基于 VQ 和 GMM 的实时声纹识别研究①

鲁晓倩, 关胜晓

(中国科学技术大学 信息科学技术学院, 合肥 230027)

摘 要: 目前声纹识别系统已经实现较高的识别精度, 但是随着目标说话人个数的增加, 一般系统很难满足实时性的要求, 由此提出一种双层识别模型. 在第一层识别模型中, 采用基于 VQ-VPT(Vector Quantization-Vantage Point Tree)模型进行快速匹配, 挑选出与测试者声纹特征最相近的 K 个目标说话人声纹模型. 在第二层识别模型中, 采用 GMM-UBM(Gaussian Mixture Model-Universal Background Model)模型, 精确匹配上层模型得到的 K 个目标说话人声纹模型, 并做出最终的判决. 实验验证, 双层识别模型在确保高识别精度的前提下, 大幅度的提高了系统的识别速度.

关键词: 声纹识别; 矢量量化; 优势节点树; 高斯混合模型; 通用背景模型

Real-Time Voiceprint Recognition Based on VQ and GMM

LU Xiao-Qian, GUAN Sheng-Xiao

(School of Information Science and Technology, University of Science and Technology of China, Hefei 230027, China)

Abstract: At present, the voiceprint recognition system has achieved high identification precision. But with the increase of the number of target speakers, general system has difficulty in satisfying the need of real time. Therefore, a two-layer recognition model is raised in this paper. The first layer based on VQ -VPT model quickly sorts out K target speakers' voiceprint models which are most similar to the speaker's voiceprint characteristics. In the second layer, the GMM-UBM model matches the K voiceprint models to make a final judgment. Via experimental verification, under the premise of ensuring high recognition accuracy, the two-layer recognition model has greatly improved the recognition speed of the system.

Key words: voiceprint recognition; vector quantization (VQ); vantage point tree (VPT); gaussian mixture model (GMM); universal background model (UBM)

声纹识别[1](Voiceprint Recognition)技术属于生物认证技术的一种, 是通过人的说话声音来辨识说话人身份的技术. 与指纹识别、文字密码、人脸识别等其它认证技术相比, 声纹识别具有不会遗失、无需要记忆、实现简单等特点, 是一种非接触识别方式.

1962 年 L.G. Kesta 首次介绍了采用声纹进行识别的可行性. Bell 实验室最先采用了模板匹配的方式进行声纹识别. 1969 年 Luck JE 提出了采用倒谱的方式进行识别, 其实验结果较为理想. BS Atal 采用线性预测倒谱的方式进行说话人识别. 1972 年 Atal 提出了采用基频轮廓的方式进行声纹识别, 该种方式将数字信号处理相关的技术应用到声纹识别中, 可以从语音信号中提取出间接反映说话人特征的一些参数.

20 世纪 70 年代末到 90 年代, 声纹识别的重点研究领域为模式识别算法, 动态时间规整、矢量量化、隐马尔科夫模型、人工神经网络等逐渐得到广泛应用. 90 年代以后, 高斯混合模型因其简单有效且具有较好的噪声鲁棒性成为声纹识别的主流技术, 将声纹识别引入新的发展阶段. 2000 年以后, Reynolds 提出高斯混合模型-通用背景模型, 降低了说话人模型对训练集合

① 通讯作者:关胜晓 Email:guanxiao@ustu.edu.cn

收稿时间:2014-01-03;收到修改稿时间:2014-03-03

依赖, 增强了系统鲁棒性. 随后机器学习中的区分式分类器支持向量机开始被应用到声纹识别中来, 其中以 2006 年提出的高斯混合模型均值超向量-支持向量机最具影响力. 近年来, 文本无关的声纹识别任务中信道环境失配问题成为研究重点, 一系列基于因子分析的信道补偿技术被应用到文本无关的声纹识别系统中, 有效的改善了信道失配带来的问题.

目前国内的一些高校和高科技单位也在声纹识别领域进行了广泛深入的研究, 中科利信技术公司推出了说话人识别(TSIE)引擎, 科大讯飞公司也推出了声纹识别引擎.

GMM-UBM 虽然识别精度高, 但是由于模型比较复杂, 计算量大, 在说话人数目增多时识别速度比较慢, 难以满足实时性的要求[2]. 针对上述问题, 本文提出了一种双层决策识别模型. 第一层采用 VQ-VPT 模型进行快速匹配. 与红黑树和 B+树相比, VP 树可以用于索引高维特征向量, 用 VP 树索引语音特征向量的时间复杂度为对数时间复杂度, 而传统的 VQ 全搜索的时间复杂度为线性时间复杂度, 因此采用 VQ-VPT 可以快速挑选出与测试者声纹特征最近的 K 个目标说话人声纹模型. 第二层采用 GMM-UBM 模型, 精确匹配第一层模型得到的 K 个目标说话人, 并做出最终的判决.

1 模型研究基础

1.1 VPT 的建立

优势节点树[3-5](Vantage Point Tree, VPT)是一种基于距离的度量空间上的索引结构, 只能采用静态的方式进行创建, 不能进行动态的插入和删除元素. 同时 VPT 是一种基于距离的二叉平衡树[4], 其搜索时间复杂度为 $O(\log n)$, 而传统的 VQ 全搜索时间复杂度为 $O(n)$, 因此采用 VPT 结构可以大大提高搜索效率.

VPT 通过超球体将多维点空间划分成两个互斥的区域, 超球体的几何中心点被称为优势点(Vantage Point), 并同半径一起存放在 VPT 的非叶子节点中[3]. 落入球体内部的点归在优势点的左子树上, 外部的点归在右子树上. 再在互斥区域中分别找出优势点, 继续划分子空间, 生成子节点. 经过层层划分, 每个叶子节点代表划分后的子区域, 并存放属于该区域的点. 在如图 1 所示的一维点空间中, 为数轴 1-12 点建立 VPT, 如图 2 所示.

在图 2 中, 以优势点 1 为例, 距离 1 的小于半径 5 的点归为左子树, 否则归为右子树. 在高维向量空间中, 构建 N 个数据点集的 VPT 的算法如下:

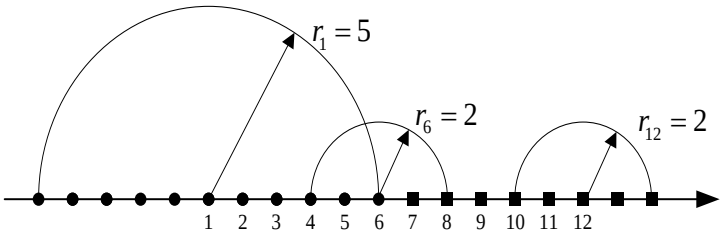

图 1 一维点空间

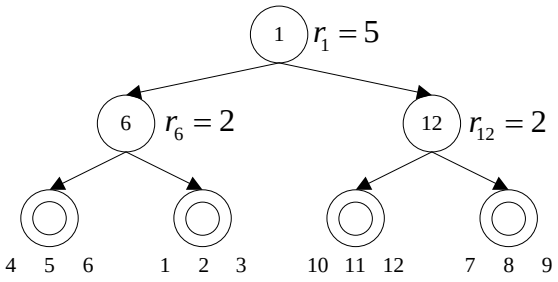

图 2 一维点的 VPT

①在 N 个点中, 选取优势点, 计算剩余点到该点之间的距离并按照距离的大小进行排序.

②选取距离的中值 r，将小于 r 的一半数据点放在左子树，将大于该值的一半数据点放在右子树，并将 r 的值作为半径信息保存在根节点之中.

③针对左子树以及右子树，采用如步骤①和步骤②的方式递归的建立子树，直到吊桶的大小满足阈值要求或距离半径 r 满足阈值要求.

采用该种方式构建 VPT 的时间复杂度为 $O(n\log n)$，空间复杂度为 $O(\log n)$.

对于给定的数据点集合 S，在其上建立的 VPT，将以 $O(\log n)$ 的搜索效率搜到查询点 q 所在的叶子节点. 因此，可以快速的定位查询点 q 与周围点的相对位置.

1.2 GMM

目前，GMM[6,7]是声纹识别的主流模型，得到广泛的应用. 若干高斯函数的线性组合可以逼近任意曲线如图 3 所示，故 GMM 作为一种概率统计模型可以精确的描绘说话人特征参数的概率分布. VQ 模型属于几何模型，用来描述特征空间几何分布特性的，描述精度不及 GMM[8]. 对于一个混合度为 M 的 D 维模型参数为 λ 的 GMM，特征矢量为 X，则 X 在该 GMM 下的似然度为:

$$p(X\mid\lambda)=\sum_{i=1}^{M}\omega_i N_i(X) \tag{1}$$

其中，ω_i 是混合权值，且满足 $\sum\omega_i=1$；$N_i(X)$ 表示第 i 个混合分量的 D 维高斯密度函数，具体形式为:

$$N_i(X)=\frac{1}{(2\pi)^{D/2}\left|\Sigma_i\right|^{1/2}}\exp\left\{-\frac{1}{2}(X-\mu_i)^T\Sigma_i^{-1}(X-\mu_i)\right\} \tag{2}$$

其中，μ_i 表示均值向量，Σ_i 表示协方差矩阵. 为了方便计算会将协方差矩阵对角化，通常认为语音特征参数各维度之间是独立，即协方差矩阵中的对角元素为各维度的方差. GMM 的参数包含混合权值、均值向量、协方差矩阵，即，$\lambda=\{\omega_i,\mu_i,\Sigma_i\}$，$i=1,2,...,M$. 参数 λ 采用最大期望(Expectation Maximization, EM)算法[9]通过迭代的方式完成估计.

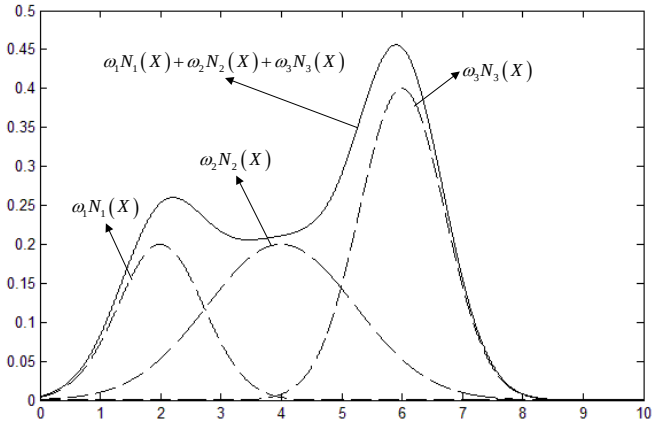

图 3　高斯混合模型逼近任意曲线

2　双层识别模型

双层识别模型将决策分为两步进行，首先对待识别的特征向量进行一次快速匹配，挑选出与其最接近的 K 个声纹模型，淘汰掉不可能的声纹模型，减小第二步的计算量. 然后将得到的 K 个声纹模型进行精确匹配，得出最终结果. 在第一层模型中采用基于 VQ 的方式将目标说话人的声纹模型构建为码书的形式，并采用 VPT 的形式将码书中的码字索引为平衡二叉树的结构，快速识别的过程类似矢量量化，但决策依据并不是量化误差，而是查询与测试特征向量最近的若干个码字中哪一个码书的命中率最高，以此挑选出 K 个声纹模型. 然后利用 GMM-UBM 计算 K 个声纹模型的似然度，选择似然度最大的声纹模型作为识别结果. 示意图如下图所示:

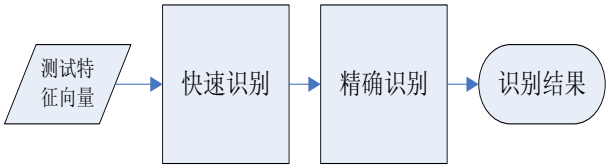

图 4　双层识别模型示意图

2.1 基于 VQ-VPT 的快速识别模型

VQ[10]实质上是对多维语音特征向量在特征空间中聚类的过程，降低冗余信息，识别时的计算量较小. 假设特征向量为 D 维，N 帧语音信号组成特征向量集为 $X=\{X_1,X_2,...,X_N\}$. 那么对该特征集进行量化，先将 D 维欧氏空间按照量化的级数进行划分，假设划分 M 个区域，分别表示为 $R_1,R_2,...,R_M$，每一个被划分点区域 R_j 被称之为一个胞腔. 输入矢量 X_i 与每个胞腔 R_j 的边界进行比较，并将 X_i 归到与之距离最小的胞腔中. 然后在每个胞腔 R_j 内找出距离边界最小的矢量 X_i 作为代表矢量 Y_j，则 M 个代表矢量组成的集合 $Y=\{Y_1,Y_2,...,Y_M\}$ 就构成了一个量化器. Y 称作码书，即该说话人的声纹模板. Y_j ($j=1,2,...,M$)称为码字，M 称为码书长度.

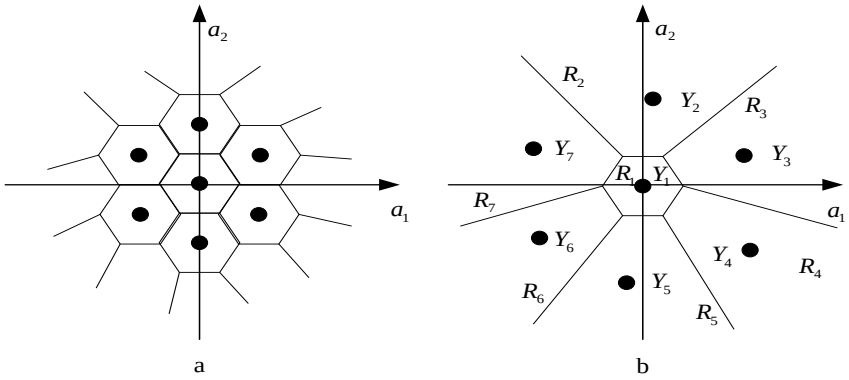

图 5 矢量量化原理示意图

测试者的特征向量集 $T=\{T_1,T_2,...,T_N\}$ 与模板 Y 之间的失真测度由公式(3)定义:

$$D=\frac{1}{T}\sum_{j=1}^{M}\sum_{i=1}^{N}min\,d\left(X_i,Y_j\right) \tag{3}$$

在训练语音受到污染或者训练数据不充分的情况下，某些胞腔里包含极少数特征点. 此类胞腔必然会影响到失真测度的计算影响精度. 此外，对空胞腔计算也会增加计算负担. 考虑精度和计算速度，对胞腔容量进行排序，选取排在前 90%的胞腔作为合理胞腔.

快速识别模型的训练过程如下:

①对目标说话人训练语音进行特征提取，获得特征向量集.

②对特征向量集进行训练，生成代表每个目标说话人声纹特征的码书.

③采用 VPT 的构建算法，对所有目标说话人码书中的码字进行索引，生成 VPT.

快速识别模型的识别过程如下:

①初始化 $Scores\lceil i\rceil=0$，$i=1,2,...,n$

②对测试语音进行特征提取，获得特征向量集.

③特征向量集中选取一个特征向量，在 VPT 中查找与其距离最近的 M 个码字.

④对 M 个码字分别查找其对应的码书，并对其所对应的目标说话人进行加分，$Scores\lceil i\rceil=Scores\lceil i\rceil+1$.

⑤重复③至④直到遍历完测试特征向量集中的所有码字.

⑥在 $Scores\lceil i\rceil$ 中挑选出得分最高的 K 个目标说话人用于精确识别.

传统识别算法需要采用全搜索的方式计算测试特征向量与全体码书的每一个码字之间的距离，系统的计算复杂度正比于码书的个数以及每个码书中码字的个数的乘积，即为 $O(NL)$. 其中 N 为目标说话人的个数，L 为矢量量化的级数. 而采用建立 VPT 索引的方式，系统的计算复杂度对数正比于码书的个数与每个码书中码字的个数的乘积，即为 $O(log\,NL)$，因而采用该种方式可以大大的提高系统的识别速度.

2.2 精确识别模型

精确识别采用 GMM-UBM，即高斯混合模型-通用背景模型. UBM 由所有目标说话人的语音训练得到，代表所有说话人的声纹特征，实质上是一个大型的 GMM. 一般而言，UBM 的训练语句时长为 1 个小时，混合度为 1024. 通常目标说话人的训练数据是有限的，训练得到的 GMM 不能真实反映说话人特征. 为此，有人提出基于 MAP(Maximum A Posteriori)自适应

UBM 得到目标说话人的 GMM, 用来弥补训练数据的不足[6,7].

由于训练大型 GMM 对硬件环境要求较高, 一般采用小型的 UBM, 即训练语句时长大约 30 分钟, 混合度为 128, 但其对统计空间的描述不如大型 UBM 细致. 在这种情况下, 采用 MAP 训练 GMM 模型不能充分的描述目标说话人的特性, 故本文采用 GMM-UBM 作为精确识别模型. GMM 和 UBM 两个模型的训练是分别独立进行的. 由于训练 UBM 的数据集比训练目标说话人 GMM 的数据集大的多, 所以其混合度比 GMM 也要大的多.

精确识别阶段对在快速识别阶段筛选出的 K 个最可能的说话人模型采用 GMM -UBM 进行精确识别. GMM-UBM 的系统框图如图 6 所示:

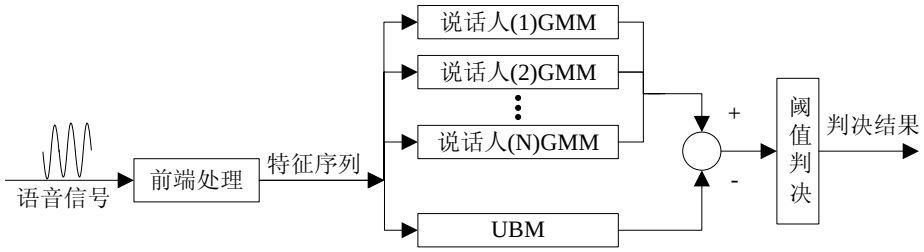

图 6　GMM-UBM 系统框图

由上图 6 可知, 测试语音的特征矢量序列 $X=\{X_t\},t=1,2,...,T$ 的对数似然比可以由式(4)来计算:

$$S(X)=\frac{1}{T}\sum_{t=1}^{T}\left(log\left(p\left(X_t\mid\lambda_s\right)\right)-log\left(p\left(X_t\mid\lambda_{UBM}\right)\right)\right) \tag{4}$$

其中, λ_s 是目标说话人的 GMM 模型参数, λ_{UBM} 是 UBM 的模型参数. 采用似然比的方式打分是一种归一化处理, 可以对不同的目标说话人设置统一的判决阈值. 在精确识别时, 分别计算测试特征向量与快速识别阶段筛选出的 K 个目标识别模型之间的相似度, 并选取最大似然度值所对应的目标说话人模型作为识别结果[11]. 由于不需要计算测试特征向量与所有 N 个说话人模型的似然度, 减小了计算量, 提高了识别速度.

3　实验结果及分析

本文实验数据来源于 TIMIT 语音数据库, 该语音库包含来自美国的 8 个地区的 630 个人的语音, 每人录有 10 句话, 共 6300 个句子, 每句时长大约 3s. 对每个目标说话人取前 5 句录音训练其 GMM, 后 5 句用作测试. 任意选取每个人的 1 句录音组成约 30s 的录音用于训练 UBM. 本文采用 30ms 汉明窗进行加窗, 帧移 16ms, 由双门限法去除静音段, 再用预加重滤波器 $H(z)=1-0.95z^{-1}$ 提升高频成分. 选取 11 阶 MFCC 作为特征参数[12].

3.1 快速识别算法的时间性能

选取某测试者的一段语音, 提取特征得到特征矢量序列 $X=\{X_t\},t=1,2,...,T$, 分别采用传统的 VQ 算法和 VQ-VPT 的算法提取 10 个最近的目标说话人, 实验结果如表 1 所示.

表 1　VQ 与 VQ-VPT 的搜索相似模板性能比较

码字个数	6800	11000	19600	22000	29000	31000	36000	43000
VQ (s)	0.061	0.094	0.146	0.167	0.208	0.241	0.256	0.307
VQ-VPT (s)	0.001	0.0019	0.0029	0.0030	0.0039	0.0029	0.0041	0.0049

VPT 与 VQ 算法的时间对比如图 8 所示. 从该图可以看出, 采用 VPT 的方式进行搜索, 其时间消耗远小于采用 VQ 的全搜索方式. VQ 算法的时间消耗与码字数目之间是线性的关系, 而 VPT 搜索的时间消耗与码字数目之间是对数的关系, 因而随着码字数目的增加, 两者之间的性能差异会更大, VPT 的性能优势会变得更加明显. 当系统中有 150 个目标说话人, 矢量量化的码书中码字个数为 256 个时, VPT 树搜索的速度为 VQ 全搜索速度的 60 倍以上, 因而采用 VPT 索引 VQ 码字的方式进行快速识别具有较好的时间性能.

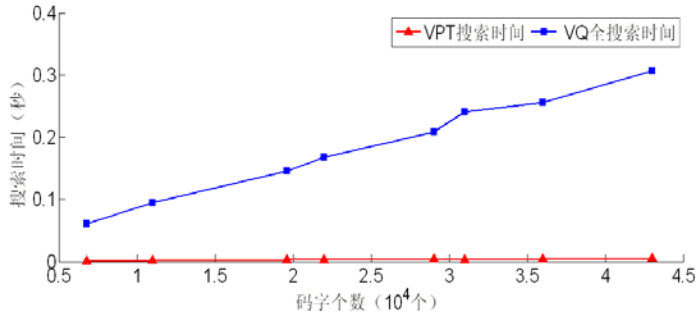

图 7 VQ 与 VQ-VPT 算法的时间对比

3.2 模型筛选准确率

图 8 和表 2 给出了对三段来自不同测试者的语音的快速识别结果. 挑选 10 个最近的目标说话人, 每个目标说话人码书大小为 256, 选取 30 个目标说话人的模板. 第一段测试语音的为 1 号说话人的录音信息, 第二段测试语音为 5 号说话人的录音信息, 第三段测试语音为 11 号说话人的语音信息. 从最终的得分结果来看, 真实目标说话人的得分通常明显高于其它目标说话人的得分. 经大量实验, 在 VQ-VPT 算法中只要从所有目标说话人模型中挑选出 10%的目标说话人模型即能包含真实目标说话人模型.

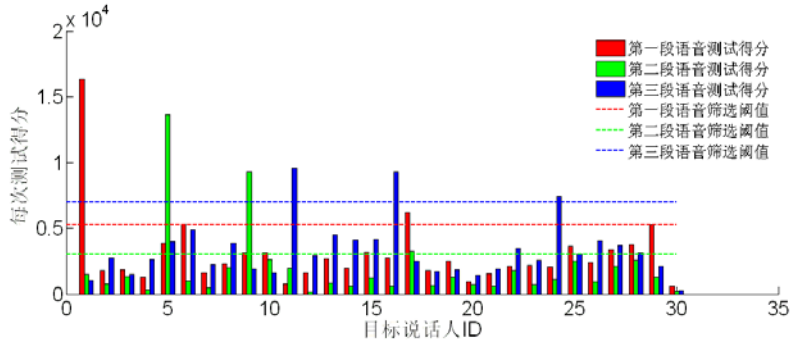

图 8 三个不同测试者的快速识别结果

表 2 得分前三名的目标说话人分数所占百分比

前三名得分所占百分比	1st	2nd	3rd
第一段测试语音	17.8068%	6.7178%	5.7409%
第二段测试语音	23.5307%	16.0929%	5.5611%
第三段测试语音	9.5677%	9.3205%	7.4400%

为了测试快速识别算法的核心模型筛选准确率, 我们测试了当系统中存在不同数目目标说话人模型时, 取前 15%的目标说话人模型数目时该算法的正确筛选率, 如表 3 所示:

表 3 快速识别算法的正确筛选率

目标说话人个数	15	50	100
正确筛选率	100%	97%	95%

3.3 双层决策系统的性能

表 4 给出了采用双层决策模型的方式和采用 GMM-UBM 的方式进行识别的速度以及识别的准确率的对比结果. 其中, GMM 混合度设置为 32, UBM 的混合度设为 128, 采用 VQ 算法生成的码书中包含 256 个码字. 在快速识别中挑选得分前 10%的目标说话人.

表 4 GMM 模型与双层决策模型的对比

人数	20	50	120
GMM-UBM 识别率	95.2%	92.3%	87.2%
双层决策模型识别率	94.7%	91.5%	85.4%
GMM-UBM 识别时间(s)	0.884	2.42	7.25
双层决策模型识别时间(s)	0.105	0.352	0.872

从上述结果可以看出, 采用双层决策模型可以在不损失识别率的前提下大幅的提高识别速度. 由于采用快速算法的识别时间很大程上依赖 K 值的选择. 选取较小的 K 值, 可以在快速识别阶段过滤掉更多的目

标说话人, 获得更高的识别速度. 当 K 值选取为目标说话人数的 10%到 20%时, 通常可以在损失 2.5%以下识别精度的同时将算法的识别速度提高 8 倍左右.

4 结语

本文主要针对声纹识别中常用到的两种模型, 根据各自模型的优点, 提出了双层识别模型. 在确保识别准确率的情况下, 进一步提高识别速度. 该模型将声纹识别系统的决策划分为两步, 首先采用快速识别算法进行粗识别, 挑选出最有可能是识别结果的 K 个目标说话人, 然后对这 K 个目标说话人采用精确识别算法进行进一步的识别.

本文结合 VPT 的思想, 提出 VQ-VPT 模型. 该算法采用 VPT 的形式对码字建立索引, 通过搜索与测试特征向量最近邻的码字的方式对目标说话人进行评分, 然后选出得分最高的 K 个模型. 当系统中目标说话人数为 150 人时, 该算法的执行速度为 VQ 全搜索算法执行速度的 60 倍以上.

经实验验证, 双层识别模型与 GMM-UBM 算法相比, 可以在使识别率降低不超过 2.5%的情况下, 使算法的识别速度提高 8 倍左右.

参考文献

1 朱浩冰,郭东辉.声纹识别系统原理及其关键技术.计算机安全,2007:15–17.

2 杨迪,戚银城,刘明军,张华芳子,武军娜.说话人识别综述.电子科技,2012,25(6):162–165.

3 Kibriya AM. Fast algorithms for nearest neighbour search[Master Thesis]. Hamilton: The University of Waikato, 2007.

4 Yianilos PN. Data structures and algorithms for nearest neighbor search in general metric spaces. Proc. of the Fourth Annual ACM-SIAM Symposium on Discrete Algorithms. New York. ACM Press. 1993. 311–321.

5 Liu T, Moore AW, GRAY A, Yang K. An investigation of practical approximate nearest neighbor algorithms. Advances in Neural Information Processing Systems. 2004. 825–832.

6 Reynolds DA, Rose RC. Robust text-independent speaker identification using Gaussian mixture speaker models. IEEE Trans. on Speech and Audio Processing. 1995. 72–83.

7 Reynolds DA, Quatieri TF, Dunn RB. Speaker verification using adapted Gaussian mixture models. Digital Signal Processing. 2000. 19–41.

8 许东星.基于 GMM 和高层信息特征的文本无关说话人识别研究[硕士学位论文].合肥:中国科学技术大学,2009.

9 Moon TK. The expectation-maximization algorithm .Signal Processing Magazine, 1996, 13(6): 47–60.

10 Gersho A, Gray RM. Vector Quantization and Signal Compression. MA: Springer, 1992: 37–90.

11 侯钰,刘轶,郑方,蒋丹宁,秦勇,黄石磊,刘勇.基于VP树结构的多层匹配算法在哼唱识别中的应用.清华大学学报(自然科学版),2009:1419–1424.

12 韩纪庆,张磊,郑铁然.语音信号处理.北京:清华大学出版社, 2010.