

北京大学软件与微电子学院信息安全工程

2017秋季课程项目

|  |  |
| --- | --- |
| 题 目 | Web声纹认证系统实现 |
| 成 员 | 曹路，李晖，杜思佳，储贤 |
| 指导老师 | 孙惠平 |

2018年1月2日

**目 录**

[摘要 3](#_Toc502494961)

[第一章 选题背景与意义 4](#_Toc502494962)

[1.1声纹认证简介 4](#_Toc502494963)

[1.2声纹认证优缺点 4](#_Toc502494967)

[1.3声纹认证应用领域 5](#_Toc502494967)

[1.3.1 信息领域 5](#_Toc502494975)

[1.3.2 银行、证券 5](#_Toc502494976)

[1.3.3 公安司法 5](#_Toc502494975)

[1.3.4 军队和国防 5](#_Toc502494976)

[1.3.5 保安和证件防伪 5](#_Toc502494975)

[第二章 产品与国内外研究现状 6](#_Toc502494968)

[2.1 现有产品 6](#_Toc502494970)

[2.2 国内外研究现状 6](#_Toc502494971)

[第三章 基本原理、实现思路与安全机制 7](#_Toc502494977)

[3.1 基本原理 7](#_Toc502494978)

[3.2 实现思路 8](#_Toc502494981)

[3.3 安全机制 9](#_Toc502494986)

[第四章 实验运行环境 1](#_Toc502494962)0

[4.1系统环境 1](#_Toc502494963)0

[4.2依赖包 10](#_Toc502494967)

[4.3开发IDE 10](#_Toc502494967)

[4.4核心框架（包） 11](#_Toc502494967)

[第五章 实验测试 11](#_Toc502494962)

[5.1基本实现目标 1](#_Toc502494963)1

[5.2实现流程 11](#_Toc502494967)

[5.2.1 搭建框架 11](#_Toc502494975)

[5.2.2 测试 11](#_Toc502494975)

[第六章 实验结果 1](#_Toc502494962)3

[第七章 结论 1](#_Toc502494962)6

[第八章 研究前景 1](#_Toc502494962)7

[参考文献 19](#_Toc502494991)

**摘 要**

声纹认证作为生物特征认证的一种，是根据说话人的声波特性进行身份辨识的服务。身份辨识与口音无关，与语言无关，可以用于说话人辨认和说话人确认，广泛应用于金融安全、智能家居、智慧建筑等领域。不同的任务和应用会使用不同的[声纹认证技术](https://baike.baidu.com/item/%E5%A3%B0%E7%BA%B9%E8%AF%86%E5%88%AB%E6%8A%80%E6%9C%AF)，如缩小刑侦范围时可能需要辨认技术，而银行交易时则需要确认技术。声纹认证就是把声信号转换成电信号，再用计算机进行识别。声纹认证可以说有两个关键问题，一是[特征提取](https://baike.baidu.com/item/%E7%89%B9%E5%BE%81%E6%8F%90%E5%8F%96)，二是[模式匹配](https://baike.baidu.com/item/%E6%A8%A1%E5%BC%8F%E5%8C%B9%E9%85%8D)([模式识别](https://baike.baidu.com/item/%E6%A8%A1%E5%BC%8F%E8%AF%86%E5%88%AB))。

在本次课程项目中，我们实现了基于MFCC和GMM的声纹认证系统，MFCC：梅尔频率倒谱系数 Mel-Frequency Cepstral Coefficients，Mel频率是基于人耳听觉特性提出来的, 它与Hz频率成非线性对应关系。MFCC则是利用它们之间的这种关系, 计算得到的Hz频谱特征, 一定程度上模拟了人耳对语音的处理特点。GMM：高斯混合模型 Gaussian Mixture Model，GMM模型是单一的高斯密度函数的扩展，可以逼近任意形状的概率密度分布，被广泛应用到语音识别领域。语音信号的特征向量被提取之后，需要对提取的特征向量建模模型训练根据特征参数建立高斯混合模型GMM。

**第一章 选题背景与意义**

近几十年来，语音处理技术已经得到了迅猛的发展，特别是在语音传输和数字语音存储方面的发展给人类的生活带来了极大的便利。传统的身份确认系统通常利用个人所知道的信息来作为身份确认的依据，如使用密码或者个人物品来判断，比如身份证、密保卡等。但是随着个人账号的不断增多，与之相对应的记忆成本与日俱增，传统方法正面临着巨大的考验。

生物学特性作为身份认证已经被越来越多的运用到人们的日常生活中。常见的生物特征有指纹、人脸、视网膜和声音等[1]，在所有的生物特征中，语音作为人类交流最自然的特征之一，其具有与指纹一样的能够识别身份的特性。因此，研究声纹识别(Voiceprint recognition)将可以给人们的日常生活带来极大的便利。

**1.1声纹认证简介**

所谓声纹(Voiceprint)，是用电声学仪器显示的携带言语信息的声波频谱[2]。人类语言的产生是人体语言中枢与发音器官之间一个复杂的生理物理过程，人在讲话时使用的发声器官--舌、牙齿、喉头、肺、鼻腔在尺寸和形态方面每个人的差异很大，所以任何两个人的声纹图谱都有差异。每个人的语音声学特征既有相对稳定性，又有变异性，不是绝对的、一成不变的。这种变异可来自生理、病理、心理、模拟、伪装，也与环境干扰有关。尽管如此，由于每个人的发音器官都不尽相同，因此在一般情况下，人们仍能区别不同的人的声音或判断是否是同一人的声音[3]。

**1.2声纹认证的优缺点**

**优点：**

* 获取方便、自然，声纹提取可在不知不觉中完成，用户接受程度较高
* 语音的识别成本低廉，使用简单，一个麦克风即可
* 适合远程身份确认，通过通讯网络或互联网络即可实现远程登录
* 声纹辨认和确认的算法复杂度低

**缺点：**

* 易受身体状况、年龄、情绪等的影响
* 环境噪音对识别有干扰
* 多人说话混合环境下人的声纹特征不易提取

**1.3 声纹认证应用领域**

**1.3.1信息领域**

在自动总机系统中，把身份的声纹辨认和关键词检出器结合起来，可以在姓名自动拨号的同时向受话方提供主叫方的身份信息。前者用于身份认证，后者用于内容认证。同样，声纹识别技术可以在呼叫中心(Call Center)应用中为注册的常客户提供友好的个性化服务。

**1.3.2银行、证券**

鉴于密码的安全性不高，可以用声纹识别技术对电话银行、远程炒股等业务中的用户身份进行确认，为了提供安全性，还可以采取一些其他措施，如密码和声纹双保险，如随机提示文本用文本相关的声纹识别技术进行身份确认(随机提示文本保证无法用事先录好的音去假冒)，甚至可以把交易时的声音录下来以备查询。

**1.3.3公安司法**

对于各种电话勒索、绑架、电话人身攻击等案件，声纹辨认技术可以在一段录音中查找出嫌疑人或缩小侦察范围；声纹确认技术还可以在法庭上提供身份确认（同一性鉴定）的旁证。在监狱亲情电话应用中，通过采集犯人家属的声纹信息，可有效鉴别家属身份的合法性。

**1.3.4军队和国防**

声纹辨认技术可以察觉电话交谈过程中是否有关键说话人出现，继而对交谈的内容进行跟踪(战场环境监听)；在通过电话发出军事指令时，可以对发出命令的人的身份进行确认。

**1.3.5保安和证件防伪**

机密场所的门禁系统等。又如声纹识别确认可用于信用卡、银行自动取款机、门、车的钥匙卡、授权使用的电脑、声纹锁以及特殊通道口的身份卡，把声纹存在卡上，在需要时，持卡者只要将卡插入专用机的插口上，通过一个传声器读出事先已储存的暗码，同时仪器接收持卡者发出的声音，然后进行分析比较，从而完成身份确认。同样可以把含有某人声纹特征的芯片嵌入到证件之中，通过上面所述的过程完成证件防伪。

**第二章 产品与****国内外研究现状**

**2.1现有产品**

基于语音处理技术，国内外在语音识别方面发展的非常迅速，例如我国的科大讯飞[4]、Google的云语音API[5]、微软的Bing语音[6]，但是在声纹识别方面的发展相对缓慢。声纹认证逐渐被作为一种双因子认证应用于日常的软件中，例如微信、支付宝的声纹登录，科大讯飞的“声纹认证+人脸识别P2P转帐”等。2015年诞生于美国硅谷的speakIn，目前已经是行业领先的声纹识别与身份安全解决方案提供商。

除了现有的产品，各大公司和科研机构还提供了一些声纹识别库，例如Microsoft Speaker Recognition API[7]、IDIAP Research Institute的Bob[8]、Will Drevo基于Python和Numpy开发的*Dejavu*[9]等。

1. Microsoft Speaker Recognition API

微软开发出的这个API识别较为精准，而且处理速度较快，但由于其用到的口令串是标准的，攻击者能提前录制这些短语，从而入侵账户。此外，该服务是非开源付费使用的。

1. Bob

Bob是由Switzerland的一家名为Idiap的研究机构里面的生物特征安全和隐私保护小组开发的一套声纹识别工具箱。但是该工具箱安装会占用大量空间和内存，并且语音处理的时长较长。Dejavu在匹配声音样本上十分精确，并且其对音频的处理时间非常的短，

1. Dejavu

Dejavu是一个轻量级、易用、准确的Python音频指纹库，由麻省理工大学的Will Drevo基于Python and Numpy开发，因此本系统就是基于该音频指纹库实现。

**2.2国内外研究现状**

1976年德州仪器（Texas Instruments）制作了第一个说话人识别的原型[10]。后来，NIST（National Institute of Standard and Technology）在语音处理方面做出了极大的贡献。声纹识别的发展归功于特征提取和建模两种技术的进步，早期的与文本相关的说话人识别使用动态时间弯曲（DTW）和模板匹配技术。最近对说话人识别的研究主要集中在与文本无关的说话人识别方面。特征提取技术主要基于短时语音帧分析，语音信号被设定为准平稳的，一般情况下语音的帧长为8-30ms，采样频率一般为8kHz-16kHz。倒谱分析和梅尔倒谱分析（MFCC）是说话人识别中最常用的短时分析方法，线性预测（LP）并不常用，但是很多时候常和MFCC结合来使用。

常用的建模方法包括高斯混合模型（GMM）[11]，隐马尔科夫模型（HMM）[12]，支持向量机模型（SVM）[13]，矢量量化模型（VQ）[14]和人工神经元网络（ANN）[11]。

HMM常被用做与文本相关的说话人确认，而GMM、SVM、VQ主要用做与文本无关的说话人识别。其中GMM模型被认为是现在最优秀的建模方法。高斯混合模型是一种高斯概率密度函数（PDF）加权和的向量集。常被看做是单状态连续隐形马尔科夫模型，或者看作为“软”矢量化模型。

随着高斯混合模型技术的快速发展，支持向量机技术逐渐被取代。现在运用最多的技术是使用MFCC特征向量和基于GMM模型的识别，这两项技术到目前仍然是最先进的技术。所以本系统采用的就是采用MFCC特征向量和基于GMM模型的声纹识别。

**第三章 基本原理、实现思路与安全机制**

**3.1基本原理**

声纹认证一般分为用户注册过程和用户登录认证过程：

* 在用户注册过程中，首先对用户输入的用户名和文本口令进行记录，之后对用户所处的背景噪音进行记录，从语料库中随机抽取单词生成临时口令串（temporary password string），提示用户朗读该口令串进行语音注册。使用LTSD(long-term spectral divergence)[16]对所记录到的语音进行背景噪声去除。对去噪后的音频进行语音识别，将识别结果与临时口令串进行模糊匹配（Fuzzy matching），若达到判断阈值则接受这段语音，反之，则重新生成新的口令串让用户朗读。提取音频信号的梅尔频率倒谱系数[17]（Mel-Frequency Cepstral Coefficients, MFCC）特征并进行正则化表示。根据特征参数用最大期望算法（Expectation Maximization Algorithm, EM算法）为注册用户建立高斯混合模型[18]（Gaussian Mixture Model, GMM）。
* 在用户登录认证过程，提示待认证用户输入用户名和文本口令，从语料库中随机抽取单词生成临时口令串，提示待认证用户朗读该口令串，使用LTSD对所记录到的语音进行背景噪声去除。对去噪后的音频进行语音识别，将识别结果与临时口令串进行模糊匹配，若达到判断阈值则接受这段语音，反之，则重新生成新的口令串让待认证用户朗读。提取音频信号的MFCC特征并进行正则化表示。根据特征参数用EM算法为注册用户建立GMM模型,与注册用户库的GMM模型进行对数似然估计，若满足阈值，则认为用户是合法用户准许用户进入，反之，则拒绝用户登入系统。

**3.2 实现思路**

* 用户注册过程

1. 输入用户名和文本口令。

2、使用Python中的***RandomWords***从语料库中随机抽取1个单词。用户需朗读该单词进行语音注册。

3、使用Python中的***Google Speech Recognition***对采集音频进行语音识别，使用Python中的***FuzzyWuzzy***进行模糊匹配（计算Levenshtein距离），设定模糊匹配阈值分数为85分。除此之外，若用户在口令串生成后2秒内没有进行朗读，则系统判定超时，重新生成口令串。

4、使用Python中的***python\_speech\_features***提取音频信号的梅尔频率倒谱系数MFCC。

1. 使用Python中Sci-kit learn所实现的GMM中的***fit***方法，可以得到用户的GMM模型。

* 用户登录认证过程

1、 输入用户名和文本口令。

2、使用Python中的***RandomWords***从语料库中随机抽取1个单词。用户需朗读该单词进行语音录音。

3、使用Python中的***Google Speech Recognition***对采集音频进行语音识别，使用Python中的***FuzzyWuzzy***进行模糊匹配（计算Levenshtein距离），设定模糊匹配阈值分数为85分。除此之外，若用户在口令串生成后2秒内没有进行朗读，则系统判定超时，重新生成口令串。

4、 使用Python中的***python\_speech\_features***提取音频信号的梅尔频率倒谱系数MFCC。

5、 使用Python中Sci-kit learn所实现的GMM中的***fit***方法，可以得到待认证用户的GMM模型。

6、使用Python中Sci-kit learn所实现的GMM中的***score\_samples***方法得到待认证模型与注册用户的对数似然估计值，选择最佳对数似然估计值，若满足阈值，则准许用户进入系统，反之，则拒绝用户登入系统。

**3.3安全机制**

1、与文本有关和与文本无关

根据对识别时所用的语音内容要求的不同，可将声纹识别分为2类[19]：与文本有关(text-dependent)、与文本无关(text-independent)。与文本有关的声纹识别需要在用户注册时就确定识别所用的发音内容，由于文本内容是已知的，攻击者则可以通过悄悄录音、诱导用户说指定文字等手段，窃取到用户的登录声纹，基于此，攻击者就能轻而易举的进入系统。而本文设计的系统采用的是与文本无关机制，用户在注册、登录时使用的临时口令串都是随机生成的，系统是通过提取用户的声音特征并与之前录入的特征进行匹配，即使攻击者对用户进行录音也无法窃取用户的声纹信息从而登录系统。

2、中间人攻击

有一种潜在的攻击类型就是中间人攻击，攻击者可能监听该系统的客户端和服务器端，当用户进行注册时，客户端会将用户音频文件发送回后台服务器端，中间人在这时会拦截该音频文件，并将其篡改成自己的音频文件发送回服务器端，这样，攻击者可以通过自己的声纹信息登录系统。为此，本系统设计了一种防止中间人攻击的机制，客户端每次将音频文件发送给服务器端之前，进行数字签名操作，用自己的RSA私钥签署音频文件的SHA256散列[20]，并将音频文件连同数字签名一起发送给服务器端，若中间人对该音频文件进行篡改，则服务器端会验证失败。

**第四章 实验运行环境**

**4.1系统环境**

Linux环境，Fedora 27 64bit

Python2.7，Mysql

VMware® Workstation 14 Pro，WIN

**4.2依赖包**

pyaudio for grabbing audio from microphone

• ffmpeg for converting audio files to .wav format

• pydub, a Python ffmpeg wrapper

• numpy for taking the FFT of audio signals

• scipy, used in peak finding algorithms

• matplotlib, used for spectrograms and plotting

• MySQLdb for interfacing with MySQL databases

**4.3开发IDE**

Pycharm

**4.4核心框架（包）**

Flask（http://docs.jinkan.org/docs/flask）

Dejavu (<https://github.com/worldveil/dejavu>)

**第五章 实验测试**

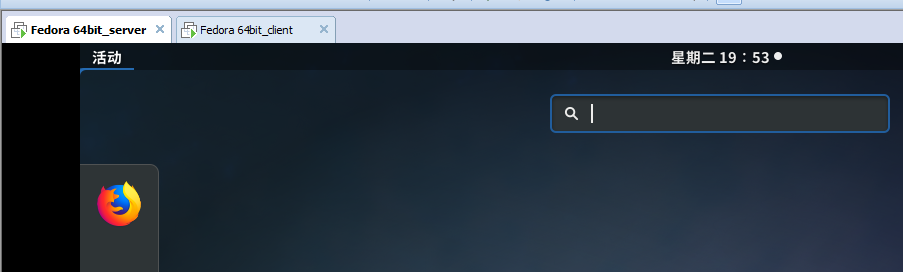
**5.1基本实现目标**

实现一个BS架构的声纹入库与采集系统（B端负责显示和声纹采集，S端负责声纹入库、匹配、计算等多个操作）

**5.2实现流程**

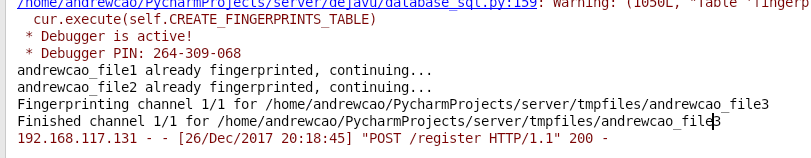
**5.2.1搭建框架**

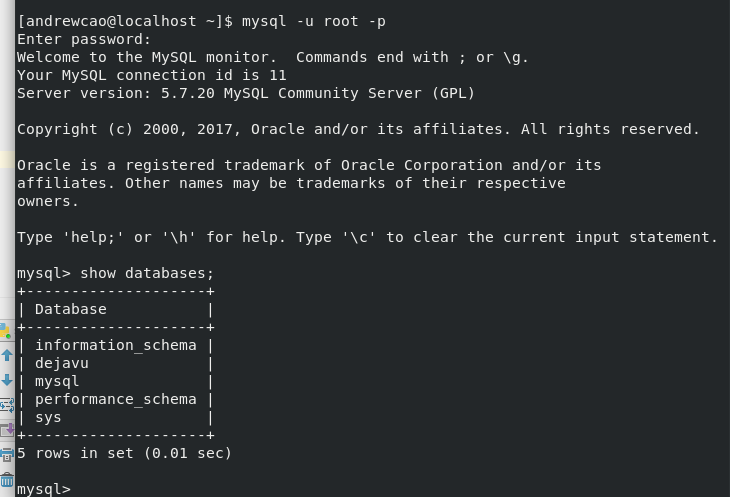
分成Server端和Client，为了方便模拟演示，我们使用两个Fedora系统分别模拟，添加好所有依赖包

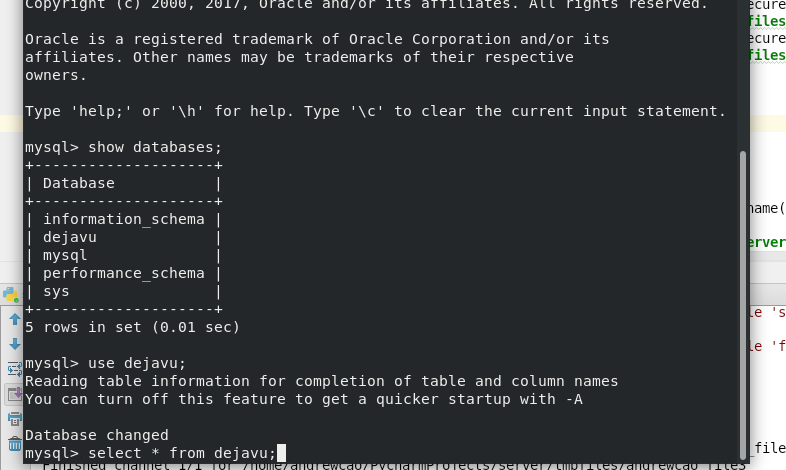


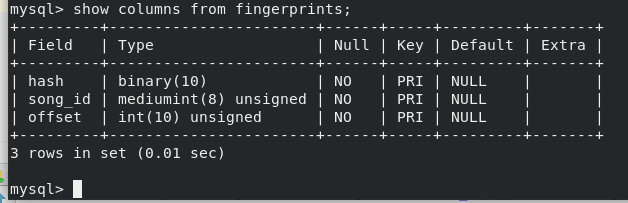
基于Flask框架实现了一个BS架构，具体原理看Flask官方文档，简单来说就是搭建一个不断轮询、等待C端连接的B端服务机，

**5.1.2测试**



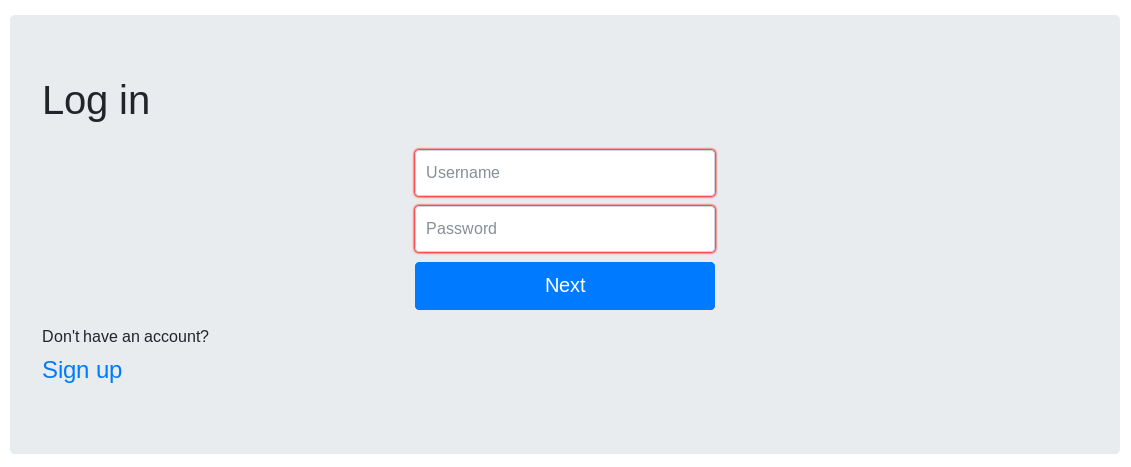




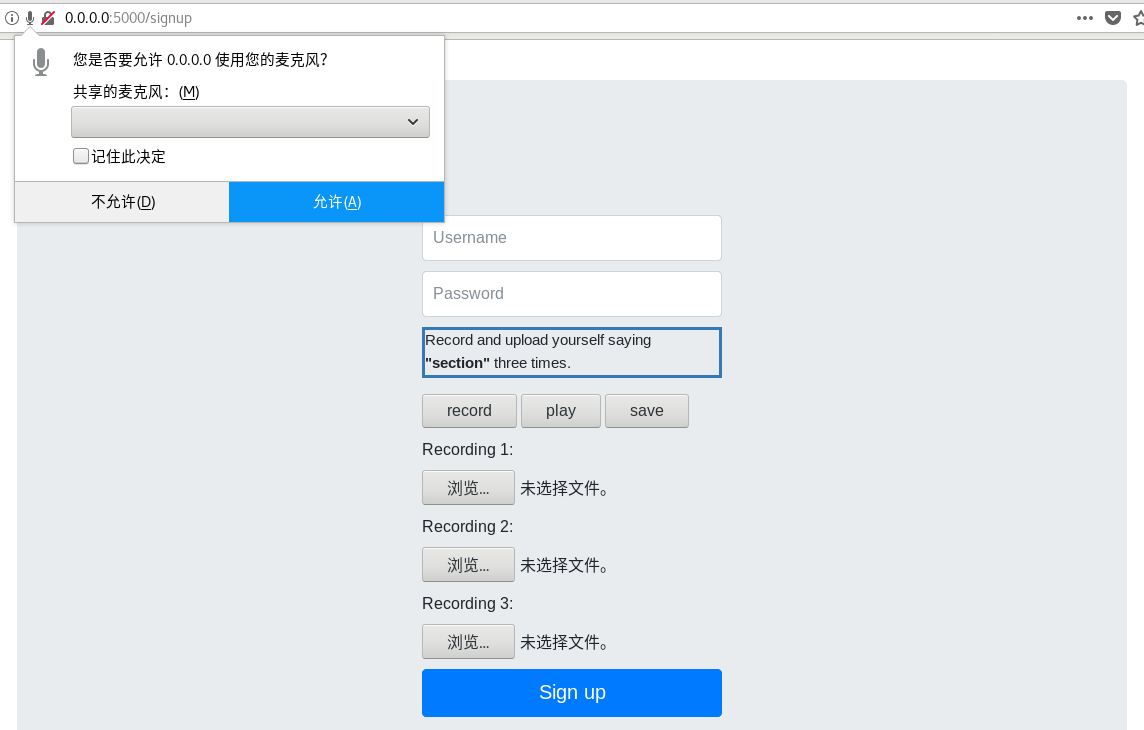


**第六章 实验结果**

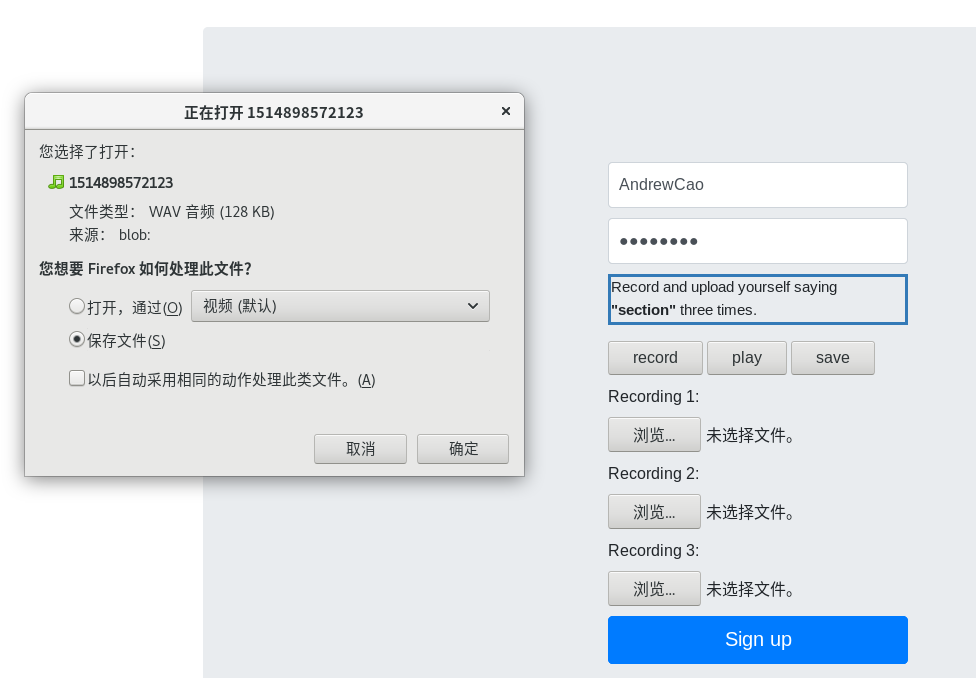
登录界面：



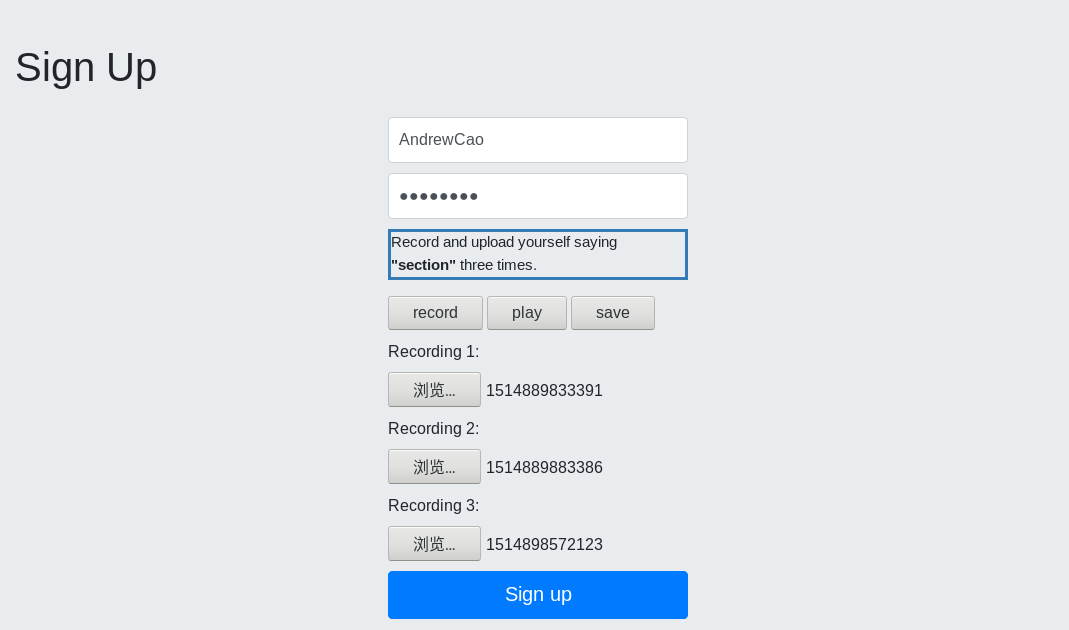
注册界面：



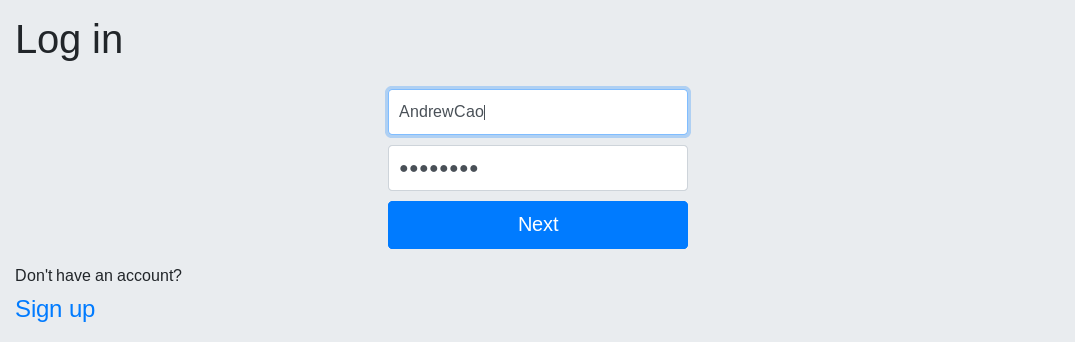
录音：



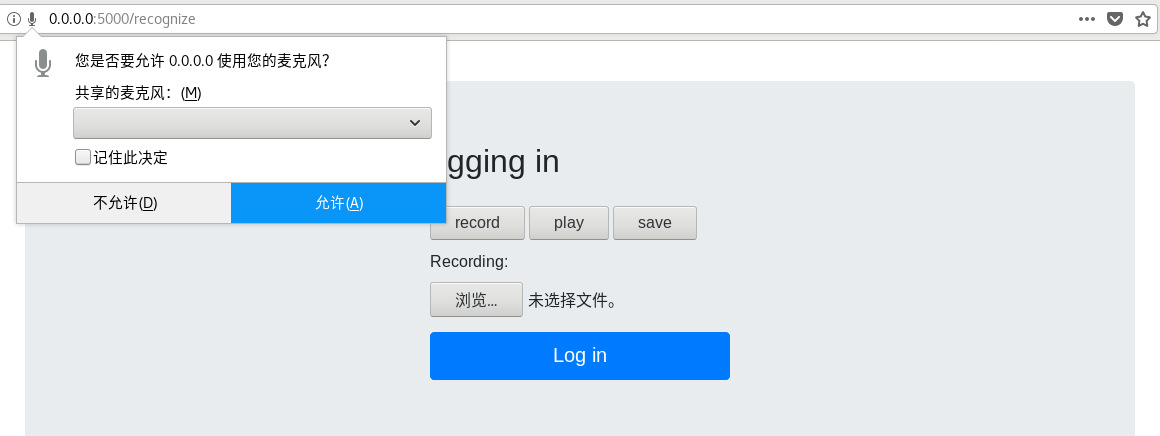
上传录音文件：



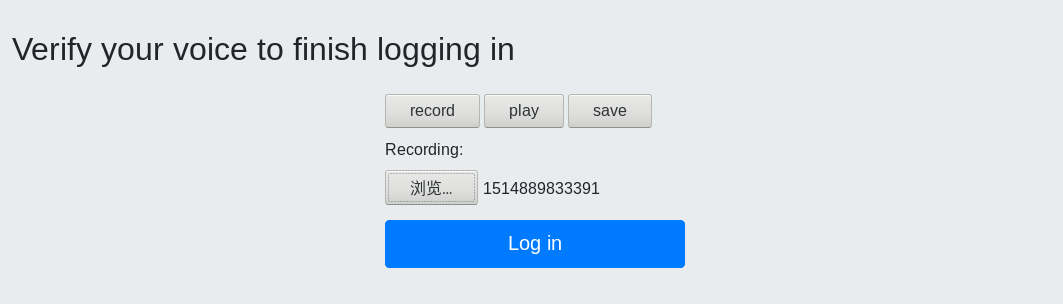
登录：



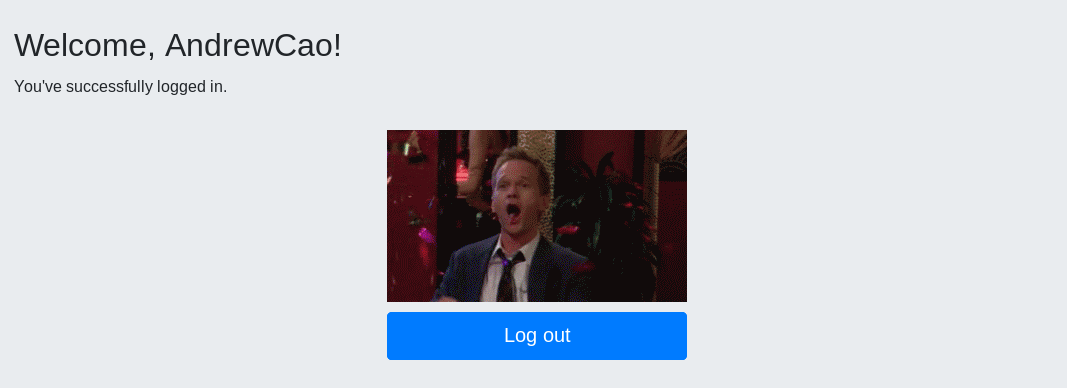
上传录音文件：



音频识别：



登录成功



**第七章 结论**

声纹认证技术为现在的身份鉴别提供了非常有力的、可靠的解决途径。有着深远的 社会价值、经济价值。也对我国的战略安全，信息安全有着重要的意义。是各国正在发 展的重点战略技术。声纹认证技术作为生物认证技术的重要分支已经引起科研机构和企 业界的高度重视，前景非常乐观。 在前人理论的基础上，经过几个月的努力工作，本文在详细的分析了声纹认证理论的基础上，完成了与文本有关声纹辨认模式和算法的构建，并进行了具体的实验。

本文研究的内容是基于MFcc倒谱系数和高斯混合模型(GMM)的说话人识别系统。其中主要是针对说话人识别所采用的特征向量作了研究，提出了在普通倒谱系数的提取过程中使用音调特性来进行特征提取的方案，即用Mel频率刻度对频率轴进行弯折，从而得到了Mel频率倒谱系数(MFCC)，并给出了MFCC倒谱系数的具体提取过程和算法。由于传统的倒谱特征参量提取方法主要是基于LPcc特征，本文通过实验验证了MFcc倒谱特征的系统识别性能略好于LPcc倒谱特征的性能，所以在MFcc特征的基础上对识别过程中所使用的特征作了改进。

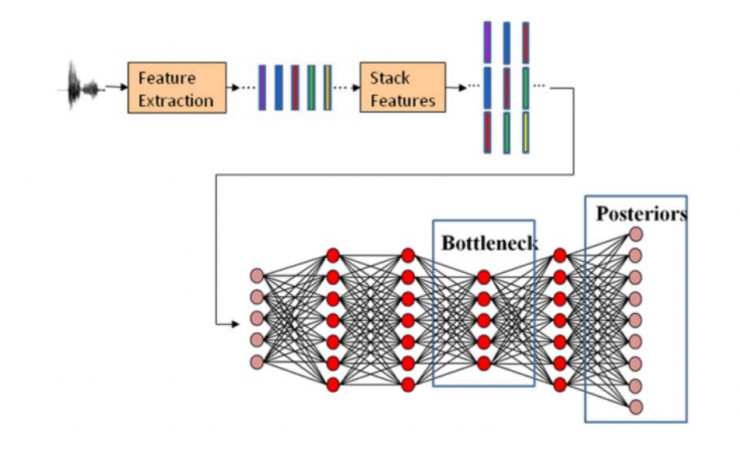
本文讨论的基于高斯混合模型的说话人识别系统，该系统的一个主要特点就是文本无关性。本文详细分析了梅尔倒谱系数的提取，以及高斯混合模型的建立、识别过程。从实验中我们也发现，识别的性能还和麦克的质量有很大关系，一个优质的麦克对提高识别率有着关键性的影响；环境的因素也非常重要，最理想的环境是宽敞的安静的开放空间，这里不仅没有回音，噪音也相对较小。 尽管声纹认证技术已经进行了很多年，也取得了很大的发展，但总的来说，想要真正的广泛应用，还有很长的路要走。

**第八章 研究前景**

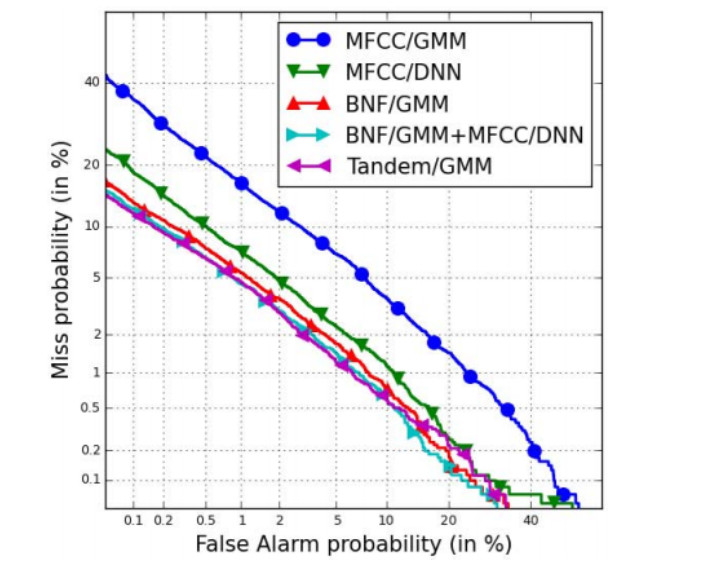
由于说话人认证是一个比较复杂的问题，受时间，空间和环境的影响较大，要解决这个问题，一方面要选取较高鲁棒性的特征参量，另一方面要联合多个不同特征的优势，将不同的特征结合起来构成复合特征向量来使用，而所有的这些都需要在今后的工作中加以研究和实践。

声纹认证也和其他认证一样，也向着深度学习的方向发展，但是又和语音识别稍有差异，传统算法和模型在声纹认证中还占有相当大的比重。

　　下图是Fred Richardson在论文中提出的声纹识别的深度学习模型示意



　　深度学习的效果还是有的，下图就是各种方法的一种比较，也就说，将来实时声纹识别将会有比较大的突破。



　　但是难度也很大，因为深度学习是基于数据驱动的模型，需要庞大的数据，这些数据最好是真实场景的数据，以及对数据的精确标注，这些都是很费钱很费人的事情。而且声纹识别训练库的建立，至少要保证性别比例分布为50%±5%，包含有不同年龄段、不同地域、不同口音、不同职业。同时，测试样本应该涵盖文本内容是否相关、采集设备、传输信道、环境噪音、录音回放、声音模仿、时间跨度、采样时长、健康状况和情感因素等影响声纹识别性能的主要因素。

　　也就是说，声纹认证对数据的要求其实比语音识别还要高很多，这本身就是个很大的门槛，也是突破声纹认证，真正能让声纹认证落地千家万户的核心因素。希望有一天，说话人认证技术能够真正成熟起来，与其它技术结合起来在现实生活中大放异彩。

**参考文献**

1. 杨佳东. 与文本无关的嵌入式声纹识别门禁系统[D]. 吉林大学, 2004.
2. 谷志新. 基于声纹信息的身份认证模式与算法的研究[D]. 东北林业大学, 2005.
3. Zhang W. METHOD OF ANSWERING INCOMING CALL, AND MOBILE TERMINAL:, WO/2014/169644[P]. 2014.
4. 科大讯飞. 探索语音识别技术的前世今生[J]. 科技导报, 2016, 36(9):76-77.
5. <https://cloud.google.com/speech/>CLOUD SPEECH API
6. <https://azure.microsoft.com/zh-cn/services/cognitive-services/speech/?cdn=disable>
7. <https://azure.microsoft.com/en-us/services/cognitive-services/speaker-recognition/>
8. <https://pythonhosted.org/bob.bio.spear/index.html>
9. <http://willdrevo.com/fingerprinting-and-audio-recognition-with-python/>
10. 蒋晔. 基于文本无关的说话人识别技术研究[D]. 南京理工大学, 2008.
11. Reynolds D A. Speaker identification and verification using Gaussian mixture speaker models [J]. Speech Communication, 1995, 17(1–2):91-108.
12. Che C W, Lin Q, Yuk D S. An HMM approach to text-prompted speaker verification[J]. 1996, 2:673-676 vol. 2.
13. Campbell W M, Campbell J P, Reynolds D A, et al. Support vector machines for speaker and language recognition[J]. Computer Speech & Language, 2006, 20(2–3):210-229.
14. Soong F, Rosenberg A, Rabiner L, et al. A vector quantization approach to speaker recognition[J]. AT&T Technical Journal, 1987, 66(2):387-390.
15. Ramı́Rez J, Segura J C, Benı́Tez C, et al. Efficient voice activity detection algorithms using long-term speech information[J]. Speech Communication, 2004, 42(3–4):271-287.
16. Chen Y, Wang J Z. A region-based fuzzy feature matching approach to content-based image retrieval[J]. Pattern Analysis & Machine Intelligence IEEE Transactions on, 2002, 24(9):1252-1267.
17. 胡峰松, 张璇. 基于梅尔频率倒谱系数与翻转梅尔频率倒谱系数的说话人识别方法[J]. 计算机应用, 2012, 32(9):2542-2544.
18. 于娴, 贺松, 彭亚雄,等. 基于GMM模型的声纹识别模式匹配研究[J]. 通信技术, 2015, 48(1):97-101.
19. 马纯艳. 一种基于GMM的汽车声纹识别锁算法研究[D]. 南京理工大学, 2014.
20. 韩冰, 刘丰, 王平. 基于RSA和SHA-256算法实现电子文档数字签名[C]// 中国西部地区信息技术学术研讨会. 2006.