实验结果表明, 与传统 MCE 算法相比, 改进的 MCE 方法不仅能大大减小计算量, 而且能在一定程度上提高系统性能.

参考文献

1 吴朝晖,杨莹春.说话人识别模型与方法.北京:清华大学出版社,2009.
2 胡政权.说话人识别中语音参数提取方法的研究[硕士学位论文].南京:南京师范大学,2013.
3 徐永华.基于 GMM-UBM 模型的语种识别[硕士学位论文].昆明:云南大学,2010.
4 王秋雯.基于 GMM-UBM 的快速说话人识别方法[硕士学位论文].哈尔滨:哈尔滨工业大学,2011.
5 王琳琳.说话人识别中的时变鲁棒性问题研究[硕士学位论文].北京:清华大学,2013.
6 李荟.基于自适应和 MCE 的说话人识别模型训练技术[博士学位论文].哈尔滨:哈尔滨工业大学,2007.