特征音方法在说话人识别中的应用①

李 荟[1], 赵云敏[2]

[1](东北石油大学 计算机与信息技术学院, 大庆 163318)
[2](大庆油田 第一采油厂, 大庆 163318)

摘 要: 针对现实中训练数据不足的特点, 在说话人建模时采用高斯混合模型—通用背景模型(Gaussian Markov Model-Uniform Background Model, GMM-UBM), 主要从说话人识别模型的自适应方法和参数估计方法两个方面, 研究如何提高说话人识别系统的识别率. 在说话人识别模型自适应方面, 改进传统的用最大后验概率 MAP (Maximum A Posterior Probability) 得到说话人模型的方法, 将语音识别中的最大似然线性回归 MLLR (Maximum Likelihood Linear Regression)和基于特征音(EigenVoice, EV)的自适应方法, 应用到说话人识别模型自适应当中, 并将其与 MAP 方法进行比较.

关键词: 说话人辨认; 最大后验概率; 最大似然线性回归; 特征音

EigenVoice Means Used in Speech Recognition

LI Hui[1], ZHAO Yun-Min[2]

[1](School of Computer and Information Technology, Northeast Petroleum University, Daqing 163318, China)
[2](Firstoil Production Plant of Daqingoilfield, Daqing 163318, China)

Abstract: This thesis adopts GMM-UBM when model speaker recognition system considering of lacking data. In the aspect of adapting in speaker recognition system modeling and parameter estimating, attentions are put on researching in how to improve recognition rate. In the side of adapting in speaker recognition system modeling, we will ameliorate conventional MAP (Maximum A Posterior Probability) means to get speaker recognition model, apply MLLR (Maximum Likelihood Linear Regression) and EigenVoice adaptation ways which used in speech recognition into adapting in speaker recognition system modeling, and compare the results with MAP means.

Key words: speaker identification; MAP; MLLR; eigen voice

1 引言

为了提高信息的安全性, 同时保证用户服务的便利性, 利用人自身的生物特征进行用户身份认证技术—生物特征识别技术, 越来越引起人们的关注. 尽管在实验室理想语音环境下, 说话人识别系统已经取得了比较好的效果. 但是在现实应用当中, 说话人识别系统的性能受到很多因素制约, 系统的识别结果还不能让人满意.

在这些影响因素中训练语音的不充分是致使系统性能下降的主要原因之一[1]. 概括地说, 训练数据不足, 说话人模型得不到充分的训练, 就不能真实的反映说话人的语音信息, 这样导致的结果是用它来做识别误识率会比较高. 因此, 如何消除由训练数据不足导致的识别率低的问题, 是提高说话人识别系统实用性不得不考虑的一个关键, 它的研究一直是说话人识别领域的一个重要课题.

目前, 对训练数据不足问题的解决方法主要有自适应背景模型的方法和改进模型参数估计算法两大类方法[2]. 通常把说话人自适应方法大致分为以下四类[3]: 说话人归一化(Speaker Normalization)、模型参数自适应(Model Parameter Adaptation)、说话人聚类(Speaker Clustering)和谱变换(Spectral Transforming). 在上述四种自适应方法中, 说话人归一化、说话人聚类和谱变换都比较简单易行, 但是由于过于粗略, 不能充分的刻画

① 收稿时间:2013-01-30;收到修改稿时间:2013-03-18

说话人的个性特征, 而模型参数自适应因其良好的性能应用最为广泛. 模型参数自适应通过提高模型对目标说话人的精确建模来增强系统性能. 在实际的应用中, 绝大多数系统都是综合使用多种自适应技术的. 比如说话人聚类和模型参数转换就常一起使用; MAP 和向量域平滑(Vector Field Smoothing, VFS)的自适应方法, 在系统的渐进性和快速性两方面都取得了较好的效果等等.

在语音识别领域, 上述自适应方法的应用已经使得系统的识别率有了很大的提高. 虽然说话人识别和语音识别有一定差异, 但是本文将自适应方法修改以后, 应用到说话人识别领域, 并提高了说话人识别系统的性能.

2 说话人识别基线系统

GMM 是目前实现说话人识别的主要技术之一. 对 GMM 参数的估计通常采用 MLE (Maximum Likelihood Estimation) 算法, 其中较为典型的是 Baum-Welch 算法[4]. 当训练语料较少时, 为每一个人的语音整体建立一个 GMM 模型是不准确的, 因为它没有被充分的训练过, 不能正确的反映说话人语音的概率分布. 所以一般考虑用多说话人的语音训练出一个高阶的 GMM, 称为通用背景模型 UBM, 用来表示说话人无关的语音特征分布. 然后分别用每个说话人的语音自适应 UBM 得到当前说话人的模型. 目前, 通常所选择的自适应方法都是传统的 MAP 方法. 这种方法在训练语料不足时效果不太理想. 可以尝试将语音识别中常用的自适应方法引入到说话人识别模型自适应中.

GMM-UBM 是性能非常优秀的说话人识别模型, 该模型已经成为 PC 平台上说话人识别中性能最好的模型. 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. 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. 因此 GMM 中每个高斯“隐式”对应的声学特征得到了充分地描述. 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话者特征分布的并集, 具有背景意义.

3 基于特征音的模型自适应方法

主元分析法[5](Principal component Analysis, PCA), 或称主分量分析, 是多元统计分析方法中一种最主要的分析方法, 它是建立在矢量表示的统计特性基础上的变换. 它研究如何将多指标的问题转化为较少的综合指标的一种重要方法, 即将高维空间的问题转化到低维空间去处理, 使问题变的比较简单、直观. 而这些较少的综合指标之间互不相关, 又能提供原有指标的绝大部分信息.

PCA 通常又被称为是特征向量法或是 Karhunen-Loeve(K-L)变换法. 它是一种基于目标统计特性的最佳正交变换. 称其为最佳变换是因为它具有重要的优良性质: 使变换后产生的新的分量正交或不相关; 以部分新的分量表示原矢量均方误差最小; 使变换矢量更趋稳定、能量更趋集中等, 这使它在特征选取, 数据压缩等方面都有极为重要的应用, 而近几年也开始应用于特征语音自适应中.

主元分析法能有效地降低数据的维数. 假定一组 n 维的数据向量 $X=\{x_1,x_2,\cdots x_n\}, x_i\in R^n, n>>1$. 主元分析就是找到从 n 维到 r 维的映射: $X\to Y, X\in R^{r\times n}$, $Y\in R^{r\times n}, r<<n$ 使 Y 与 X 的误差最小, $r<n$ 从而达到降维的目的.

3.1 特征语音

特征语音自适应方法从本质上来说, 也是一种基于模型的说话人自适应算法, 但它与其它基于模型的算法又有很大的不同(如 MAP 重估算法是用来修改模型的参数). 它是一种基于说话人空间(或特征空间)映射的隐马尔可夫模型自适应方法. 这种方法把自适应模型认为是从一组参考说话人离线获得的少数低维基本矢量(特征矢量或主元)的线性组合, 因此可大大减少自适应数据估计的参数数目. 这种方法又与其它基于说话人空间的算法不同, 它保证基本矢量互相正交同时又保留了参考说话人之间差异的最主要的成分.

假设 $\{\mu_1,\cdots\mu_R\}$ 是一组训练好的 R 个特定说话人模型. 其中, $\mu_r=[\mu_{r,11}^T,\cdots\mu_{r,NK}^T]^T$ 为展开的第 r 个说话人模型所有高斯均值矢量的 D 维超矢, 维数 $D=n\times N\times K$, n,N,K 分别为语音帧矢量的维数、GMM 的状态数和每个状态包含的混合高斯成分的数目, 而 $\mu_{r,NK}^T$ 表示第 r 个参考说话人第 j 个状态的第 k 个高斯均值矢量. 特征语音方法就是要寻找由 $\varpi(i), i\in\{1,\cdots,p\}$ 所张成的 p 维线性子空间(特征空间), $\varpi(i)$ 被称为第 i 个基本矢量, 又称为特征语音.

为求出 R 个特定说话人的特征空间基本矢量, 定义协方差矩阵如下式

$$S=\frac{1}{R}\sum_{r=1}^{R}(\mu_r-\bar{\mu})(\mu_r-\bar{\mu})^T \tag{1}$$

其中 $\bar{\mu}$ 为 R 个特定说话人的统计平均高斯均值矢量, 即

$$\bar{\mu}=\frac{1}{R}\sum_{r=1}^{R}\mu_r \tag{2}$$

为找到特征空间的 $p(p<R<<D)$ 维基本矢量(特征矢量或主元), 我们用主元分析法(PCA) 求取特征矢量矩阵 $W=[\varpi(1),\cdots\varpi(p)]$. $\varpi(i),i\in\{1,\cdots,p\}$ 即为求得的特征空间基本矢量(特征语音). 那么目标说话人自适应模型的高斯均值矢量(超矢)估计值 $\hat{\mu}$ 能够 p 从个特征空间基本矢量的线性组合得到, 如下式

$$\hat{\mu}=\sum_{i=1}^{P}\varpi(i)x(i)+\bar{\mu}=WX+\bar{\mu} \tag{3}$$

式中 $X=[x(1),\cdots x(p)]^T$ 是特征语音的权系数矢量.

图 1 说明了特征空间描述和特征语音自适应的基本思想.

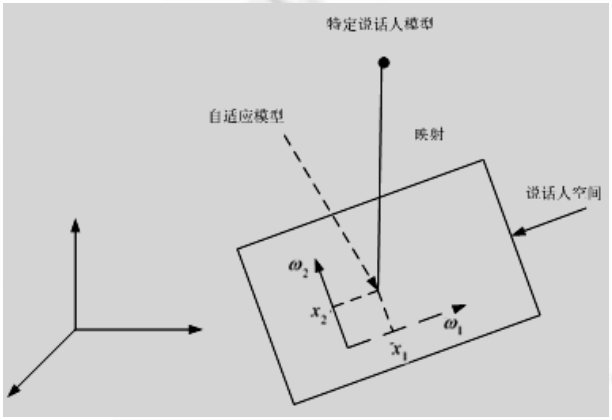

图 1　特征空间和特征语音自适应示意图

3.2 语音权系数的求解(最大似然特征分解)

求解特征语音的权系数即为用目标说话人的自适应数据来估计式(3)中的 $x(i).i\in\{1,\cdots,p\}$. 它可以通过给定的自适应数据(观察矢量序列) O 用最大似然特征分解求得, 如下式

$$\hat{x}=\arg\max_{x}\log p(O\mid W_x) \tag{4}$$

为说明具体求解权系数的过程, 首先定义一个似然辅助函数[6], 为使似然辅助函数 $Q(\lambda,\hat{\lambda})$ 最大化, 用高斯消元法, 即可求出 p 个特征语音权系数. 这样就可以得到说话人自适应的高斯均值估计值, 再用得到的这一估计值, 计算概率的新值, 反复进行迭代, 直到权系数收敛为止(也就是权系数的绝对误差值小于某个设定的值).

图 2 描述了特征语音自适应方法全过程. 首先离线训练一组个特定说话人的 GMM 和一个非特定人模型, 然后再从每一个特定说话人模型中提取一个包含自适应参数(如 GMM 输出高斯均值矢量)的超矢, 在超矢中参数排列的顺序是任意的, 但必须保证在所有特定说话人中提取的参数的数目 D 和参数的顺序相同. 再用降维技术—主元分析法(PCA), 对维数为 D 的 R 个超矢求取 R 个特征矢量(主元), 按照它们方差贡献率的大小保留前 $p,(p<R<<D)$ 个最大特征值对应的特征矢量, 称这些保留的特征矢量为特征语音. 在在线自适应阶段, 通过离线获得的特征语音和非特定人的 GMM, 对新说话人的自适应数据, 利用最大似然特征分解反复进行迭代, 直到特征语音权系数收敛, 最后就可以得到说话人自适应的高斯均值矢量(超矢)估计. 值得一提的是, 在用最大似然估计进行第一次迭代时, 概率值 $\gamma_m^{(s)}(t)$ 是用非特定人模型参数计算, 而在随后的迭代过程中, 它们是用自适应的模型参数(估计的自适应模型均值参数)估计, 而每一个高斯分布的方差 $C_m^{(s)}$ 来自非特定人模型, 但在迭代过程中, 保持不变.

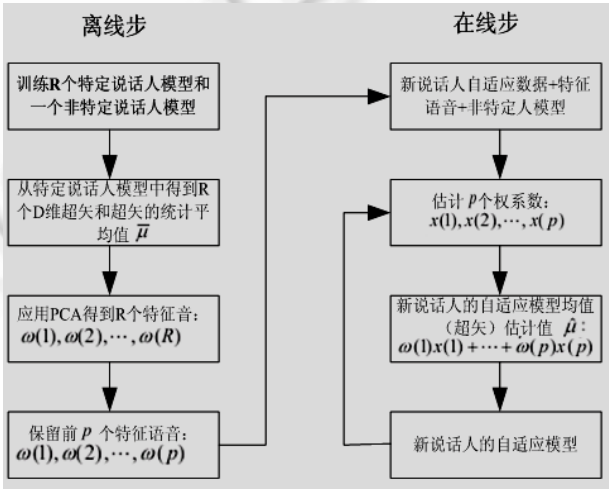

图 2　特征语音说话人自适应方块图

从图 2 也可以看出, 在开始识别前, 特定说话人的训练和主元分析这些大量的计算处理是离线进行的, 这就使得新说话人的自适应过程既相对简单又使得识别系统数据处理时间较少, 适合实际应用的要求.

4 实验结果及分析

本文中使用的实验数据库是在实验室环境下录制的. 包括 30 个人, 15 男 15 女, 分为训练集和测试集, 采样频率 11025Hz, 16bit 量化, 端点检测, 分帧帧长 23.2ms(256 点), 帧移 11.6ms(128 点), 加 30ms 汉明窗以克服 Gibbs 现象, 最后逐帧计算语音特征参数.

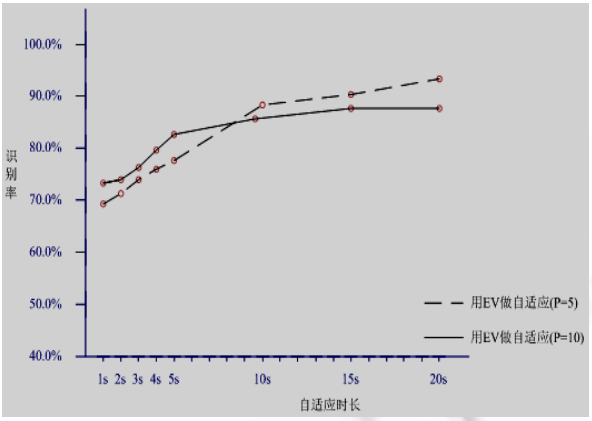

图 3 使用 EV 方法作自适应时在不同训练时长情况下的识别率对比

图 3 中虚线和实线分别为特征音维数 P 为 5 和 10 时, 通过 EV 自适应以后系统的识别率. 从图可看出, 用 EV 做自适应时, 无论 P 的取值为多少, 识别率都随着训练语料的增加呈现增长的趋势. 特征音维数为 10 时系统的识别率比特征音维数是 5 时要高. 这是由于特征音维数越多, 用 PCA 降维以后对说话人信息的保留就越多, 能够更准确地刻画说话人的模型. 当 P 取 10 时, 自适应时长为 20s 时, 系统的识别率较好, 为 93.5%.

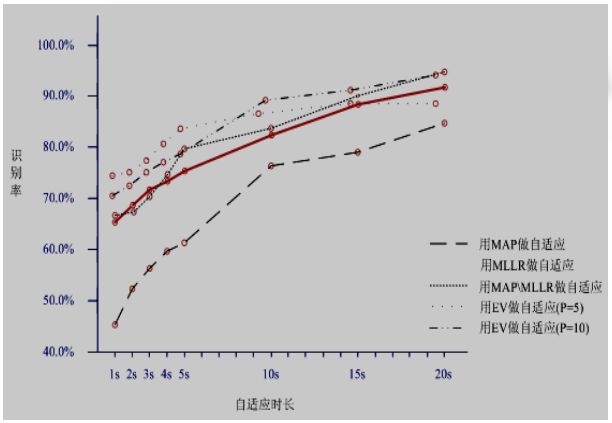

图 4 几种自适应方法在不同训练时长情况下的识别率对比

从图 4 可以看出, 用 MAP、MLLR、MAP\MLLR 相结合、EV 分别做自适应时, 效果最差的是 MAP 方法, 它在训练数据较少时, 性能不理想. 但是随着自适应数据的增多, 它增长的速度是最快的. MLLR、MAP\MLLR、EV 方法, 其识别率随着自适应数据的增多也增多, 但是增加的相对缓慢, 逐渐趋近于一种饱和状态. 当自适应时长小于 20s, 采用这三种方法做自适应时, 识别率都比用 MAP 有不同程度的提高. 自适应数据小于 4s 时, MAP\MLLR 相结合的自适应方法效果最好, 达到了 74.6%; 当自适应数据达到 5s 以上时, 特征音方法的效果最好, 明显要优于其它 3 种方法, 当特征空间维数取 10、自适应时长为 20s 时, 识别率达到最高, 为 93.5%.

5 结语

针对训练数据不足的特点, 本文在对说话人建模时采用 GMM-UBM, 改进传统的用最大后验概率 MAP 得到说话人模型的方法, 并将语音识别中的最大似然线性回归 MLLR 和基于特征音的自适应方法, 应用到说话人识别模型自适应当中, 将其与 MAP 方法进行比较. 实验结果表明, 针对不同训练语料的数量, 应用合适的自适应方法进行说话人识别模型的自适应, 系统的识别率会有显著的提高.

参考文献

1 李荟.基于自适应和 MCE 的说话人识别模型训练技术[学位论文].哈尔滨:哈尔滨工业大学,2007.

2 俞一彪,王朔中.文本无关说话人识别的全特征矢量集模型及互信息评估方法.声学学报,2009,(6):125–129.

3 Zhou G, Mikhael WB. Speaker Identification Based on Adaptive Discriminative Vector Quantisation.Vision, Image and Signal Processing, IEE Proc, 2012(6): 754–760.

4 Wolf MB, Park WKO, Blowers JC, Misty K. Toward Open-Set Text-Independent Speaker Identification in Tactical Communication. Computational Intelligence in Security and Defense Applications, Syracuse University, 2011:7–14.

5 韩纪庆,张磊,郑铁然.语音信号处理.北京:清华大学出版社,2004.12–14.

6 Barras C, Meignier XZ, Gauvian S, Multistage JL. Speaker Diarization of Broadcast News. Audio, Speech an Language Processing, IEEE, 2011(5):1505–1512.