文章编号:1002-8684(2007)06-0053-03

基于GMM的实时说话人识别系统*

·论文·

胡益平，蔡　骏，洪青阳
(厦门大学 计算机科学系，福建 厦门 361005)

【摘　要】介绍了一个基于GMM实时说话人识别系统的设计与实现,系统具有实时说话人辨认和实时说话人确认功能。在实验室条件下,对不同的高斯混合密度个数及采样率进行了测试,测试了模型的自适应性能。实验表明系统具有较好的识别准确率。

【关键词】说话人识别；实时系统；高斯混合模型
【中图分类号】TN912　　【文献标识码】A

A Real Time Speaker Recognition System Based on GMM

HU Yi-ping, CAI Jun, HONG Qing-yang
(Department of Computer Science, Xiamen University, Xiamen Fujian 361005, China)

【Abstract】 The design and implementation of a real-time speaker recognition system which is based on GMM (Gaussian Mixture Model) are presented. The system has the characteristics of real time speaker identification and real time speaker verification. In the lab environment, the performance of the system, as well as the model adaptation, has been fully tested with GMMs of different numbers of Gaussian mixtures and different sampling rates. The testing results show that the GMM-based system has a satisfactory correctness in performing speaker recognition.

【Key words】speaker recognition; real-time systems; GMM

1　引言

随着信息技术的发展,特别是人工智能技术,许多生物认证技术都可以用计算机来实现,如指纹识别、虹膜识别等，基于声纹认证的说话人识别技术也得到了飞速发展。与其他生物认证技术相比,声纹识别有其独特的优越性:(1)获取声音的方法非常简单,只要一个传声器即可,而指纹、虹膜等都需要较专业的仪器;(2)声纹识别的算法复杂度较低,能够做到实时处理、实时识别,因而具有广泛的实际应用价值;(3)支持远程识别,可以通过电话或网络远程认证。

自20世纪60年代以来，说话人识别技术在一些特定领域得到了广泛应用,如门禁系统、网上银行、刑事侦查等。目前中国的许多科研机构在研究基于汉语说话人识别技术上,取得了一定成就。说话人识别的难点主要有:(1)一个人的声音不是固定不变的,身体状况不同、年龄不同都会带来不同的声纹特征;(2)相对于英语,基于汉语的说话人识别有个突出的难题,那就是汉语的方言非常多,就算是同样讲普通话,口音也会有很大差别，因此如果训练语音与测试语音的说话人来自不同的方言区,那么有可能造成较大的识别误差。

目前,说话人识别的方法主要有模板匹配,矢量量化(Vector Quantization,VQ)、人工神经网络(Artificial Neutal Network,ANN)、隐马尔科夫模型(Hidden Markov Model,HMM)和高斯混合模型(Gaussian Mixture Model,GMM)[1]等。GMM方法具有独特的优越性,特别是对于文本无关(Text-independent)的说话人识别系统。研究[1-3]表明,GMM在文本无关的说话人辨认和说话人确认中都取得了较好的识别效果。另外,GMM算法的复杂度较低，特别是一些改进的GMM算法显著加快了处理速度,使说话人识别系统能够实现实时处理。

2　实时说话人识别系统框架

实时说话人识别系统的软件系统结构如图1所示，系统可以分为2个主要模块：训练模块和识别模块。识别模块中又可分为说话人辨认(speaker identification)和说话人确认(speaker verification)子模块。在说话人确认模块子模块中，需要用到事先训练好的通用背景模型(Universal Background Model,UBM)[3]。

*[基金项目]厦门大学“985工程”二期“信息技术”创新平台项目资助,项目编号0000-X07204

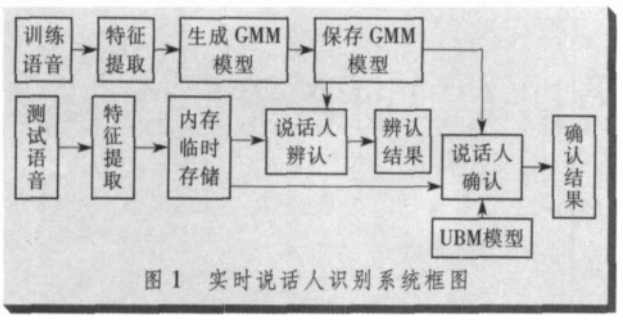

图1 实时说话人识别系统框图

3 基于GMM模型实时说话人识别系统

3.1 GMM模型

GMM模型是目前被广泛应用的说话人识别模型，它具有与文本无关、处理速度快、识别效果好等优点。在GMM模型[3]中，观测语音的概率密度函数是M个加权的高斯概率密度函数之和，表示为

$$p(x|\lambda)=\sum_{i=1}^{M} p_i b_i(x) \tag{1}$$

其中，$p(x|\lambda)$是某个说话人λ观测值x的概率密度函数；x是一个D维的随机向量；M是高斯分量的个数；p_i是各高斯分量的权重，满足$\sum_{i=1}^{M} p_i=1$；每个$b_i(x)$是1个高斯概率密度函数，表示为

$$b_i(x)=\frac{1}{(2\pi)^{D/2}|\boldsymbol{\Sigma}_i|^{1/2}}\exp\left[-\frac{1}{2}(x-\boldsymbol{\mu}_i)^{T}\boldsymbol{\Sigma}_i^{-1}(x-\boldsymbol{\mu}_i)\right] \tag{2}$$

其中，$\boldsymbol{\mu}_i$为数学期望；$\boldsymbol{\Sigma}_i$为方差矩阵。完整的GMM模型(用λ表示)由所有高斯分量的期望、方差矩阵和权重组成，可以表示为

$$\lambda=\{p_i,\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i\},\quad i=1,2,\cdots,M \tag{3}$$

3.2 语音信号的前期处理与特征提取

说话人识别首先需要将语音模拟信号转变成数字信号并提取出特征参数。语音的前期处理主要包括分帧、加窗、傅里叶变换、提取MFCC参数等步骤。由于文中系统对语音信号进行实时采样训练、实时识别，因此特征参数在提取后不保存成MFCC文件，而是将参数存放在内存中，直接用于系统的训练和识别。

3.3 UBM模型

在说话人确认系统中，需要好的竞争者模型，文中系统采用的方法是训练1个由UBM来充当竞争者的模型。由于各种汉语方言存在差异，在数据集构建时应尽可能采用属于同一个方言区的训练语音与测试语音，在系统中采用了南方口音进行训练和测试。UBM的基本思路是利用大量的非假设说话人语音来训练一个单独的GMM模型。训练方法有很多种，可以将所有的语音放在一起用最大似然(Maximum Likelihood, ML)方法对模型加以训练，也可以用一些区别性训练方法，如最大互信息法(Maximum Mutual Information, MMI)[4]、最大模型距离法(Maximum Model Distance, MMD)[5]、最小分类误差法(Minimum Classification Error, MCE)[6]等方法来获得区别性更好的UBM。

3.4 说话人模型的自适应生成

训练单个说话人的模型通常可简单采用基于自身语音进行训练的ML方法。另外，在说话人辨认系统中，还可采用区别性训练的方法来得到整体区分度更好的各个说话人模型。而在说话人确认系统中，假设的说话人模型一般须采用自适应方法来生成。常用的自适应方法主要有最大似然线性回归(Maximum Likelihood Linear Regression, MLLR)和最大后验概率(Maximum A Posteriori, MAP)。笔者将对使用MAP和不使用MAP的不同识别率进行测试、对比。

3.5 实时说话人辨认系统的实现

说话人辨认是指判断1段测试语音是属于1个封闭说话人集合中的哪一个说话人。即给定1个说话人集合$\{1,2,\cdots,S\}$，对应的GMM模型为$\{\lambda_1,\lambda_2,\lambda_S\}$，对1段测试语音$X$，说话人辨认要找到具有最大后验概率的$\lambda_k$，即

$$\hat{S}=\arg\left[\max_{1\le k\le S}\Pr(\lambda_k|X)\right]=\arg\left[\max_{1\le k\le S}\frac{p(X|\lambda_k)\Pr(\lambda_k)}{p(X)}\right] \tag{4}$$

其中，$\Pr(\lambda_k)$为先验概率。假设所有测试者有相同的先验概率，即$\Pr(\lambda_k)=1/S$；另外，对于每个说话人模型，$p(X)$都是相同的，那么，式(4)可简化为

$$\hat{S}=\arg\left[\max_{1\le k\le S}\Pr(X|\lambda_k)\right] \tag{5}$$

进一步假设测试语音X采样后的x是相互独立的，且为了防止概率值过小，对式(5)取对数，得到

$$\hat{S}=\arg\left[\max_{1\le k\le S}\sum_{i=1}^{T}\lg p(x|\lambda_k)\right] \tag{6}$$

其中，$p(x|\lambda_k)$用式(1)进行计算。

在系统中，当传声器读入测试语音后，直接对其进行预处理和特征提取，并计算每个说话人模型的对应分数，分数最大的模型即为识别结果。如果每个人的分数都低于某个阈值，则认为该测试语音不属于已有说话人集合中的任何人。

3.6 基于GMM-UBM的实时说话人确认系统

说话人确认是指事先假设一个待确认的说话人集合，通常该集合只包含某个特定说话人S，对于1段测

试语音 X,判断其是不是这个特定的说话人 S 所说的。首先定义 2 个符号:H_0, 这段测试语音来自假设的说话人 S;H_1, 这段测试语音不是来自假设的说话人 S。然后,给定 1 个确认的阈值 η,定义 1 个对数似然分数比为

$$S(X)=\lg[p(H_0)/p(H_1)] \quad (7)$$

当 $S(X)\geq\eta$ 时,H_0为真,表示认同测试语音属于假设说话人 S;当 $S(X)<\eta$ 时,H_1为真,表示不认同测试语音属于假设说话人 S。

从式(7)可以看出,实现说话人确认的关键是计算 2 个概率 $p(H_0)$和 $p(H_1)$。对于 1 段测试语音 X,将假设的说话人模型记为 λ_{hyp}, 将假设说话人以外的所有人所形成的模型记为 $\lambda_{\overline{hyp}}$。那么,式(7)可写为

$$S(X)=\lg p(H_0)-\lg p(H_1)=\lg p(X|\lambda_{hyp})-\lg p(X|\lambda_{\overline{hyp}}) \quad (8)$$

其中,$\lambda_{\overline{hyp}}$理论上需要假设说话人之外的所有说话人语音来训练,以形成 1 个冒充者模型,但是在现实中这是无法实现的。常用的实际方法是采用一定数量的其他说话人语音作为训练集,以形成一个冒充者模型。λ_{hyp}只要用假设说话人 S 的训练语音进行训练就能得到,但更好的方法是采用模型自适应方法, 系统中分别测试了利用 MAP 方法由 UBM 自适应地训练出的 λ_{hyp}, 以及利用 ML 方法训练出的 λ_{hyp}。

4 系统性能测试与分析

4.1 说话人辨认系统

训练语音采用 24 维 MFCC 系数作为特征参数,训练语音时长 24 s。为实现说话人辨认系统,在实验室环境中录制了 30 个人的语音作为候选说话人语音,并且选择其中 15 人的语音进行测试, 测试语音长度为 4 s,识别正确率结果如表 1 所示。由表 1 可看出, 由于训练语音时间较短,每个人只有 24 s,所以增加高斯混合密度函数的个数并没有带来性能上的改善, 而采样率的提高使获得的语音信息更多,因而识别率有一定提高。

表 1 说话人辨认系统正确率

采样率(8 bit量化)/kHz	正确率/%	
	高斯混合密度个数	
	16	24
8	80	80
16	93	93

4.2 说话人确认系统

说话人确认系统需要预先训练一个 UBM,然后再调整阈值来获得较好的识别效果。在确认过程中,真实说话人的确认错误率称为错误拒绝率(FRR)或遗漏率(miss probability),而冒充者的错误确认率称为错误接受率(FAR)或错误报警率(false alarm probability),这 2 种类型的错误率可以通过阈值来调整。由于是实时系统,系统需要现场录音测试,所以只选择了较少的人数(15 人)来进行测试。在实验中,分别测试了采用 ML 方法进行训练和采用 MAP 方法进行训练的假设说话人模型,对确认结果进行了对比,对比结果如表 2 所示。

表 2 说话人确认系统

错误率	正确率/%	
	ML	MAD
FRR	6.7	6.7
FAR	20.0	6.7

表 2 表明,模型自适应方法由于利用了 UBM 的信息,所以提高了识别正确率。但是由于所使用的 UBM 数据还不够充分,测试人数也较少,所以系统整体性能还有待进一步提高。另外系统中的测试语音和训练语音都来自同一个传声器,所以基本上没有信道差异。为了使软件系统更为通用, 后续的研究还须考虑信道差异的影响,并考虑使用信道差异消除技术。文中系统能实现语音数据的实时处理及识别操作的实时完成, 因此具有很好的实时性,这为系统的实际应用奠定了基础。

参考文献

[1] REYNOLDS D A, ROSE R C. Robust text-independent speaker identification using Gaussian mixture speaker models[J]. IEEE Trans. Speech Audio Processing, 1995, 3(1): 72-83.

[2] REYNOLDS D A. Speaker identification and verification using Gaussian mixture speaker models[J]. Speech Communication, 1995, 17: 92-108.

[3] REYNOLDS D A, QUATIERI T F, DUNN R B. Speaker verification using adapted Gaussian mixture models[J]. Digital Signal Processing, 2000, 10: 19-41.

[4] BAHL R, BROWN P F, DE SOUZA P V, etc. Maximum mutual information estimation of hidden Markov model parameters for speech recognition[C]// Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing. [S.l.]: IEEE Press, 1986, 1: 49-52.

[5] HONG Q Y, KWONG S. Discriminative training for speaker identification based on maximum model distance algorithm [C]// Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing. [S.l.]: IEEE Press, 2004, 1: 25-28.

[6] WU Chou. Discriminant-function-based minimum recognition error rate pattern-recognition approach to speech recognition [C]// Proceedings of the IEEE. [S.l.]: IEEE Press, 2000, 88: 1 201-1 223.

[责任编辑] 潘浩然 [收稿日期] 2007-02-13