

# 基于声纹识别的 Web 身份认证系统设计

曾 斌, 姚 路, 陈志诚

(海军工程大学管理工程系, 武汉 430030)

**摘 要:** 设计一个基于声纹识别的 Web 身份认证系统。在浏览器端利用自主开发的 ActiveX 录音控件录取封装使用者的声纹特征并传送给服务器, 服务器端使用隐马尔科夫模型表述单字, 单字之间通过增加静音状态分割以及语音训练形成稳定的声纹特征模型作为身份验证的基础库。实际测试结果表明, 该系统识别性能好、可移植性强, 适用于网络资源的远程声控。

**关键词:** 身份认证; 声纹识别; 隐马尔科夫模型; 动态时间校正

## Design of Web Identity Authentication System Based on Voiceprint Recognition

ZENG Bin, YAO Lu, CHEN Zhi-cheng

(Department of Management Engineering, Naval University of Engineering, Wuhan 430030, China)

**【Abstract】** This paper designs Web authentication system based on voiceprint recognition. An ActiveX control is developed in the browser to record and encapsulate the voice properties of the users and then transfer to the server. In the server side, a Hidden Markov Model(HMM) is designed to represent the Chinese words which are separated by an inserted silence state. And then a stable voiceprint database is constructed after the enough training of speech samples. Actual test results show that this system is suitable for controlling Web resource and has good identification performance and strong portability.

**【Key words】** identity authentication; voiceprint recognition; Hidden Markov Model(HMM); Dynamic Time Warping(DTW)

DOI: 10.3969/j.issn.1000-3428.2011.15.047

### 1 概述

近年来, 由于因特网和电子商务的迅速发展, Web 服务器需要存取大量资源并提供各种信息服务。目前, 网络上最普遍的身份认证方式属于密码验证存取行为, 文字密码容易被窃取或遗失, 甚至造成个人或单位财产的遗失。为了增加网站上数据的安全性, 在认证过程中, 除了使用密码方式进行验证外, 语音是人类身上最自然的特征之一, 声卡及麦克风是目前个人计算机的基本配置之一, 所以, 利用声音进行身份认证是所有生物验证系统中所需花费最小的一种。

声纹识别属于生物识别技术的一种, 它通过分析语音波形计算出说话人生理和行为特征的语音参数, 自动识别说话人身份。与语音识别不同, 声纹识别利用的是语音信号中的说话人信息, 而不考虑语音中的字词意思, 能够很好地满足因特网的身份认证要求<sup>[1]</sup>。如今, 声纹识别技术已逐渐走入实际应用, 其技术手段主要包括对声音特征进行线性或非线性计算、数据挖掘<sup>[2]</sup>和模式匹配, 如动态时间规整、隐马尔可夫模型(Hidden Markov Model, HMM)<sup>[3]</sup>、神经网络和多特征组合等技术。目前, 通常使用的算法主要有模板匹配法、最近邻方法、神经网络法、隐马尔可夫模型方法和 VQ 聚类方法等。实际证明使用隐含马尔可夫模型效果更好。它能很好地描述语音信号的时变性和平稳性, 例如, 基于 HMM 的声纹识别技术在环绕智能中的成功应用<sup>[4]</sup>。

选取声纹识别的特征参数在声纹识别系统中非常重要。作为准平稳的随机过程, 语音信号在 10 ms~25 ms 内可以认为是平稳的, 因此, 可以对语音信号进行分帧分析。目前, 比较有效的识别参数为 Mel 频率倒谱系数(Mel Frequency Cep-

strum Coefficient, MFCC), 在有信道噪声和频谱失真时, MFCC 参数表现比较稳定。由线性预测系数导出倒谱系数是一种常用的语音识别参数, 在安静的环境下, 线性预测倒谱系数与 MFCC 系数的性能相差不多。研究表明, 采用感觉加权的线性预测倒谱系数有更好的识别稳健性。

本文在网站资源的存取过程中加入一种声纹验证的机制, 以控制网站使用者的访问及重要资源的存取, 从而达到加强数据安全的目的。

### 2 基于声纹识别的 Web 身份认证系统设计

#### 2.1 系统结构

整个系统包括浏览器端录音程序、浏览器端与服务器端之间声纹数据的传递以及声纹训练与识别程序等部份。在录音程序方面需要具备能够在 Web 页面上运行程序的能力, 这方面的标准语言很多, 例如, Java 语言、嵌入式脚本语言和微软所提供的以组件对象模型(Component Object Model, COM)为基础的 ActiveX 技术等。其中, ActiveX 控件技术能够支持现有网站技术及其他因特网的标准, 具有自动下载并自行安装功能, 可开发功能强大的软件控件, 并提供一套安全问题的解决方案, 具备较好的集成功能, 这些优点正符合开发浏览器端录音程序的设计需求, 因此, 本文采用 ActiveX 控件的技术开发浏览器端录音控件。常见的因

**基金项目:** 湖北省自然科学基金资助项目(ZRY1086)

**作者简介:** 曾 斌(1970—), 男, 副教授, 主研方向: 信息安全, 身份认证系统; 姚 路, 讲师、硕士; 陈志诚, 讲师、博士研究生

**收稿日期:** 2011-02-23 **E-mail:** trueice@public.wh.hb.cn

特网应用程序设计包括采用公共网关接口(Common Gateway Interface, CGI)和应用服务提供商(Application Service Provider, ASP)等,由于 ASP 对 ActiveX 控件具有更好的支持,因此将以 ASP 技术开发浏览器端与服务器端之间的数据传递方式。常见的声纹验证技术有动态时间校正(Dynamic Time Warping, DTW)与隐藏式马尔科夫模型等,动态时间校正正在识别连续语音上的效果不佳且运算费时,当字词量增大时辨识速度较慢,这方面隐马尔科夫模型明显优于动态时间校正<sup>[5]</sup>。因此,本文选择隐马尔科夫模型作为声纹识别系统的设计方案。基于声纹识别的 Web 身份认证系统架构见图 1。

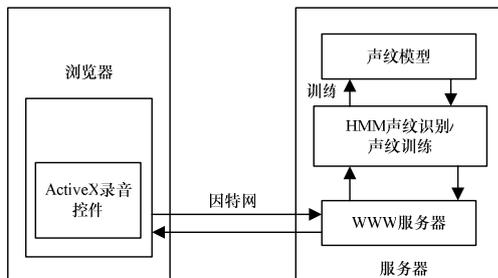


图 1 基于声纹识别的 Web 身份认证系统架构

2.2 浏览器端录制功能设计

根据图 1 的系统架构,在浏览器端安装一个 ActiveX 录音控件录制使用者的语音,一个简单的 ActiveX 控制控件只需具备自行注册能力与 IUnknown 接口即可。因此,本文设计的录音控件必须具备 IUnknown 接口及自行注册的能力,并能通过 IUnknown 接口对外公开录制的声音特征数据,提供浏览器存取录音控件所录制的声纹特征,并与网页上的窗体相结合,将数据上传服务器端处理。录音控件采用低阶波形应用程序接口(Application Program Interface, API)开发录音功能。为了能够与大多数声卡兼容,本录音程序设置采样频率为 11.025 KHz、单声道、8 位的取样质量以录制识别用的声音,同时加入能量计算程序以便能够具备自动静音功能,其能量  $E$  计算如下:

$$En(n) = \sum_{m=0}^{N-1} S^2(n+m) \quad (1)$$

录音程序所记录的数字化语音数据可通过数字信号处理算法抽取其特征值,本系统采用十阶倒频谱系数作为声纹特征参数。设置一个音框大小为 256 个取样点,相邻的 2 个音框重迭 81 个取样点。计算语音数据的自相关系数  $r(m)$ :

$$r(m) = \sum_{n=1}^{N-m} S(n)S(n+m) \quad 0 \leq n \leq N-1, 0 \leq m \leq N-1 \quad (2)$$

自相关系数再经由 Levinson-Durbin 算法计算其线性预估参数:

$$E^{(0)} = r(0) \quad (3)$$

$$k_i = \frac{r(i) - \sum_{j=1}^{i-1} a_j^{(i-1)} r(i-j)}{E^{(i-1)}} \quad 1 \leq i \leq P \quad (4)$$

$$a_i^{(i)} = k_i \quad (5)$$

$$a_j^{(i)} = a_j^{(i-1)} - k_i a_{i-j}^{(i-1)} \quad 1 \leq j \leq i-1 \quad (6)$$

$$E^{(i)} = (1 - k_i^2) E^{(i-1)} \quad (7)$$

$$a(j) = a_j^{(P)} \quad 1 \leq j \leq P \quad (8)$$

线性预估参数再转换成倒频谱参数:

$$\hat{c}_1 = -a_1 \quad (9)$$

$$\hat{c}_n = -a_n - \sum_{m=1}^{n-1} (1 - m/n) a_m \hat{c}_{n-m} \quad 1 < n \leq P \quad (10)$$

$$\hat{c}_n = -\sum_{m=1}^P (1 - m/n) a_m \hat{c}_{n-m} \quad P < n \quad (11)$$

最后获得代表语音频谱的包络,即倒频谱系数<sup>[6]</sup>,为一十阶的倒频谱系数矩阵,这一序列的倒频谱系数数组,如果一个一个对外发布,将造成存取上的不便,必须使用较简单的格式包装这一数据序列,再对外发布。字符串是最简单的格式之一,这一序列的声纹特征可使用格式化 I/O 的方式串接成一个字符串,并透过 IUnknown 接口对外发布该字符串变量,供浏览器存取,其转换方式如图 2 所示。

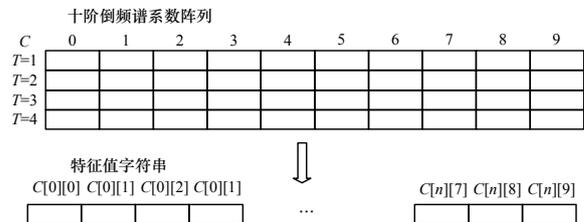


图 2 倒频谱系数数组转换为一个特征值字符串的示意图

在浏览器端,使用 ActiveX 控制控件录下的声纹特征,透过窗体的包装传送到 Web 服务器,窗体所传送数据将由内嵌的 ASP 程序存取。ASP 程序取得的数据是编码过的格式,需要译码为成对的字符串数值,并转换字符串中的加号与 16 进制数字才能还原为原来上传的数据,这个译码后取得的字符串是经过包装的,必须依浏览器端包装字符串的方式,反向取出字符串内的每一个数值放置于一个二维数组中(图 2 的逆向操作),这样以音框为单位的声纹特征序列,将作为声纹识别或训练使用。ASP 程序的另一个特色为动态产生 Html 文件并响应至浏览器端浏览器上,利用这个特性与声纹识别的结果相结合,可动态响应适当的网页内容到浏览器端,进而达到控制网页存取的目的。

2.3 服务器端声纹识别功能设计

服务器端取得声纹特征即可进行声纹模型的训练与识别。把隐马尔科夫模型应用在声纹训练上,可将声纹的变化看成是一连串状态的改变。对于每一个字,可以使用数个状态建立每个字的隐式马尔科夫模型,所取的状态数越多其声纹变化的表示将更详细,本系统所建立的隐马尔科夫模型架构,使用 9 个状态建立每个字的声纹模型,每个状态使用高斯分布描述其概率分布。对于一句语音信号,除了建立每个字的模型之外,还在首、尾及字与字之间加上一个静音状态,并配合隐马尔科夫模型的相关算法训练声纹模型。

在训练声纹模型前,需要设置隐马尔科夫模型的初值,隐马尔科夫模型通常以  $\lambda=(A, B, \pi)$  代表,参数  $A$  为状态转移概率矩阵,表示由某一个状态转移至另一个状态的概率值,其初值设为:

$$A = \begin{bmatrix} 1/2 & 1/2 & 0 & \dots & 0 \\ 0 & 1/2 & 1/2 & \dots & 0 \\ 0 & 0 & 1/2 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & 1 \end{bmatrix} \quad (12)$$

$\pi$  为起始状态的概率矩阵,表示语音初始状态的概率值,其初值设置为:

$$\pi = [1 \ 0 \ 0 \ \dots \ 0] \quad (13)$$

参数  $B$  为声音的观测序列符号在某个状态下的概率矩阵,其数学表示式为:

$$B = \{b_j(O_k)\} \quad b_j(O_k) = \Pr(O_k \text{ 在 } t \mid q_j \text{ 在 } t) \quad (14)$$

在式(14)中,  $b_j(O_k)$ 是 Gaussian 概率密度分布函数, 在训练前必须先设置每个状态的平均值及方差。开始训练前无法得知每个声纹数据的最优状态序列, 因此, 初值的设置只能先采用平均分配的方式, 将一序列的音框按次序平均分配到每一个状态上, 求其平均值及方差, 以建立初步的声纹模型状态。声纹训练主要采用 Viterbi 算法<sup>[6]</sup>。Viterbi 算法的主要目的在于找出训练语音的最优状态序列以及产生此状态序列的概率值, 声纹数据只要运行过一次 Viterbi 算法, 即可取得一条最优的状态序列, 记录将每个声纹序列产生的最优状态序列, 重新计算每个状态的平均值及方差, 获得一个更好的声纹模型。重复评估模型中的每个状态, 直到状态稳定为止, 即可求得最优的声纹模型, 如图 3 所示。

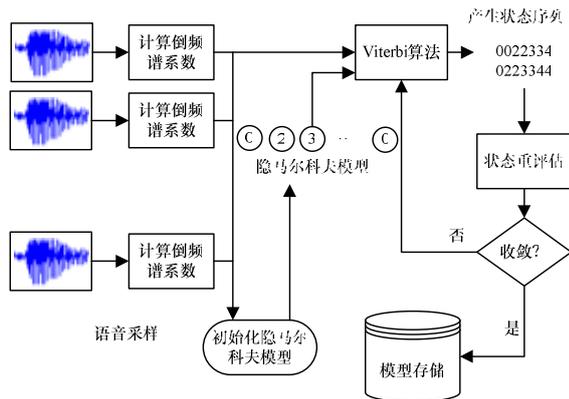


图 3 隐马尔科夫模型的声纹训练流程

身份认证系统通过声音识别以确认使用者的身份, 因此, 身份认证系统的第 1 步必须取得说话者的身份, 透过网页窗体的方式要求登入网站的使用者输入其身份, 取得其身份后即可加载进行比对的声纹模型, 再透过 Viterbi 算法计算声纹模型与语者输入声音的得分, 识别其身份。如果这个分数大于某一个阈值表示, 则说话者身份正确, 即动态产生网页响应到浏览器端, 提供身份正确的说话者可以存取网站资源, 其流程如图 4 所示。

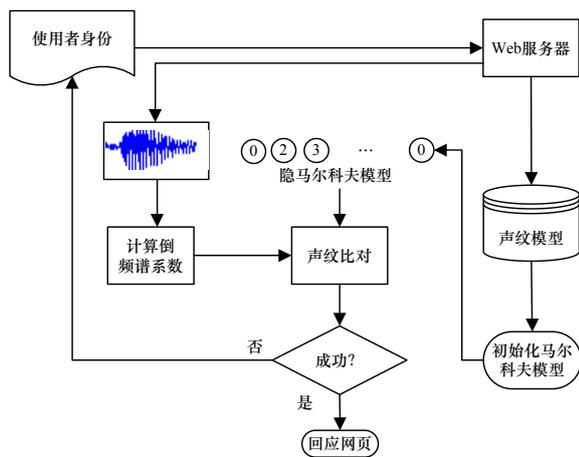


图 4 声纹识别流程

### 3 实验与结果分析

将声纹识别系统安装到 Web 网站上进行实际使用前, 先在本地端进行测试, 调整声纹识别系统。实验中所使用的声纹数据库是由 12 位学生于实验室中所录制采集的, 通过前面介绍的本地版录音程序, 录取 12 位学生的声纹特征。由于识别核心采用固定语句开发, 因此录下固定的语句作为实验中

测试使用的声纹数据, 每位语者录下“大家好”、“身份认证”以及“努力学习”3 句话, 每句话各录下 150 次。所录下的语句在第 1 个实验中, 将以每一个人的前 100 句声音训练声纹模型, 并观察声纹模型的收敛情形。在第 2 个实验中, 依第 1 个实验决定的训练语句数目建立声纹模型, 并研究识别阈值的设置。

第 1 个实验观察声纹训练所需语句数对于声纹模型的影响, 并计算出声纹训练所需的最少语句数目。首先选取声纹数据库中任意一位学生“大家好”这句话的前 100 句作为训练用语音样本, 分别依次训练每句话的声纹模型, 最后可以得到不同训练语句数目所产生的声纹模型。从声纹模型的前面、中间以及后面各挑出一个状态, 观察其收敛情形。从实验数据发现大约在 20 句话左右其值已趋于平缓, 到 50 句左右稍有变化, 50 句后趋于稳定, 因此, 将声纹识别系统所需的最小语句数目定为 50 句。

第 2 个实验讨论真实说话者被系统拒绝的错误率(类型 1 错误)及伪装者被系统接受的错误率(类型 2 错误), 通过这 2 项系统错误率探讨最优的识别阈值。学生 1 “大家好”这句话的类型 1 错误与类型 2 错误曲线如图 5 所示。可以看出, 大约在 -24 的位置为 2 条曲线交叉点, 这个交叉点说明了识别系统的阈值设置为这个数值时, 将得到最低的类型 1 及类型 2 的错误率, 即识别阈值设置为该数值就可得到最优的系统性能。以同样的方式观察其他实验者最优识别阈值的数值, 如表 1 所示。

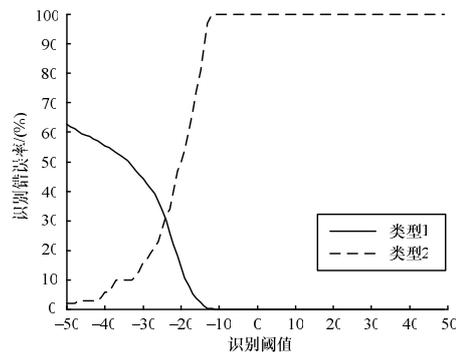


图 5 类型 1 及类型 2 识别错误率曲线

表 1 12 位学生不同语句的最优识别阈值

序号	大家好	身份认证	努力学习
1	-24	-10	-20
2	-25	-18	-20
3	-19	-24	-23
4	-20	-21	-20
5	-22	-15	-20
6	-27	-19	-27
7	-16	-10	-22
8	-19	-3	-16
9	-22	-13	-16
10	-50	-20	-35
11	-25	-19	-20
12	-24	-15	-20

训练完毕后的声纹识别系统安装在 Web 网站上, 使用者必须通过密码识别及声纹识别 2 道识别程序, 才能登录网站, 存取网站资源, 并且注意这些资源无法以超级链接的方式直接存取, 一定要通过这 2 道识别程序才能存取。

### 4 结束语

该系统具有很强的移植性和方便性, 给 Web 资源提供了双重身份认证手段, 若对其进一步完善和改进, 可广泛应用于电子商务、电子政务、网络安全, 特别是保密性质的服务

(下转第 167 页)