基于 VQ 和 HMM 的双层声纹识别算法

赵 峰[1], 于 洋[2]

(1. 桂林电子科技大学 计算机与信息安全学院, 广西 桂林 541004;
2. 桂林电子科技大学 信息与通信学院, 广西 桂林 541004)

摘 要: 为了给盲人设计一种可靠度高、功耗低、方便快捷的嵌入式说话人身份识别系统, 提出一种基于 VQ 矢量量化和 HMM 隐马尔科夫模型的双层声纹识别算法。该算法采用复杂度相对低的 VQ 矢量量化算法, 对模板库进行快速初次筛选, 得出有效匹配结论或选出可能符合的模板, 缩小模板库范围, 并采用离散 HMM 隐马尔科夫模型进行精确识别。仿真结果表明, 该算法在保证识别精度的同时, 缩短了识别时间, 提高了识别效率。

关键词: 声纹识别; 矢量量化; 隐马尔可夫模型; 双层

中图分类号: TN912.34 **文献标志码**: A **文章编号**: 1673-808X(2017)01-0008-07

Two-level voiceprint recognition algorithm based on VQ and HMM

ZHAO Feng[1], YU Yang[2]

(1. School of Computer Science and Information Security, Guilin University of Electronic Technology, Guilin 541004, China;
2. School of Information and Communication, Guilin University of Electronic Technology, Guilin 541004, China)

Abstract: In order to design a reliable, low-power and efficient speaker recognition embedded system which is used to help the blind, the two-level speaker recognition model based on VQ and HMM is put forward. The VQ algorithm, which has low complexity, is adapted to filter the library quickly. The no matching result can be made or the possible templates can be selected after this. Then the HMM is employed to make an accurate conclusion. The simulation result shows that the algorithm can reduce the time of recognition with good recognition rate, which improves the recognition efficiently.

Key words: speaker recognition; vector quantization; hidden Markov model; two-level

盲人用户受自身条件所限, 失去了视觉这一人类感知的重要信息来源, 因此只能“听音辨人”。相关研究表明, 人类听觉记忆相对较弱, 衰退较快, 所以如何可靠地判断对话人的身份, 是盲人生活中要解决的一个重要问题。而目前市场上的盲人辅助用具主要是基于导航和阅读的应用。本研究拟为盲人用的身份识别嵌入式设备设计一种算法, 使该设备可以以较高的可信度、较快捷的速度和较低的功耗来完成身份识别任务。

声纹特征相比较于其他生物特征, 具有提取方便、设备简单、算法复杂度低以及可执行远程操作的优点, 已经在信息检索、身份认证、安保刑侦、个性化定制等领域得到了广泛的应用。考虑以上技术特点, 本系统选择声纹作为生物识别元素。但现今的声纹识别算法往往是在 PC 机上运行, 通常将识别的准确率作为主要的性能参数, 而对识别耗时、系统资源的占用率等性能考虑较少[1]。鉴于此, 提出了基于矢量量化(vector quantization, 简称 VQ)和隐马尔可夫模型(hidden Markov model 简称 HMM)的双层声波识别算法。该算法可在模板库无匹配的情况下, 快速给出提示, 节约系统资源。在模板库有匹配的情况下, 可以先进行筛选, 再进一步给出精确结果, 减少精确匹配的计算量, 从而在一定程度上提高了识别的效率, 符合对该嵌入式盲人辅助设备可信度高、快捷、低功耗的设计要求。

1 声纹识别系统概述

声纹识别系统根据识别对象的不同, 可以分为与

收稿日期: 2016-01-15
基金项目: 国家自然科学基金(61471135)
通信作者: 赵峰(1974—), 男, 山东日照人, 研究员, 博士, 研究方向为无线通信理论及信息处理技术。E-mail: zhaofeng@guet.edu.cn
引文格式: 赵峰, 于洋. 基于 VQ 和 HMM 的双层声纹识别算法[J]. 桂林电子科技大学学报, 2017, 37(1): 8-14.

文本相关和与文本无关两大类。前者需要说话人在训练和识别时提供相同的文本;后者对此没有特定要求。相对而言,与文本相关的声纹识别便于实现,且识别精度高;与文本无关的声纹识别用户使用方便,应用范围广。

声纹识别的过程主要包括训练和识别[2]。训练即根据语料,经过预处理和特征提取,建立相应的模板。识别是将待识别的语料,经过预处理和特征提取,与之前建立的模板进行匹配。声波识别流程如图 1 所示。

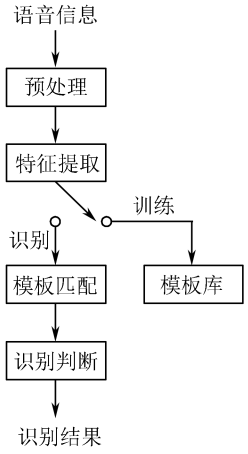

图 1　声纹识别流程图

Fig. 1　The process of speaker recognition

据此整个声纹识别系统可以分解为以下 2 个部分[3]:

1)语音信号的预处理。包括采样量化、预加权、分帧、加窗、端点检测等过程。整个过程根据语音信号的"短时平稳性"进行。

2)语音信号的特征提取。理想的特征参数首先要有较好的识别性,在受到干扰时还要有较好的顽健性,计算不能过于复杂,还要不易被模仿等。但目前尚未有特别符合上述要求的参数。较常用的有线性预测倒谱系数(linear prediction coding coefficient,简称 LPCC)、Mel 频谱倒谱系数(Mel frequency cepstrum coefficient,简称 MFCC)和感知线性预测系数(perceptual linear predictive,简称 PLP)等。

声纹识别的核心算法:

1)模板匹配法。利用特定的特征矢量集合和待测矢量通过距离测度来实现比较,根据累计距离来判断。计算量小,模型简单。但训练数据不足时精度较低。常用的算法有动态时间规整(dynamic time warping,简称 DTW)、矢量量化等。

2)概率模型法。根据特征矢量的统计特性建立数学模型,从概率统计的角度进行匹配识别。模型精确,识别度高,但训练复杂,计算量大。如 HMM、高斯混合模型(Gaussian mixture model,简称 GMM)。

3)新机器学习算法。如近些年研究较为深入的人工神经网络(artificial neural network,简称 ANN)、支持向量机(support vector machine,简称 SVM)等。这种算法可以自我学习,自我完善,将是今后发展的方向[4]。

2　基于 VQ 和 HMM 的双层识别模型

2.1　预处理

2.1.1　预加重

语音信号的平均功率谱受生门激励和口鼻辐射的影响,高频端约在 800 Hz 以上按照 6 dB/oct(2 倍频)跌落,所以要采用具有 6 dB/oct 的高频提升特性的预加重数字滤波器实现,一般为一阶:

$$H(z)=1-uz^{-1}。\tag{1}$$

其中 u 取值为 0.93～0.97。

2.1.2　分帧

一般假设语音信号在 10～30 ms 短时间内是平稳的。每个短时间为一帧,分帧可以采用连续分段,但为了保持语音信号的连续性,避免数据丢失,使过渡平滑,一般采取交叠分段,使相邻的两帧之间有 50%～70%的重叠,重叠部分称为帧移,本算法采用 50%的帧移。

2.1.3　加窗

为了减少截断效应在分帧过程中的影响,降低帧两端的坡度,要对语音帧进行加窗操作。矩形窗的谱平滑性较好,但波形细节易丢失,并且会产生泄露。汉明窗的主瓣宽度是矩形窗的 2 倍,衰减也比矩形窗快,因此采用汉明窗实现加窗操作。窗函数为:

$$w(n)=\begin{cases}0.54-0.46\cos(2\pi n/(N-1)),\\ \qquad 0\leqslant n\leqslant N-1;\\ 0,\ \text{其他}。\end{cases}\tag{2}$$

2.1.4　端点检测

端点检测主要为了检测语音数据中的噪声段和静音段等无效片段,确定语音信号的起点和终点,从而缩短处理时间,减少噪声的干扰,提高识别效率[5]。本算法采用基于特征的双门限检测法,利用语音信号的短时能量和短时过零率这 2 个特征的 4 个门限联合检测。

短时能量是指一帧内样点值的加权平方和,

$$E(n)=\sum_{m=n}^{n+N-1}(x(m)w(n-m))^2,\tag{3}$$

表示从第 n 个点开始加窗的短时能量。

短时过零率为一帧内语音信号穿过时间轴的次数,其定义为:

$$Z(n)=\frac{1}{2}\sum_{m=-\infty}^{\infty}|\mathrm{sgn}(x(n))-\mathrm{sgn}(x(n-1))|\times w(n-m)。\tag{4}$$

其中 sgn(x) 为符号函数，x 为非负时为 1，为负时为 −1。但按定义计算常常会使结果容易受到低频的干扰，所以一般再设定一个门限 T，计算跨过正负门限 (T，$-T$) 的次数，这样若有小的随机噪声，只要在正负门限的带内，就不会产生虚假过零，

$$Z'(n)=\frac{1}{2}\sum_{m=-\infty}^{\infty}(|\mathrm{sgn}(x(m)-T)-\mathrm{sgn}(x(m-1)-T)|+|\mathrm{sgn}(x(m)+T)-\mathrm{sgn}(x(m-1)+T)|)w(n-m)。\tag{5}$$

但这样可能导致在正负门限附近的波动产生多次过零记录，而且当相邻帧同时跨越 2 个门限时，会产生 2 次过零记录，因此采用滑动门限 S 进行改进，即相邻帧必须满足异号，且差值的绝对值大于门限 S，才算有效的过零记录，

$$Z''(n)=\frac{1}{2}\sum_{m=-\infty}^{\infty}(|\mathrm{sgn}(x(m)-T)-\mathrm{sgn}(x(m-1)-T)|\mathrm{sgn}(|x(m)-x(m-1)|-S))w(n-m)。\tag{6}$$

双门限端点检测法，首先要给短时能量和短时过零率设定一个低门限和一个高门限，并将语音帧划分为语音段、非语音段和过渡段 3 种。初始默认为非语音段，当 2 个量中有一个量超过了低门限，则开始标记，此时并不确定是否为语音段，故定义为过渡段；在过渡段中，若 2 个量都回到了低门限以下，则认为过渡段还是处在非语音段，重新标记为非语音段；若在过渡段中有一个量超过了高门限，则认为过渡段实际进入了语音段，正式标记为语音段；在语音段中，当 2 个量都低于低门限时，开始标记，当此状态持续一定时间后，则认为从标记时进入了非语音段；若持续时间不满足，则认为仍然处于语音段；若开始标记后满足了持续时间，即信号确实进入了非语音段，但检测之前这段语音的长度不满足最低要求，则认为该段语音段可能为噪声造成，将其重新标记为非语音段。

一般门限的初始值依据经验给出，如短时过零率常设定为 10 或 5。这里针对短时能量的高低门限，给出一种自适应的门限设置。

首先计算所有帧中能量的最大值 $E_{\max}$ 和最小值 $E_{\min}$，然后按照以下表达式求低门限 E_{l} 和高门限 E_{h}：

$$E_{\mathrm{l}}=E_{\min}(1+1.8\lg\frac{E_{\max}}{E_{\min}}),\tag{7}$$

$$E_{\mathrm{h}}=0.85E_{\mathrm{l}}+0.15E_{\mathrm{s}}。\tag{8}$$

其中 E_{s} 为所有帧中能量大于 E_{l} 的帧的能量平均值。

除了 2 个状态量的 4 个门限外，还要设定非语音段需满足的最短时长和语音段需满足的最短时长。根据经验，设定非语音段的最短时长为 60 ms，语音段的最短时长为 160 ms。

2.2 特征参数的提取

人耳接收到的频率与实际频率不是线性的对应关系，人耳对于低频声音比高频声音更敏感，这一主观感觉的度量就是梅尔频率。其与频率的换算公式为：

$$B(f)=2595\times\lg\left(1+\frac{f}{700}\right)。\tag{9}$$

MFCC 是基于 Mel 频率的倒谱系数，首先要通过 Mel 滤波器组转换为 Mel 频谱，再进行倒谱计算[6]。Mel 滤波器组的中心频率在梅尔频率刻度上是均匀排列的，常用三角带通滤波器实现。MFCC 提取过程如图 2。

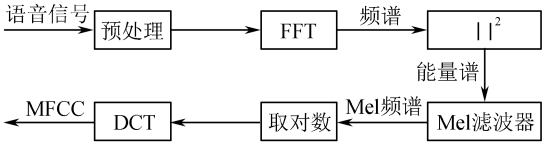

图 2 MFCC 提取过程

Fig. 2 The process of solving MFCC

MFCC 可反映语音信号的静态特性，但人耳对语音信号的动态特性比较敏感，所以再取 MFCC 的一阶差分，来增强特征参数的动态特性。差分公式为：

$$d(n)=\frac{1}{\sqrt{\sum_{i=-k}^{k}i^2}}\sum_{i=-k}^{k}i\cdot c(n+i),\tag{10}$$

其中 k 一般取 2，表示差分是当前帧和前两帧以及后两帧的线性组合。

相关研究表明，并非所有 MFCC 分量都对语音特征提取有较大的参考价值，而前 12 维中包括了绝大部分有价值的信息，因此选取 12 维 MFCC 和其 12 维一阶差分组合，合成 24 维的特征参数。

2.3 双层识别模型

2.3.1 VQ 算法

矢量量化是将 D 维空间划分为 K 个区域边界，每个区域被称为胞腔，然后将输入的特征矢量和每个胞腔的边界进行比较，并量化成“距离”最小的胞腔的中心。这些中心被称为码字，而码字的组合称为码书。

“距离”一般采用欧式距离：

$$d(y_i,C_j)=|y_i-C_j|^2=\sum_{k=1}^{D}(y_{ik}-C_{jk})^2。\tag{11}$$

其中:y_i 为 D 维待测矢量;C_i 为一个码字。一个特征矢量序列的总距离等于各个矢量的距离之和。训练码书是一个迭代的过程,其基本思想是在每次迭代时,利用最小距离准则对待测矢量序列进行重新分类,使整个距离总和不断减小[7]。其步骤如下:

1)设置量化失真阈值 δ、初始距离 $d^{(0)}$、最大迭代次数 $K_{\max}$ 以及初始码字 $C_j^{(0)}(j=1,2,\cdots,M)$。

2)以码字为中心,根据最临近准则划分为 M 个空间 $S_1^{(m)}$, $S_2^{(m)}$,…, $S_M^{(m)}$。若 $d(y_i,C_j^{(m)})\leqslant d(y_i,C_k^{(m)})$,$\forall k\neq j$,则 y_i 归入 $S_j^{(m)}$,且 $d_i=d(y_i,C_j^{(m)})$,其中 m 为迭代次数。

3)计算总距离:

$$d=\sum_{i=1}^{T}d_i,\tag{12}$$

其中 T 为总帧数。计算改进量的相对值:

$$\delta^{(m)}=\frac{\Delta d^{(m)}}{d^{(m)}}=\frac{|d^{(m)}-d^{(m-1)}|}{d^{(m)}}。\tag{13}$$

4)若 $\delta^{(m)}<\delta$ 或者 $m\geqslant K_{\max}$,则输出码字集合,否则转步骤 5)。

5) $m=m+1$,计算新的聚类中心

$$C_j^{(m)}=\frac{1}{N_j}\sum_{y_i\in S_j^{(m)}}y_i,\ j=1,2,\cdots,M。\tag{14}$$

然后转步骤 2)。

初始码字的选择将直接影响最后的结果。在输入矢量序列中随机选取 M 个作为初始码字,但这样选取分布不均匀,效果不稳定。通常采用分裂法选取,传统的分裂法将最初的样本质心作为初始码字,然后通过一个微扰 ε(增减)分裂成 2 个码字,依此聚类,再重复求各类的质心,通过 ε 对每个质心进行分裂,重复进行即可得到需要的码书[8]。但该算法对微扰 ε 的选取要求较高,而且一般 ε 为一定值,不具有较好的自适应性。因此,采用如下方法来确定初始码字。

首先找到最初样本的质心作为初始码字,然后找到畸变最大的矢量,即与质心的距离最远的矢量 C_j,然后再寻找一个与 C_j 误差最大的矢量 C_i,以这 2 个矢量为中心聚类,将矢量集分成 2 个类,再继续分裂 N 次,即可得到 2^N 个类,将每类的质心作为码字构成码书。当出现空类(即只有 1 个元素)时,直接将这个类删除,最后将所含元素最多的类按照上述方法分裂成 2 个类进行替换。

进行识别时,将待测样本与各个码书按照最邻近距离准则进行聚类,求得距离总和,即可根据不同码书的距离进行判断。

2.3.2　HMM 算法

HMM 过程是一个双重随机过程,其一用来描述状态的转移,另一个描述状态和观察值之间的统计关系。而从观察者的角度来看,只能看到观察值,而状态的变化则成了“隐含”的。而这一过程与人类的发声特点相符,可以较好地描述语音信号。

对于有 N 个状态($\theta_1,\theta_2,\cdots,\theta_N$)和 M 个观察值($V_1,V_2,\cdots,V_M$)的 HMM 过程,其主要参数为:

1)状态转移矩阵:

$$\boldsymbol{A}=\{a_{ij}=P(q_{t+1}=\theta_j\mid q_t=\theta_i)\},$$

其中 q_t 为 t 时刻的状态,$1\leqslant i,j\leqslant N$;

2)观察值概率矩阵:

$$\boldsymbol{B}=\{b_{jk}=P(o_t=V_k\mid q_t=\theta_j)\},$$

其中 o_t 为 t 时刻的观察值,$1\leqslant j\leqslant N$,$1\leqslant k\leqslant M$;

3)初始概率矩阵:

$$\boldsymbol{\pi}=\{\pi_i=P(q_1=\theta_i)\},$$

其中,$1\leqslant i\leqslant N$。离散 HMM 与连续 HMM 的主要区别是参数 $\boldsymbol{B}$,前者的 $\boldsymbol{B}$ 是一个概率矩阵,后者的 $\boldsymbol{B}$ 是由所有状态上的观察概率密度函数共同构成[9]。

HMM 算法主要解决识别过程中的匹配问题、解码问题和训练问题。3 个问题的求解如图 3 所示。

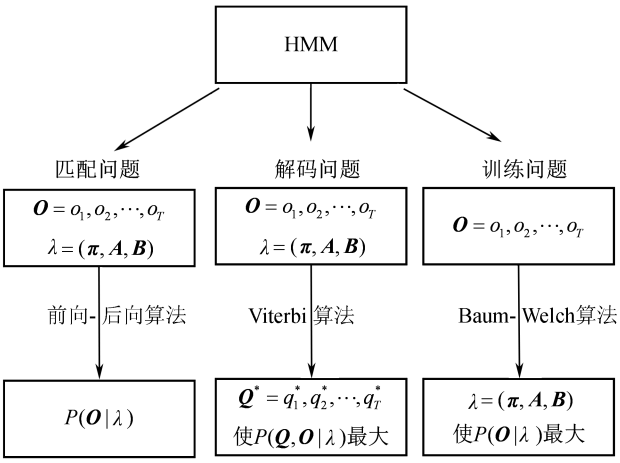

图 3　HMM 三个问题的求解

Fig. 3　The solution of HMM

训练问题中迭代的初值对结果十分重要,特别是 $\boldsymbol{B}$ 的初始值,因此先随机设定 λ,利用 Viterbi 算法求出状态序列,再根据状态序列估计 $\tilde{\lambda}$,将 $\tilde{\lambda}$ 作为重估公式的初始值,利用 Baum-Welch 算法求得参数 $\bar{\lambda}$。

2.3.3　VQ-HMM 双层识别模型

本算法应用于嵌入式盲人辅助声纹识别系统,因此希望算法在保证精度、可靠度的前提下,具有低功耗、低存储、快捷便利的嵌入式设备特点。因此采用

文本相关的声纹识别，识别精度较高，计算量适中，适合嵌入式设备。VQ 矢量量化采用若干离散的数值来表示矢量，有很大的压缩度，减少了数据的存储量，算法复杂度低，便于快速进行识别，但识别精度有时易受影响[10]。为了进一步提高精度，将 VQ 算法作为初级识别，可以快速淘汰差异比较大的样本，筛选出可能的样本，缩小识别的范围。然后采用 HMM 作为后一级识别模型进行精确识别，选择离散 HMM，简化模型复杂度，从而减少了占用计算量最大的匹配计算；同时可以充分利用第一级别的量化算法，直接将特征矢量序列量化为用码字符号表示的符号序列，供离散 HMM 模型使用，节省了系统的开销[11]。考虑到是文本相关的识别，为了保持语音信号的时间特性，采用从左到右无跨越的 HMM 模型，$\boldsymbol{\pi}$ 的初始值可设为 $(1,0,\cdots,0)$，对于四状态的 HMM，$\boldsymbol{A}$ 的初始值可设为：

$$\boldsymbol{A}=\begin{bmatrix} a_{11} & a_{12} & 0 & 0 \\ 0 & a_{22} & a_{23} & 0 \\ 0 & 0 & a_{33} & a_{34} \\ 0 & 0 & 0 & a_{44} \end{bmatrix}。 \tag{15}$$

从左到右的模型即是当 $j<i$ 时，$a_{ij}=0$；无跨越的模型即是当 $j>i+1$ 时，$a_{ij}=0$。$\boldsymbol{B}$ 的初始值可以采用均值设定。

综上，双层声纹识别系统的完整实现过程为：

首先对语音数据进行预处理，求特征向量。训练时，利用分裂法，从样本中求得 VQ 量化初值，并对样本进行 VQ 矢量量化，保存量化后得到的码书 $\boldsymbol{B}_{\text{code}}$、样本在码书下的平均失真距离 d 以及量化后的特征参数序列转换成的码字符号序列 L。L 即为 DHMM 的输入序列，根据上面的初始值设定，利用 Viterbi 算法求出状态序列，进而估计输出参数，调整模型的初始值，再利用 EM 算法求出模型。保存模型的 3 个变量 $\lambda=(\boldsymbol{\pi},\boldsymbol{A},\boldsymbol{B})$。识别时，将待测的序列与系统中保存的码书进行匹配，匹配结果与保存码书对应的失真距离 d 进行比较，若在 $0.9d\sim1.1d$，则通过初级筛选，并利用该码书将特征矢量序列量化为码字符号序列，送入 DHMM 模型。DHMM 模型对每个符合条件的量化后的序列，用其对应的模板进行匹配，相似度最大的即为最后的识别结果。系统流程如图 4 所示。

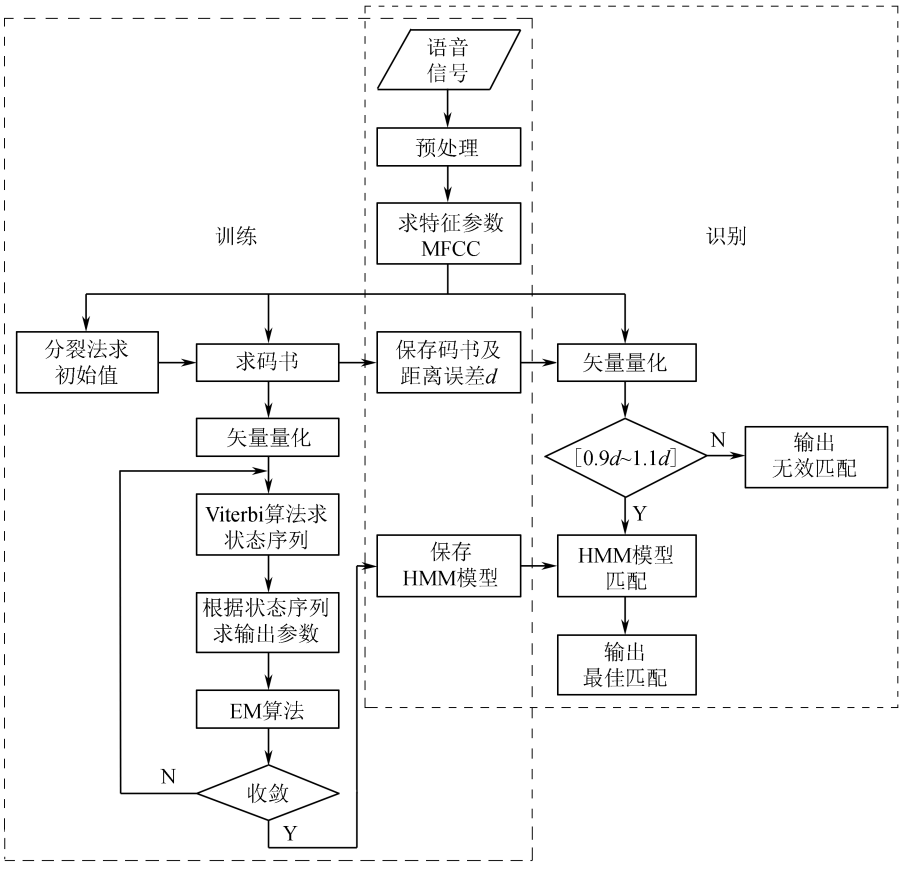

图 4 系统流程

Fig. 4 The process of the system

3　实验结果分析

采用 Matlab 进行仿真实验。实验采用美国语言数据联合会(linguistic data consortium,简称 LDC)发布的连续语音语料库进行仿真实验,该库含有 630 位不同说话者的语音信息,具有较好的性别、年龄和地区组合度,语料组成也多种多样。根据算法采用的文本相关的识别选取其中的 SA 类语料,该语料库中由不同的说话者朗读相同的语句。

3.1　改进端点检测法的实际效果

首先用传统的端点检测法和改进后的端点检测法对一段语音进行检测。检测效果见图 5 和图 6。

3.2　码书长度对系统的影响

码书中码字个数会影响 VQ 量化的速度和精度,同时码字个数也是 DHMM 模型中观察值数目,因此会影响整个系统的识别性能。选取一段文本,对含有 50 个样本的库进行识别实验,用有效样本进行匹配,分别取码字数为 16、32、64、128。采用语音识别中常用的实时率(real time factor,简称 RTF)来度量算法的运行速度。

由图 5、图 6 对比可见,经过改进优化后的自适应端点检测消除了部分噪声的干扰,效果更好。

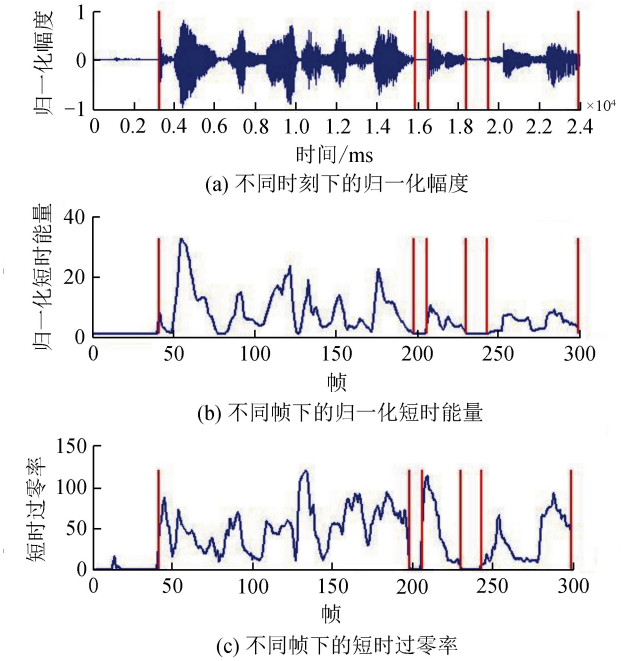

图 5　传统端点检测算法检测效果

Fig. 5　The effect of traditional algorithm for endpoint detection

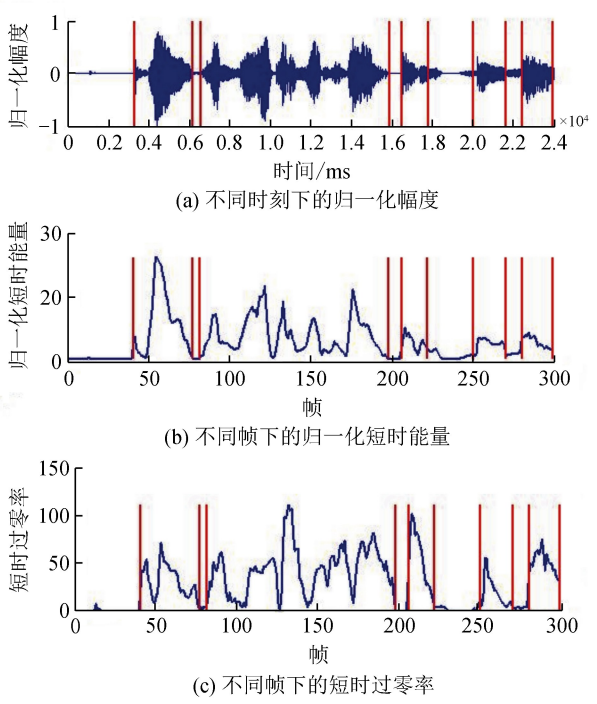

图 6　改进后自适应端点检测算法检测效果

Fig. 6　The effect of adaptive algorithm for endpoint detection

$$R_{TF}=\frac{T_o}{T}, \tag{16}$$

其中:T_o 为算法运行的时间;T 为处理的语音时长。不同码书长度下识别效率如表 1 所示。

表 1　不同码书长度下识别效率

Tab. 1　The recognition efficiency for different length of codebook

码字数	识别率/%	平均 RTF
16	80	0.037
32	86	0.054
64	84	0.087
128	88	0.152

由表 1 可见,并非一味增加码书的长度就可以带来较好的识别效果,码书长度的增加会带来运算时间的变长和存储空间的变大,据此选取码书长度为 32,以下实验均在码书长度为 32 的条件下进行。

3.3　双层识别系统的识别效率

选取 2 段文本,分别选取 20 个样本进行建模,然后用对应的测试样本进行识别。全匹配情况下识别效率如表 2 所示。

表 2 全匹配情况下识别效率

Tab. 2 The recognition efficiency under full-matching

识别算法	识别率/%	平均 RTF
VQ	87.5	0.031
HMM	90.0	0.066
VQ+HMM	90.0	0.036

从表 2 可看出，由于 VQ 算法进行了提前的快速筛选，使双层算法可以最大化地减少 HMM 的计算次数，从而减少识别的时间，同时相对于单纯的 VQ 算法又有精度上的优势。

对于前面选取的 2 段文本，分别选取 20 个样本进行建模，其中 10 个可以进行有效匹配，另外 10 个库中无匹配。非全匹配情况下识别效率如表 3 所示。

表 3 非全匹配情况下识别效率

Tab. 3 The recognition efficiency under unfull-matching

识别算法	识别率/%	平均 RTF
VQ	90.0	0.029
HMM	92.5	0.065
VQ+HMM	90.0	0.032

从表 3 可看出，并非每次均有有效的识别结果，可能库中并无有效的匹配，这样双层模型仅利用第一层 VQ 即可做出判决，无需进行复杂的 HMM 识别过程，大大提高了识别效率，更符合实际应用情况。

4 结束语

针对嵌入式盲人辅助身份识别设备，设计了 VQ-HMM 双层识别算法，提高了系统对身份识别的效率。采用了改进的端点检测算法，具有一定的抗噪能力，提高了识别的精度。进一步采用可以快速进行筛选的 VQ 量化算法，对样本进行快速处理，直接拒绝无效样本或者缩小模板库范围，再利用 DHMM 模型在缩小范围后的模板库内进行精确匹配，同时 VQ 的资源可以直接被后一层模型应用，最大程度上节省了系统的开销。系统在保持精度的前提下缩短了判决的时间，减少了系统的工作量，达到了算法对于精度、功耗、耗时的要求。实验表明，该算法在保证识别精度的同时，缩短了识别的时间，提高了识别的效率，可以满足系统的需要。鉴于今后的应用环境可能会更加复杂，抗噪能力将是下一步研究的重点。

参考文献：

[1] 韩纪庆，张磊，郑铁然. 语音信号处理[M]. 北京：清华大学出版社，2004：279-316.

[2] 赵力. 语音信号处理[M]. 北京：机械工业出版社，2003：125-188.

[3] 吴朝辉，杨莹春. 说话人识别模型和方法[M]. 北京：清华大学出版社，2008：56-108.

[4] 朱浩兵. 适用于特定人群的声纹识别研究[D]. 厦门：厦门大学，2008：25-38.

[5] 郭振兴，罗中明，王黎黎，等. 一种基于改进能零法的连续语音端点检测方法[J]. 哈尔滨理工大学学报，2009，14(1)：86-88.

[6] SAHIDULLAH M，SAHA G. A novel windowing technique for efficient computation of MFCC for speaker recognition[J]. IEEE Signal Processing Letters，2013，20(2)：149-152.

[7] 鲁晓倩. 基于 VP 树和 GMM 的说话人识别研究[D]. 合肥：中国科学技术大学，2014：43-46.

[8] SENOUSSAOUI M，KENNY P，STAFYLAKIS T，et al. A study of the cosine distance-based mean shift for telephone speech diarization[J]. IEEE/ACM Transactions on Audio，Speech，and Language Processing，2013，22(1)：217-227.

[9] REVATHI A，VENKATARAMANI Y. Speaker independent continuous speech and isolated digit recognition using VQ and HMM[C]//The 2011 International Conference on Communications and Signal Processing，2011：198-202.

[10] DILEEP A D，SEKHAR C C. GMM-based intermediate matching kernel for classification of varying length patterns of long duration speech using support vector machines[J]. IEEE Transactions on Neural Networks and Learning Systems，2014，25(8)：1421-1432.

[11] MALODE A A，SAHARE S L. An improved speaker recognition by using VQ & HMM[C]//The IET Chennai 3rd International on Sustainable Energy and Intelligent Systems，2012：1-7.

编辑：张所滨