

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 | 一种基于MFCC和GMM的声纹认证系统实现 |
| 成 员 | 曹路，李晖，杜思佳，储贤 |
| 指导老师 | 孙惠平 |

2017年10月24日

**一、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. **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. **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

**（三） 声纹认证应用领域**

**a)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。

**b)****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。

**c)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。

**d)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。

**e)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。

**二、产品与****国内外研究现状**

**（一）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。

**（二）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。

**三、****基本原理、实现思路与安全机制**

**（一）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。

**（二）实现思路**

* 用户注册过程

1. 输入用户名和文本口令。
2. 对用户所处的环境噪声进行记录，记录时长为5秒。

3、使用Python中的***RandomWords***从语料库中随机抽取8-10个单词构成口令串。用户需朗读该单词串进行语音注册。

1. 使用LTSD拟合音频包络线，消除背景噪音的干扰。

5、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。

1. 使用Python中的***python\_speech\_features***提取音频信号的梅尔频率倒谱系数MFCC。
2. 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、 对用户所处的环境噪声进行记录，记录时长为5秒。

3、使用Python中的***RandomWords***从语料库中随机抽取8-10个单词构成口令串。用户需朗读该单词串进行语音录音。

3、 使用LTSD拟合音频包络线，消除背景噪音的干扰。

5、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。

**（三）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。

**四、预期结果和成果形式**

（一）预期结果

1、能够对注册用户的声纹进行预处理、背景噪音去除、声纹特征提取以及特征存储。

2、能够实现合法用户通过输入用户名、文本口令和语音信息进行登录。

（二）成果形式

可用于用户注册或者登录的网页版声纹认证系统

**五、****项目分工与进度安排**

1. 项目分工

1、前期（选题~开题汇报）

曹路（组长）负责：选题、现有声纹库的调研、认证系统设计、开题报告的撰写。

李晖负责：课题背景的调研，开题报告的撰写。

杜思佳负责：声纹认证系统原理调研。

储贤负责：当前现有的声纹产品以及国内外研究现状的调研，开题报告的撰写。

2、后期（开题汇报后~项目展示）

曹路（组长）负责：认证系统的搭建与开发，项目展示PPT撰写

李晖负责：认证系统的开发，系统测试，项目展示PPT撰写

杜思佳负责：认证系统的开发，系统测试，项目展示PPT撰写

储贤负责：认证系统的开发，系统测试，项目展示PPT撰写

（二）进度安排

2017.11.11 明确项目选题

2017.11.18—2017.11.20 查阅相关文献，明确调研方向

2017.11.21 小组讨论调研进展，确定下一步内容

2017.11.21—2017.11.23 小组讨论，撰写开题报告和展示PPT

2017.11.24 提交开题报告和展示PPT

2017.11.25 展示PPT

2017.11.25 基于开题报告，开发相应系统，实现对应的功能

**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.