常熟理工学院

本科毕业设计（论文）

题　目　基于声纹识别的门禁应用

学　院　计算机科学与工程学院

年　级　2014　专 业　软件工程（单招）

班　级　0922141　学 号　092214109

姓　名　张立飞

校内导师　赵彩云　职 称　副教授

校外导师　　职 称　

常熟理工学院本科毕业设计(论文)诚信承诺书

本人郑重声明： 所呈交的本科毕业设计(论文)，是本人在导师的指导下，独立进行研究工作所取得的成果。除文中已经注明引用的内容外，本论文不含任何其他个人或集体已经发表或撰写过的作品成果。对本文的研究做出重要贡献的个人和集体，均已在文中以明确方式标明。本人完全意识到本声明的法律结果由本人承担。

本人签名： _____ 日期：

常熟理工学院本科毕业设计(论文)使用授权说明

本人完全了解常熟理工学院有关收集、保留和使用毕业设计(论文)的规定，即：本科生在校期间进行毕业设计(论文)工作的知识产权单位属常熟理工学院。学校有权保留并向国家有关部门或机构送交论文的复印件和电子版，允许毕业设计(论文)被查阅和借阅；学校可以将毕业设计(论文)的全部或部分内容编入有关数据库进行检索，可以采用影印、缩印或扫描等复制手段保存、汇编毕业设计（论文），并且本人电子文档和纸质论文的内容相一致。

保密的毕业设计(论文)在解密后遵守此规定。

本人签名： _____ 日期：

导师签名： _____ 日期：

基于声纹识别的门禁应用

摘　要

声纹识别是语音识别中的一种，可以通俗得称之为说话人识别。近年来，随着人工智能逐渐进入大众的视线，基于神经网络的身份认证方法得到了进一步的发展。比如人脸识别，虹膜识别以及声纹识别。这些身份认证方法逐渐从理论阶段演化为实际的应用。

传统的门禁需要用户携带一定的密钥设备如钥匙或门禁卡，用户需要将密钥设备直接接触认证系统，认证系统读取密钥信息后才能够开启。身份认证技术随着科技的发展不断得简化认证步骤，优化认证效率，为信息安全提供着保护，为身份认证增加了便捷。本课题使用声纹识别的方式为用户提供一种简便的，无需携带任何物件的智能认证方式。

本课题通过读取标准用户语音，分析出用户所说话的特征参数，为其建立概率统计模型。当待识别用户想要通过门禁时，朗读事先设置好的文字，系统通过采集待识别用户的语音，分析完特征参数后放入标准用户的模型中计算出概率匹配值。若此概率匹配值满足一定情况，则表明此待识别用户即为标准用户。识别过程结束后，系统会将识别日志与待识别者的语音跟相应识别情况通过 Socket 发送给服务器，服务器接收到消息后保存至数据库并实时通知前台刷新显示信息。

关键词：声纹识别　身份认证　声纹识别模型　概率匹配

Access control application based on voiceprint recognition

Abstract

Voiceprint recognition is a type of speech recognition that can be commonly referred to as speaker recognition. In recent years, with the artificial intelligence gradually entering the public eye, the authentication method based on the neural network has been further developed. Such as face recognition, iris recognition and voiceprint recognition. These identity authentication methods gradually evolved from the theoretical stage into practical applications.

The traditional access control requires the user to carry a certain key device such as a key or an access control card. The user needs to directly contact the key device with the authentication system, and the authentication system can only open the key information. With the development of technology, identity authentication technology continues to simplify the authentication process, optimize the authentication efficiency, provide protection for information security, and increase convenience for identity authentication. This topic uses voiceprint recognition to provide users with a simple, intelligent authentication method that does not require carrying any object.

This topic reads the standard user speech, analyzes the characteristic parameters spoken by the user, and establishes a probability and statistics model for it. When the user to be identified wants to pass through the access control, the pre-set text is read aloud, and the system collects the voice of the user to be identified, analyzes the feature parameters, and then puts it into the standard user's model to calculate the probability matching value. If this probability matching value satisfies a certain condition, it indicates that the user to be identified is a standard user. After the identification process is completed, the system sends the recognition log and the voice of the person to be identified and the corresponding recognition situation to the server.

After the server receives the message, it saves it to the database and notifies the front desk to refresh the display information.

Key Words：Voiceprint recognition authentication voiceprint recognition model probability matching

1. 目　录

1. 引言

1.1. 声纹识别的研究背景与意义

随着时代的发展，信息技术与网络通信走遍中国的每一个角落，人们生活节奏的逐渐加快，使得身份验证的数字化、快捷化显得越来越重要。传统的以密码为基础的身份验证技术开始逐渐出现弊端，已经无法再满足现在身份认证中安全性和实效性的要求。然而随着信息科学的快速发展，近几年人工智能的火热，近年来逐渐发展起来的生物认证技术正成为一种更加便捷、更加安全，更加隐形化的信息安全技术开始逐渐被人们熟知和使用。

语言是人类所具有的天然属性之一，是人类相互通信获取信息的最方便快捷的手段之一，每个人说话都具有各自的生物特征，比如，发音器官的先天性生理差异，后天形成的个人发音习惯等行为差异，因此两个人语音是不可能完全相同的；同时，声纹识别对硬件设备的成本要求不高，对声音的采集只需要一台麦克风就可以完成，在对语音数据的处理阶段，包括噪声处理，特征提取，模式匹配等也可以在一定级别的处理芯片上也可完成处理。所以搭载在应用设备上通过分析说话人语音的方法来对说话人进行识别就成为了可能，这就是声纹识别。

声纹识别的前景可以从以下几个领域概述：

（1）. 随着计算机的发展，如果将语音作为一种交互手段在某种情况下取代键盘和鼠标，通过语音命令来控制机器运算与显示，解放人的双手，人与机器之间的交流也会越来越自然、流畅。

（2）. 网络购物已经深入人们的日常生活，网上付款时人们很大程度上依赖于密码或二维码，但是随着各种不同场合的频繁使用，网络安全的隐患越来越明显，如果将声纹识别的概念应用到网络安全中，每次验证系统会随机提供提示文本，能够更有效防止复制和剽窃。

1.2. 声纹识别的发展历史与现状

上个世纪 30 年代，伴随着信息技术和计算机技术的发展，通过仪器可以实现说话人信息的识别。“声纹”的概念最早由 Bell 实验室的 L.G.Kesta 在研究声谱时提出的。人们的研究中心从听音识别和人耳的听辨实验转移到提取有利的声

纹特征上来。

上个世纪 40 年代至 70 年代是声纹识别技术的创新阶段，Bell 实验室的 S.Pruzansky 提出了基于模式匹配和概率统计方差分析的说话人识别方法，实现了人耳分辨到自动识别技术的转变，各国的专家学者开始逐渐关注语音处理技术，相继研究提出了倒谱分析技术与线性预测分析技术。

上个世纪 70 年代至 80 年代，声纹识别技术的研究重点则在于对声音中个性特征参数的非线性或线性处理技术以及寻找新的更有效的模型匹配方法。例如如今所被人熟知的动态时间规划、神经网络、支持向量机、隐马尔可夫模型等方法。

上个世纪 90 年代至今，各种模式匹配方法逐步成型，高斯混合模型技术作为声纹识别系统较为前沿的方法，收到越来越多的关注，已经成为专家研究的热点。例如国内的讯飞输入法，其快速又准确的识别效率让语音识别技术也逐渐成为街头让人津津乐道的话题。

国内研究声纹识别起步较晚，目前取得较好研究成果的机构中有中科院声学所、中科院自动化所、清华大学、北京大学、中国科学技术大学、上海交通大学等。近年来在实验室环境下的声纹识别技术已发展成熟。在实际应用环境下的声纹识别技术收到越来越多的研究者关注，尤其是在噪声环境下的声纹识别。所以，抗噪性是声纹识别的关键技术问题之一。

在相关知识库搜索声纹识别时，早期人们都是使用隐马尔可夫模型来完成模式匹配的工作，而今年来使用高斯混合模型来实现声纹识别的人越来越多，所以，本课题采用 MFCC+GMM（Mel 倒谱系数+高斯混合模型）的方式来完成。

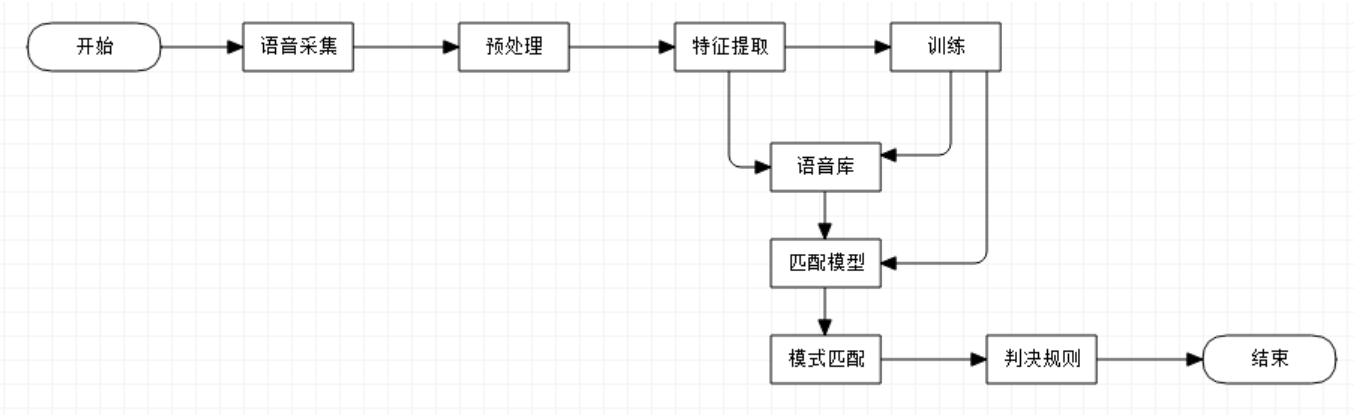
1.3. 声纹识别概述

1.3.1. 声纹识别的基本原理和系统结构

声纹识别从目的和判决方式来分类，可以分为两类：说话人确认和说话人辨认。说话人确认用来确定这段话是不是某一个特定的人说的，属于“一对一”问题，这种方式通常是获取用户输入的语音信号，将语音信号处理后放入每个库存人员的模型中，计算出匹配率值,统计出最大的匹配值，该匹配率值所对应的人员即是对应的识别结果人员。说话人辨认用来判断这段话是众多人中其中哪一个人说的，属于“多选一”问题，这种方式通常是先将标准登陆用户的语音放入标

准登陆用户的模型中，计算出概率绝对值，将此数值作为标准值，为其创建高门限和低门限。获取用户输入的语音信号，将语音信号处理后放入标准登陆用户的模型中，计算出对象贬值，如果新贬值在高门限与低门限范围内，则确定其为目标用户。。根据实际情况，不同的工作和任务会采用不同的声纹识别技术。例如网络交易时，需要采用确认（一对一）的技术；刑侦侦察过程中，为了缩小范围，需要采用辨认（一对多）的技术。

声纹识别作为语音识别中的一部分，双方存在很大的差异：语音识别通常关注的是说话的人说了些什么，而不关注说话人的声道变化和说话习惯，更不关注说话的人是谁；声纹识别关注的是说话人是谁，而忽略说话人说的是什么话，更注重说话人的说话习惯和声道变化等个性特征。声纹识别的基本原理是利用说话人的语音数据，通过统计学的方法为说话人建立一个能够描述此说话人特征的模型，作为此说话人语音信号特征参数的标准模板，然后和其余待识别者的寄存说话人特征的模型进行对比，如果达到某个标准则认定为某个识别者，从而达到判断说话人身份的目的。（如图片 1-1）



图片 1-1 声纹识别流程图

1.3.2. 声纹识别常用的特诊参数

在说话人识别系统中，原始语音信号预处理后，需要提取其特征参数，即提取一些适合分类的某些信息特征。在声纹识别中样本模型的训练和目标说话人的识别都是依赖于所选的特征参数的特性进行分析比较的。

语音信号的特征参数从各个方面体现了说话人的个性、方言和语义特性。但是由于说话内容的语义特征、说话人的情绪波动以及外界环境的干扰，到现在为止还没有找到在各种情况下都相对稳健的特征参数。所以在声纹识别系统中，存

在语音信号的易变形、训练样本数量的不确定性、外界噪声信号的干扰性等问题，声纹识别系统的特征参数提取方法需要业内人士进一步改良与探索。声纹识别的特征参数基本上可以分为三类：

（1）. 线性预测系数及其派生系数。例如：线性预测系数（LPC）、线性预测倒谱系数（LPCC）、及其组合参数。

（2）. 由语音频谱直接导出的参数。例如：共振峰、梅尔频率倒谱系数（MFCC）、感知线性预测系数（PLP）。

（3）. 混合参数。混合参数是由上述不同的特征参数通过某种合理分析所组成的特征矢量。

不同的特征参数对声纹识别系统的识别率影响不同，决定着系统的性能。一个稳定成熟的特征参数不但可以使声纹识别系统的识别率有所提高，而且可以使声纹识别系统的鲁棒性和稳定性有所增强。

1.3.3. 声纹识别常用的建模方法

在声纹识别系统中，当特征参数提取后，要用合适的模型来表征这些特征参数，使得模型能够代表语音信号的特征。因此，模型的选择应从语音信号的类型、期望的性能、计算量及存储量等方面考虑。目前主要的方法有：

（1）. 概率统计方法

语音信号在短时间内是相对稳定的，通过对这些稳定的特性如基音、共振峰频率等进行数学分析，从而选用概率密度函数和均值、方差等相关统计量对这些特征进行分类判决。概率统计建立的模型是开率了语音信号的统计特性采用某种概率密度函数的一组参数作为语音模型。概率统计建立的模型通常会从各个方面体现出语音信号的统计信息。本课题采用的方法高斯混合模型就属于概率统计方法。

（2）. 动态时间规划方法（DTW）

说话人的语音信号不但具有稳定性，而且还具有时变性。将识别模型与参考模型进行时间对比，根据一定的距离测量得出两个模型间的相似程度。

（3）. 矢量量化方法（VQ）

该方法是把所有人的特有文本编写成码本，在识别的过程中，根据

此码本对测试文本编码，并把判断标准定位量化产生的失真度。

（4）. 支持向量机方法(SVM)

支持向量机方法是近年来提出的一种新型的机器学习方法，已广泛应用在指纹识别、人脸识别等模式识别的研究上。SVM 是在统计学习理论的基础上建立起来的一种模型，该模型能够较好地解决小样本学习问题。

（5）. 融合方法

融合方法是把以上分类方法与不同参数进行组合，这样可以使系统的性能有所提升。目前常用的有识别方法组合以及识别方式的结合。

1.4.门禁系统发展历史与现状

公元之前，人们为了守护重要的财物或祈求婴孩长寿发明了锁，这种通过机械组合而达到门禁的意义。后来才逐渐演变为门锁。但是机械锁无论材料多么坚固，设计多么巧妙，人们也许可以通过一根钢丝就轻而易举的打开锁。

如今二十一世纪，人们开始使用带有红外感应的门禁系统来代替原来的机械锁。这种门禁系统需要每个人手持红外感应的卡片，当走到门禁系统时，将卡片贴向感应器，若感应器识别出卡片内的信号则放行。这种门禁方式避免了机械锁的人为破坏也从一定程度上实现了逐步的信息化。例如可以将门禁系统连入网络，这样就可以统计出进入门禁的时间与次数，这就逐步演变为了现在被人们熟知的门禁打卡机。

随着科技的发展，门禁卡或钥匙在物理世界内容易被人所遗忘或丢失，人们采用密码来代替认证设备。而人工智能锁带来的便利也进入到门禁应用中。

1.5.Windows Hello 概述

Windows Hello 是一种生物特征授权方式，用户可以轻松实时访问自己的 Windows 设备且不用担心因为不设置密码而被人窃取信息。

Windows Hello 通过使用用户的脸部、虹膜或指纹等生物特征来解锁设备，这种技术比传统密码更加安全和快捷。用户就是可以解锁 Windows、应用、数据甚至网站和服务的密钥，而不是使用容易被忘记、被破解或随手写下的一串随机排列的字母或数字。相比使用可被共享的密码，Windows Hello 可以帮助用户在

不使用、用密码的前提下直接安全地为应用、网站和网络授权。这样，用户的个人电脑与服务器上就不会存储任何密码，不给黑客可乘之机。

对于相对外行的使用者来说，Windows Hello 就是一个智能身份认证系统。举例来说，如果用户站在 Windows 设备面前，用户只需要露一下脸，动动手指，它会自动完成识别与身份验证，为你想要的服务进行授权。

2. 需求分析

2.1. 声纹识别门禁系统需求分析概述

门禁系统认证端（下全部称之为 Windows Hello 应用端）使用声纹识别来进行验证，首先需要通过录音设备录下标准用户的语音，系统分析计算好待识别者的特征参数，放入高斯混合模型中训练，得出相应的属于此人的声纹模型。将此声纹模型的数据保存下来。当待识别用户使用本系统时，先计算待识别用户的特征参数，再加载标准用户的声纹模型，将待识别用户的特征参数放入标准用户的声纹模型中计算出概率值。获取概率值后可以按照事先设置好的 1x1 或 1x 多方式进行决断结果。

在 Windows Hello 应用端中需要先计算标准用户的参数模型数据，这需要与用户直接交互，需要单独一个部分（下全部称之为桌面应用）采用图形化的界面录下用户的语音并生成 txt 文件用来保存该用户的声纹模型数据。

Windows Hello 应用端在识别时可根据用户对于系统的配置信息（例如是识别方式是 1x1 还是 1x 多）读取标准用户的声纹模型数据加载声纹模型并开始识别作业。

为了开发者能够对错误原因进行检测与分析，Windows Hello 应用端会在运行时将系统的运行状况与特征参数的数值等发送给相应的服务器，服务器接收到消息后先缓存在系统中，当 Windows Hello 应用端发出结束信息时则将数据写入到数据库中，同时通知前端实时刷新。该部分下全部成为 Web 实时日志监控系统。

2.2. 声纹识别门禁系统需求列表

表格 2-1 需求列表

应用类型	功能分类	功能	备注
桌面应用	录音功能	录音开始	通知设备开始录音
		录音结束	通知设备结束录音，并将缓存数据写入文件
		录音配置	对录音的采样频率采样位数声道数进行设置

表格 2-2 需求列表 续上表

应用类型	功能分类	功能	备注
桌面应用	文件列表功能	寄存语音文件列表	显示库存的语音文件
		寄存模型文件列表	显示库存的 txt 模型文件
	语音模型训练功能	训练功能	训练模型
		写入文件功能	将模型数据写入到 txt 文件
Windows Hello 应用	读取配置文件功能	读取配置文件功能	读取对于系统设置的相关配置
	特征参数提取功能	特征参数提取功能	计算语音中的相关特征参数
	语音模型训练建模功能	语音模型训练建模功能	语音模型训练建模功能
	识别结果决断功能	识别结果决断功能	对识别的结果按照设置的规则进行决断
	Socket 客户端发送功能	Socket 客户端发送功能	将数据信息发送给服务器
	本地日志功能	本地日志功能	将日志信息输出写入到文件
Web 实时日志监控系统	后台监听功能	监听功能	后台监听 Socket 是否被接入
		多线程功能	如果有 Socket 接入则单独开启一个线程用户接收信息
	Socket 服务器接收功能	Socket 服务器接收功能	接收来自客户端发过来的数据
	数据库写入功能	数据库写入功能	将客户端发过来的数据保存进入数据库

3. 开发工具和开发技术

3.1. 开发环境

操作系统：Windows 10 Professional 1709

集成开发环境：Visual Studio 2010 Professionnal

集成开发环境：IntelliJ IDEA 2017.3 Ultimate

JDK 版本：Java SE Development Kit 8.0.144

Web 服务器：Apache Tomcat Server 8.36

数据库服务器：Mysql 5.6

3.2. 开发技术

3.2.1. 声纹识别技术

声纹识别相比其他生物特征识别相比，使用声纹识别具有如下优势

（1）. 声纹识别能够在用户不知不觉中就能够做完身份验证，自然得获取说话人的语音，不会涉及到隐私问题，容易被使用者接受。

（2）. 认证方式简单，只需要用户提供足够的录音数据即可，不用像指纹或虹膜识别技术需要将人体器官贴近信息采集仪器。

（3）. 实现成本非常低廉，相比于各种指纹或虹膜认证技术需要单独购买特殊昂贵的扫描仪器设备，声纹识别只需要简单的语音输入设备，例如一台电脑就可以实现从采集到识别结果的所有过程。

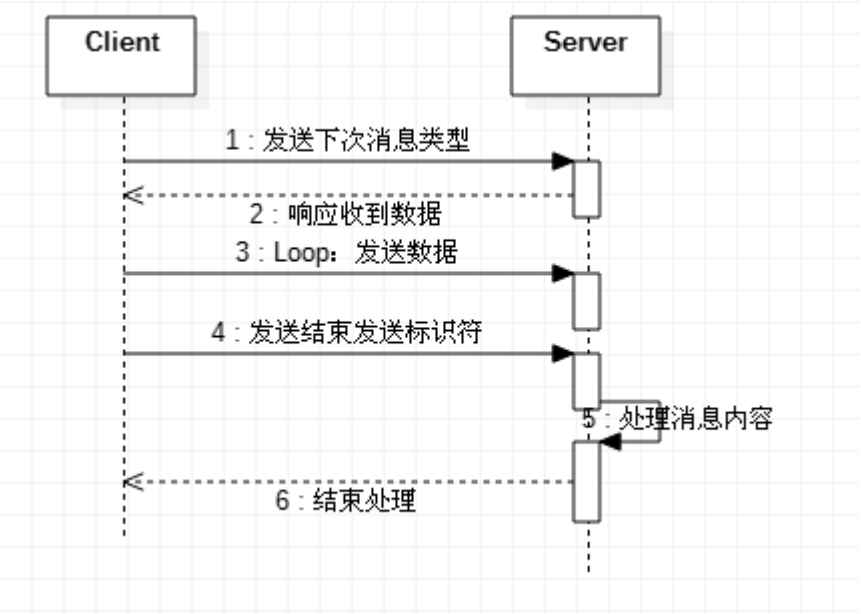
（4）. 在远距离的身份识别中。声纹识别更快捷，通过电话或手机也可实现远距离身份确认。

本课题中采用提取用户语音的 MFCC 参数放入高斯混合模型中训练，通过分析结果数据来实现声纹识别功能。

3.2.2. Socket 传输技术

Socket 又称"套接字"，应用程序通常通过"套接字"向网络发出请求或者应答网络请求。建立网络通信连接至少要一对端口号。Socket 本质是编程接口(API)，对 TCP/IP 的封装，TCP/IP 也要提供可供程序员做网络开发所用的接口，这就是 Socket 编程接口；

利用 Socket 可以在不同的机器、不同的平台间可靠地传递消息。本课题利用 Socket 在 Windows 客户端中将识别的数据集、识别的结果、识别的流程日志和识别的对象语音发送给服务器。从而实现记录日志，监控系统，排除系统异常



图片 3-1 Socket 传输数据集流程

情况的功能。

3.2.3. 后台刷新

本课题存在的两个部分，Windows 客户端与 web 实时日志监控系统，Windows 客户端将识别的结果发送 web 实时日志监控系统时，日志管理系统后台需要将信息实时的反映在 Web 前端界面当中。这需要客户端浏览器与 Web 服务器后台实时连接通信，当 Web 服务器后台接收到来自 Windows 客户端发送过来的数据时，从 Web 服务器后台发送群体通知消息给每个 Web 客户端浏览器，Web 客户端接收到消息后解析消息内容，并对前端页面进行实时刷新。这就需要 Web 服务器与 Web 客户端使用 WebSocket，双方实时连接，当需要通信时向对方发送消息，基础原理同 Socket 传输技术。

3.2.4. Struts1.3x 框架

Struts1.3x 框架是 JavaWeb 开发的一种标准化框架。在 Struts 框架还未面世之前，JavaWeb 都是采用手动 MVC 的结构来编写的。但是由于 MVC 的开放性，各个公司开发出来的功能结构并不统一，一旦项目规模扩大，新人很难着手，维护起来困难。

这个时候 Struts 出现了，Struts 封装了底层的 Servlet，用户只需要编写上层的接口事件，例如用户需要做某个事情，那么开发者只需要关心做这件事情的事件处理类是哪一个，其所用到的参数类是哪一个。Struts 规范了业内的代码结构，增加了程序的可读性，让新人也能更快上手维护系统。

3.2.5. 多线程技术

在 Web 实时日志监控系统中，使用 Socket 接收消息时，Socket 的逻辑定义是服务器类型的，即可以存在多个 Windows Hello 向 Web 实时日志监控系统发送消息。那么 Web 实时日志监控系统的后台无法同时监管 Web 服务功能与后台的数据接收功能，这就需要多线程技术。

多线程程序能够通过计算机底层硬件的支持从而实现能够在同一时间运行多个线程来处理程序，从而增加了系统资源的利用率。

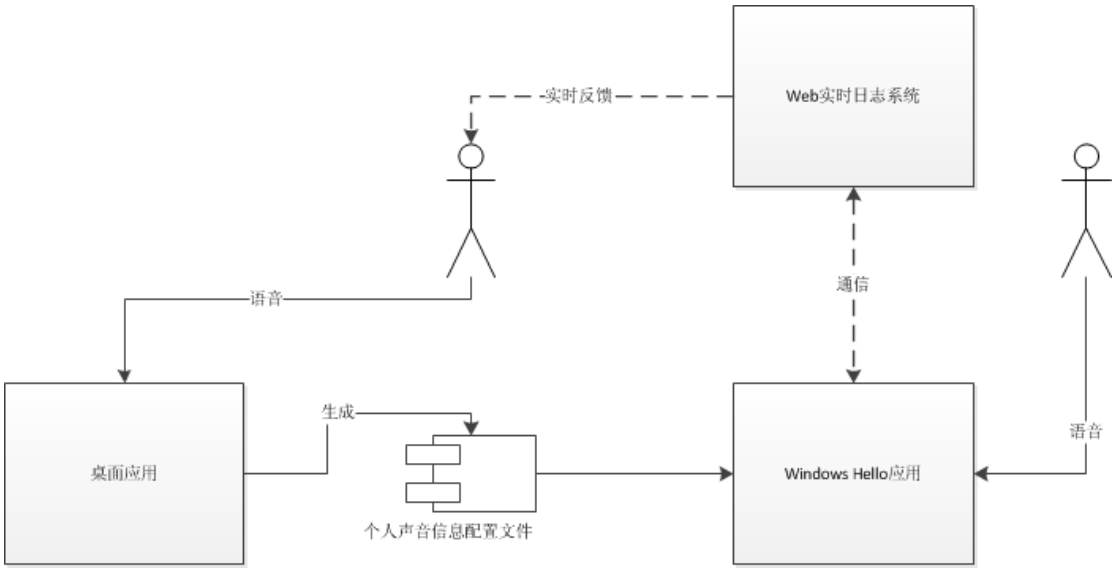
在本课题中，如果服务器后台监听到了有客户端接入 Socket 服务器，则单独创建一个线程，传入 Socket 链接句柄，在线程内进行数据的接收与保存。当客户端断开连接后，则关闭线程并清除系统资源。

3.2.6. MFC 技术

MFC 是微软提供一套图形库基础类,他以 C++语言封装了 Windows 的窗口 API 与各种控件，开发者可以在 Visual Studio 中轻松拖取页面控件布局并为控件动作编码。极大减轻了开发者的负担，使得即使一个人也能够开发出能够使用的 MFC 程序。在本课题中，我使用 MFC 制作桌面应用端，通过桌面应用端录音并生成标准用户模型参数。

4. 系统设计

4.1. 系统配置设计



图片 4-1 系统部署配置图

本课题的部署配置如图片 4-1 所示，本课题分为三大部分：桌面应用、Window Hello 应用、web 实时日志系统。

- 桌面应用：桌面应用用于测试用户设备当前是否存在合适的录音设备，并录下标准用户的语音，计算好声纹信息文件提供给后期识别认证时使用。
- Windows Hello 应用：Windows Hello 应用通过 Windows 的登陆模式，模拟门禁系统开放与闭合的情况。
- Web 实时日志系统：Web 实时日志系统用于监控当前认证设备的使用情况，防止错误识别情况的发生。

如图 4-1：桌面应用接受用户输入的语音进行计算，生产个人声音信息配置文件，Windows Hello 应用获取到标准用户的信息文件与陌生用户的语音生产的信息配置文件数据信息进行对比，从而判断是否该登陆至目标系统。同时，Windows Hello 应用与 web 实时日志系统进行通信，发送认证与匹配信息实时反馈给标准用户。

4.2. 详细设计

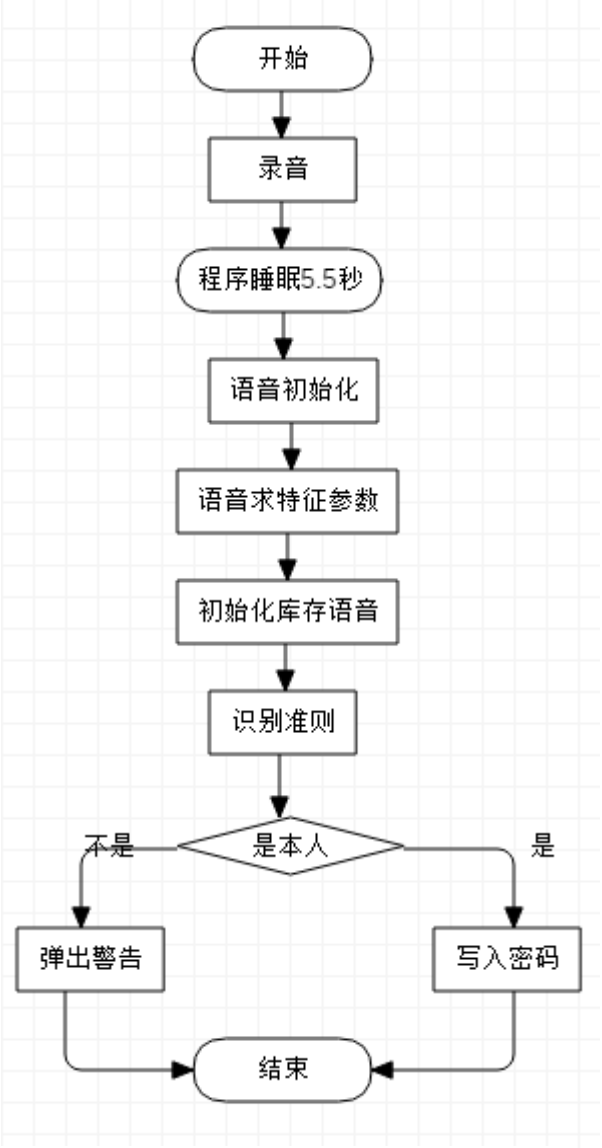
4.2.1. Windows 下桌面应用系统设计

表格 4-1 应用系统模块意义

模块分类	函数文件	功能意义
Log	LogSystem	日志系统的控制
	MessageQueue	消息队列的数据结构
	ReadConfig	配置信息的读取
	SocketClient	消息的发送与结构
VPR	WavFile_Struct	语音文件结构
	WavFile_Initial	语音文件初始化
	Model_GMM	运算模型调用
	Model_KMeans	运算模型调用
	WavData_CharaParameter	语音参数获取
	WavData_SupportFunction	共通取得
UI	CChineseCode	编码形式变换
	Shockwaveflash	flash 驱动
	Voiceprint RecognitionDlg Response	共通取得
	Voiceprint RecognitionDlg	界面函数
	WaveRecorder	录音操作

桌面应用系统使用 MFC 来开发，相关控件和意义可见第八章。

4.2.2. Windows Hello 应用设计



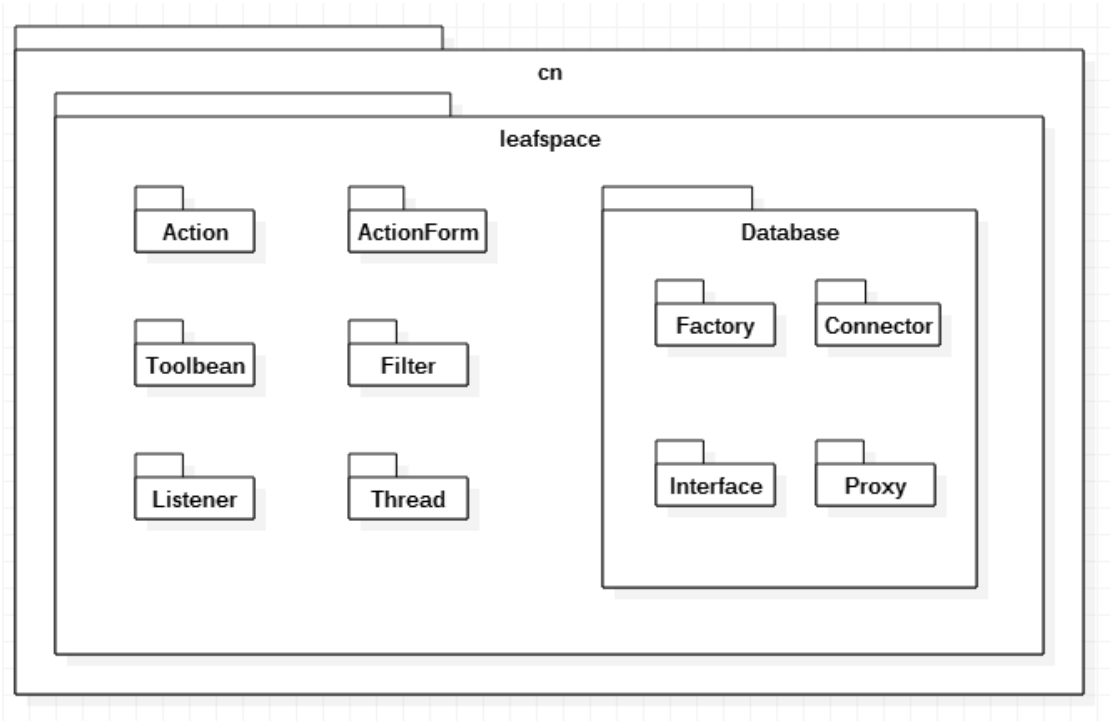
图片 4-2 Windows Hello 应用流程

4.2.3. Web 实时监控日志系统设计

表格 4-2 页面列表

页面名	页面功能
login,jsp	登陆页面
index.jsp	动态页面
table_complete	详细数据页面
form_validate	批量删除页面

Web 系统使用 struts1.3 的框架，功能只有查看与删除用。包图如图 4-3。



图片 4-3 web 日志系统包图

infolist

Fields

Field	Type	Collation	Null	Key	Default	Extra	Privileges	Comment
ID	int(11)	(NULL)	NO	PRI	(NULL)	auto_increment	select,insert,update,references	
InfoType	bit(1)	(NULL)	NO		(NULL)		select,insert,update,references	
ClientType	varchar(255)	utf8_general_ci	NO		(NULL)		select,insert,update,references	
IssueTime	datetime	(NULL)	NO		(NULL)		select,insert,update,references	
Information	varchar(500)	utf8_general_ci	NO		(NULL)		select,insert,update,references	
Result	varchar(900)	utf8_general_ci	YES		(NULL)		select,insert,update,references	
ClientIP	varchar(36)	utf8_general_ci	YES		(NULL)		select,insert,update,references	
FilePath	varchar(255)	utf8_general_ci	YES		(NULL)		select,insert,update,references	

Indexes

Table	Non unique	Key name	Seq in index	Column name	Collation	Cardinality	Sub part	Packed	Null	Index type	Comment	Index comment
infolist	0	PRIMARY	1	ID	A	0	(NULL)	(NULL)		BTREE		

Back

userlist

Fields

Field	Type	Collation	Null	Key	Default	Extra	Privileges	Comment
ID	int(11)	(NULL)	NO	PRI	(NULL)	auto_increment	select,insert,update,references	
Username	varchar(255)	utf8_general_ci	NO		(NULL)		select,insert,update,references	
Password	varchar(255)	utf8_general_ci	NO		(NULL)		select,insert,update,references	

Indexes

Table	Non unique	Key name	Seq in index	Column name	Collation	Cardinality	Sub part	Packed	Null	Index type	Comment	Index comment
userlist	0	PRIMARY	1	ID	A	1	(NULL)	(NULL)		BTREE		

图片 4-4 web 日志系统数据库结构

5. 系统实现

5.1.语音信号预处理

5.1.1. Windows 下的语音信号采集

在 Windows 下可以使用系统自带的 API 来完成录音操作；使用系统自带的 API 有两种选择方式，一种是 Windows Multimedia API(MMAPI)，这种方式是利用 WAV 录音固定的格式，先设置好音频测试设备的硬件参数，在开辟一个缓冲区，由录音设备按照设置的参数定次数定位长向缓冲区写入录音数据，再由用户定时读取缓冲区的数据，最后将读取的数据存入文件；还有一种是 Media Control Interface（MCI），使用这种方式，用户可以简单的使用字符串命令实现录音和放音。本课题中采用 MMAPI 的方法来获取语音文件。

用 MMAPI 可以实时获得音频数据，MMAPI 可以把音频缓冲起来并逐块发送给使用者，这种固定大小的音频裸流数据简称为 AudioFrame。当用户预备开始录音时，MMAPI 以一种格式打开波形输入设备，发送设备开启消息给回调函数并准备缓冲区，将缓冲区添加到设备，告诉录音设备开始录音。当使用者需要结束录音时，MMAPI 告知设备录音结束并发送消息给回调函数处理最后的数据，然后释放缓冲区，最后关闭设备。

5.1.2. 语音数据提取

RIFF 块	12	WAVE Chunk		
	标识符	4	字符	"RIFF"
	文件长度	4	长整形	文件的总字节数
	WAV 标志	4	字符	"WAVE"
格式块	24 或 26	Format Chunk		
	标识符	4	字符	"fmt"
	格式块长度	4	长整形	
	格式类别	2	整形	值＝1 表示编码方式为 PCMμ律编码
	声道数	2	整形	单声道＝1，双声音＝2
	采样频率	4	长整形	每秒的样本数
	数据传送速率	4	长整形	每秒传送的字节数
	样本字节数	2	整形	每个又称基准块＝每个样本位数×声道数÷8 样本的字节数
	样本位数	2	整形	量化位数
	附加信息	2	整形	通过块长度来判断有无
附加块	12	Fact Chunk		
	标识符	4	字符	"fact"
	块长度	4	长整形	
		4		
数据块	不定	Data Chunk		
	标识符	4	字符	
	数据长度	4	长整形	
		不定		

图片 5-1 WAV 文件数据结构

每一个语音文件（此处专指 wav 格式）都必定有四个模块（如图 5-1），它们分别为 RIFF 块、格式块、附加块、数据块，其中 RIFF、格式块和附加块的总长度必定是 24 或 26 字节，数据块长度不定。RIFF 块有如下内容：标识符、文件长度、WAV 标志；格式块有如下内容：标识符、格式块长度、格式类别、声道数、采样频率、数据传送速率、样本字节数、样本位数、附加信息，其中附加信息的有无是通过格式块长度来确定的，如果格式块长度为 24 则附加信息不存在，如果格式块长度为 26 则存在附加信息；附加块有如下内容：标识符、块长度、补丁哈希值；数据块有如下内容：标识符、数据长度、数据内容。我们在这里主要要用到数据传送速率、样本字节数、样本位数、数据长度跟数据五个部分的内容。

首先按照文件格式读取文件中我们所需要的内容，读出的数据是以字节的方式存放的，此时的数据并不一定是语音处理的数据，我们需要根据采样位数合并字节，重新换算成语音信号的数据。此处我们实验的数据都是采样位数为 16bit 的，下面我们就说明以 16bit 的数据对原字节进行变换的算法:

采样位数为 16bit，也就是说其由两个单字节的数据组成，这个两个字节分

为高字节数据与低字节数据。

如果高字节的数据大于 0，则判断低字节的数据，如果低字节的数据大于等于 0，则新生成的数为高字节的数乘 256 加上低字节的数，$C=256A+B$；如果低字节的数小于 0，则新生成的数为高字节的数乘 256 加上那个低字节的数的绝对值加上 128，$C=256A+|B|+128$；

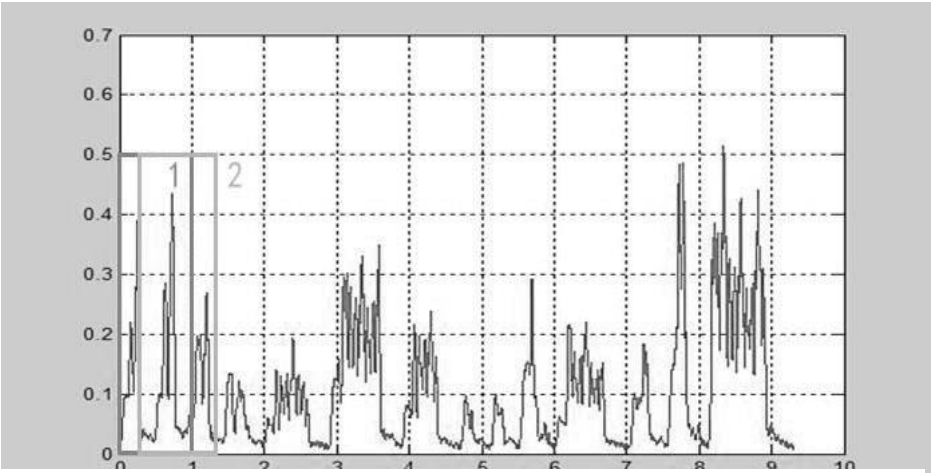
如果高字节的数据小于 0，则判断低字节的数据，如果低字节的数据大于 0，则新生成的数为高字节的数的绝对值乘 256 加上低字节的数的负数，$C=-256|A|-B$；如果低字节的数小于等于 0，则新生成的数为高字节的数的绝对值乘 256 加上低字节的数的绝对值加上 128 的负数，$C=-256|A|-|B|-128$；

注：A 表示高字节数，B 表示低字节数，C 表示生成的数

此时的数据就是语音信号可参与转换的数，其数值的范围$-2^{采样位数-1}\sim 2^{采样位数-1}-1$此时的数据已经可以参与运算了，但是从实验的角度上，相对过大的数据将会给计算机带来相对较大的负荷，所以此时我们对数值进行“归一化”，将每个数据除以$-2^{采样位数-1}$，使得最后的数据范围在-1~1 之间。

5.1.3. 分帧与加窗

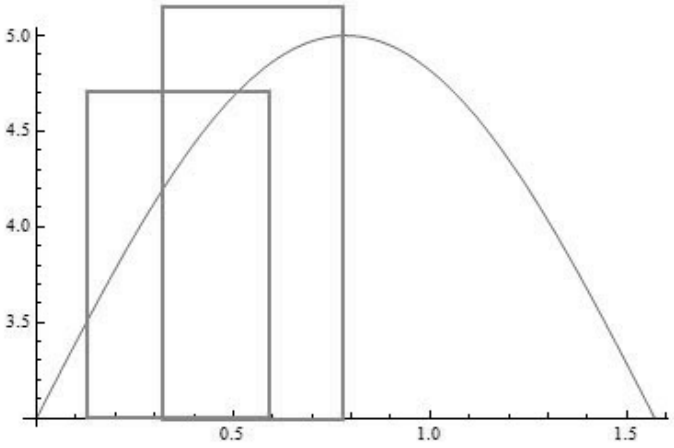
语音信号是随时间而变化的，是一个非平稳状态变化过程，但是在短时间范围内（一般在 10-30ms 内），其某些特性基本保持不变，即在短时内成相对稳



图片 5-2 宏观分帧示意图

定状态，在学术上这被称之为语音信号的短时平稳性。

任何语音信号的分析和处理必须建立在“短时”的基础上，所以在处理时需要进行分帧处理。分帧是通过可移动的有限长度窗函数进行加权的方法实现的。如图 5-2 示意图，图中为语音信号分了两个窗，窗一与窗二，窗一的开始与窗二的开始被称为窗移，若将此图放大若干倍，窗与窗之间数值的变化不大，然而却能够体现语音信号的特征。如图 5-3，就是放大后的图。



图片 5-3 微观分帧示意图

分帧通常每帧取 10~40ms 的数据，其中人们常使用 25ms 作为标准配置，但考虑到后续工作需要对数据进行快速傅里叶变换，当中需要每帧的数据个数为2^n，所以，本课题采用的帧长为 256、帧移为 125。我们通常使用汉明窗作为加窗函数，汉明窗中的数据能够为傅里叶变换提供便利减少数据量，汉明窗函数：

$$w(n)=\begin{cases}0.54-0.46*\cos\left(\dfrac{2\pi n}{N-1}\right),0\le n\le(N-1)\\0,n=\text{其他值}\end{cases} \quad (2\text{-}1)$$

其中 n 表示了当前数据在窗中处于第 n 个数据，N 表示汉明窗的窗长度。

5.1.4. 预加重

研究表明，语音信号的平均功率谱是受到声门激励与口鼻辐射的影响，高频率约在 800Hz 以上按照 6dB/倍频程跌落，导致语音信号低频段能量大，高频段信号能量明显小。而鉴频器输出噪声的功率谱密度则随频率的平方而增加（低频噪声大，高频噪声小），造成信号的低频信噪比很大，而高频信噪比明显不足。从而导致高频传输衰弱，使高频传输困难。

因此对语音信号进行预加重，把信号的高频部分进行加重，去除口唇辐射的影响，增加语音的高频分辨率，提高信噪比，改善信号的传输质量。

预加重使用具有 6dB/倍频程的提升高频特性的预加重数字滤波器来实现，它一般是一阶的数字滤波器：

$$\mathrm{H}(z) = 1 - \mu z^{-1} \quad (2\text{-}2)$$

其中μ为预加重系数，在相关研究文献实验中通常取$(0.9，\ 1.0)$。经过预加重处理后的结果为：

$$\mathrm{y}(n) = x(n) - \mu x(n-1) \quad (2\text{-}3)$$

5.1.5. 端点检测

端点检测是声纹识别中初始化阶段的主要进程，端点检测是从一段语音中判断出语音的前后两个端点。声纹识别系统中使用端点检测，一方面可以减少数据量，节约不必要的时间，另一方面可以减少噪声信号的干扰，使声纹识别更准确。

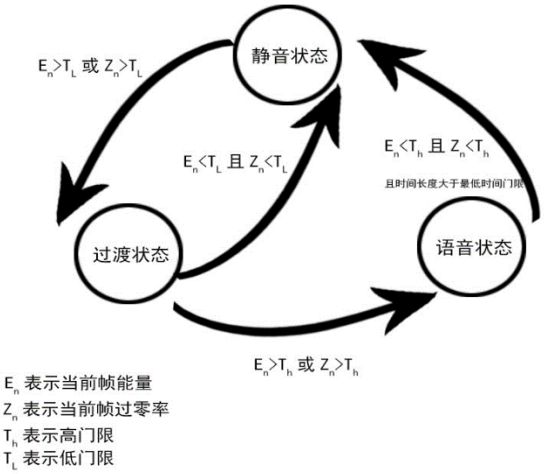
在语音处理的过程中，尤其是当针对孤立多个单词进行识别的情况下，正确的判断词语的起末端点为模型匹配和提升识别率起着不可替代的作用，既可以将信号处理时间削弱至最小，又可以清理掉无声段的噪声影响，最终提高系统的识别性能。在一定程度上端点检测的准确度直接影响着整个语音处理系统的性能。

双门限端点检测法是音频信号处理当中常用的检测手段，其需要从语音信号中提取出时域范围内的两个特征参数（计算方法见第五章：语音信号特征参数提取）：短时过零率和短时帧能量。双门限端点检测法其结合了短时过零率和短时能量的优点，因此对语音信号的起点和判决更加有效。（实验证明双门限端点检测法只在信噪比较高的情况下具有较高的鲁棒性）

双门限法就是在检测的过程中为短时帧能量和短时过零率均设置两个门限，一个高门限和一个低门限，低门限往往对信号的变化反应比较敏感，而当信号达到一定强度才能超过高门限。具体做法如下：

（1）. 计算并获得最大短时能量、最小短时能量、最大短时平均过零率和最小短时平均过零率就是语音信号的波动范围，下面依据上述的四个范围分别为短时帧能量与短时平均过零率设置双门限（分为高门限与低门限），门限的数值将直接影响到最后端点检测判断的结果。在这里，我们先将高门限设置成语音信号频率强度最高范围的 1/4，将低门限设置成语音信号最高频率强度范围的 1/8。还要设置一个最短语音长度，用于判断是否为爆破音。转到步骤 2。

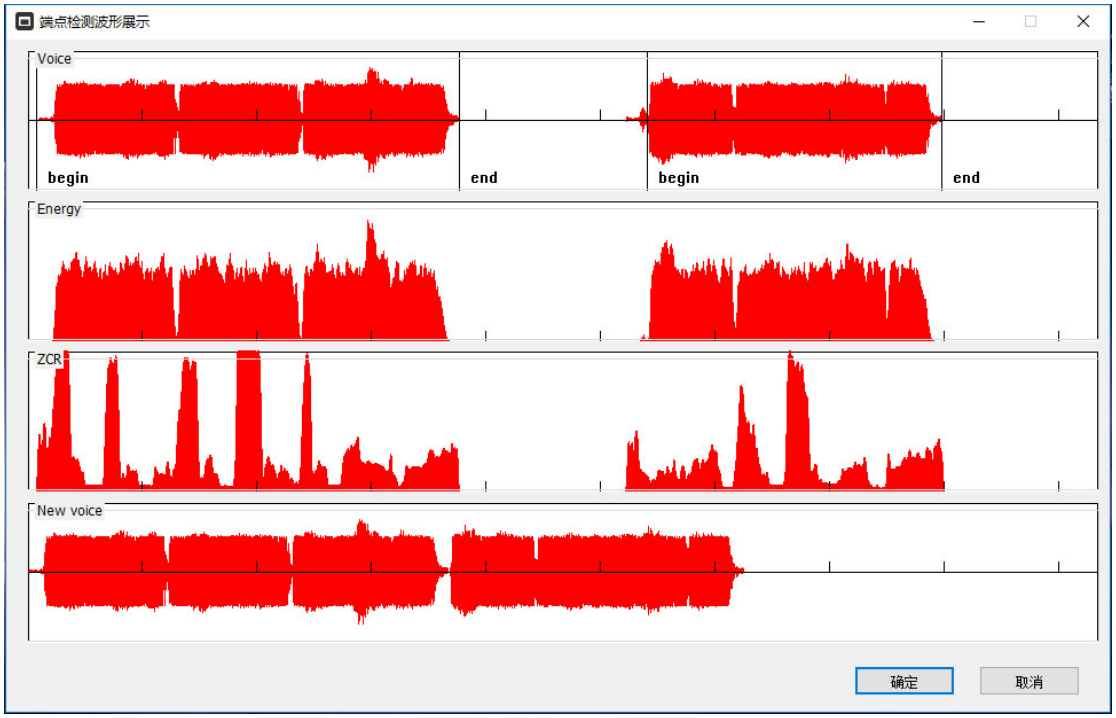
（2）. 按照窗的个数同时遍历短时帧能量与短时平均过零率，初始状态为静音状态（语音的三种状态，分别为“静音状态”、“过渡状态”、“语音状态”）。（转换图如图 5-4）转到步骤 3。



图片 5-4 语音状态转换示意图

（3）. 此时语音处于静音状态，如果当前帧的短时帧能量高于所设定的短时帧低门限或者当前帧的短时平均过零率高于所设定的短时帧平均低过零率，则标记当前为语音的起始段转到步骤过渡状态。转到步骤 4

（4）. 此时语音处于过渡状态，如果当前帧的短时帧能量小于所设定的短



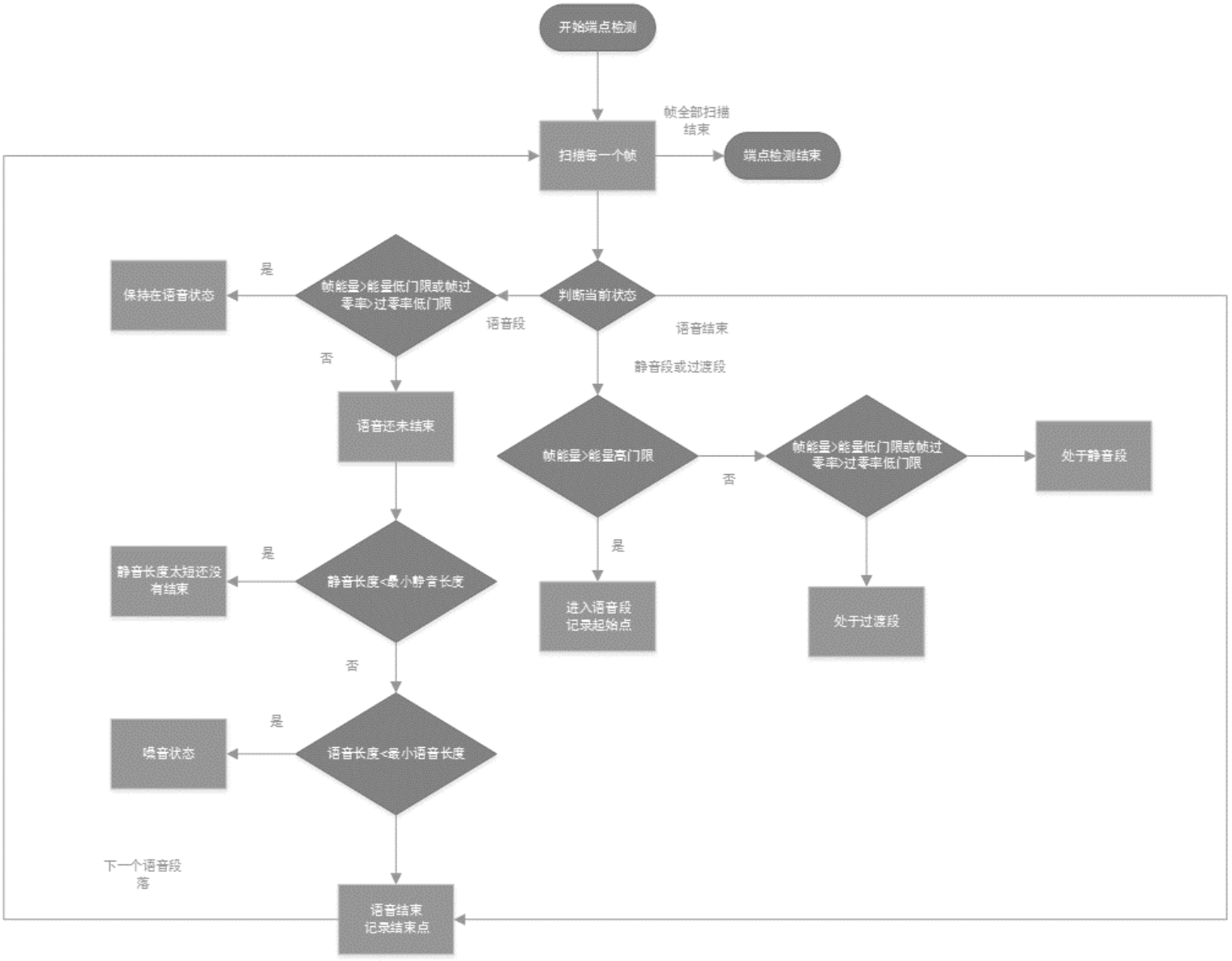
图片 5-5 一段语音的检测效果

时帧低门限且当前帧的短时平均过零率小于所设定的短时帧平均低过

零率，则状态不变（静音状态）。如果当前帧的短时帧能量高于所设定的短时帧高门限或者当前帧的短时平均过零率高于所设定的短时帧平均高过零率。则标识语音转到步骤语音状态。转到步骤 5

（5）. 此时语音处于语音状态，如果当前帧的短时帧能量小于所设定的短时帧低门限且当前帧的短时平均过零率小于所设定的短时帧平均低过零率切当前的时间长度小于最短时间门限则标识为噪音，否则标识为结束点，然后记录保存下起始点，结束点信息。转到步骤 2 直至遍历结束。

双门限端点检测流程图如图 5-6.



图片 5-6 双门限端点检测法流程

5.2.语音信号特征参数提取

5.2.1. 短时帧能量和短时过零率

语音信号的能量不是一直不变的，并且各个音之间的能量存在差异，所以可以由此判断出语音信号的个性特点。

短时过零率是指每帧内语音信号通过横轴的次数。在时域范围下横轴上可以看到连续语音信号的波形。若离散语音信号相邻的采样存在不同的代数符号，将此现象成为发生了过零，因此可以计算过零的次数。

用E_n代表第 n 帧语音信号$x(m)$的短时能量：

$$E_n = \sum_{m=n-N+1}^{n} [x(m)w(n-m)]^2 \quad (5\text{-}1)$$

用$\mathrm{Z}(n)$代表第 n 帧的过零率：

$$Z_n = \frac{1}{2} * \sum_{m=-\infty}^{\infty} |\mathrm{sgn}[x(m)] - \mathrm{sgn}[x(m-1)]| w(n-m) \quad (5\text{-}2)$$

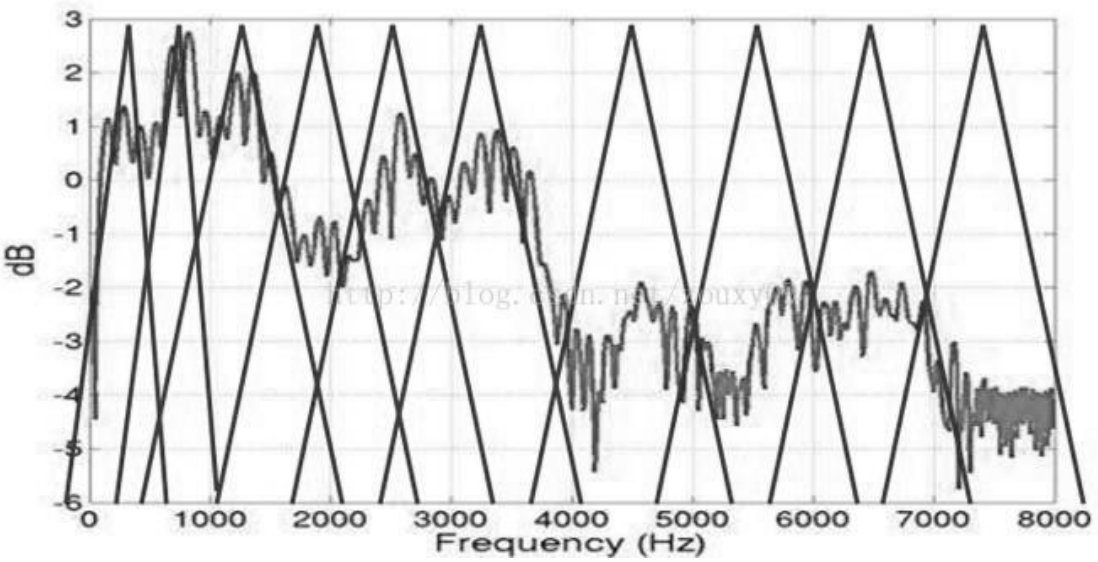
其中 sgn[]为符号函数，即

$$\mathrm{sgn}[\mathrm{x}] = \begin{cases} 1, x \geq 0 \\ 0, x < 0 \end{cases} \quad (5\text{-}3)$$

$\mathrm{w}(n)$为窗函数（见 2.3 分帧与加窗）。

5.2.2. Mel 频率倒谱系数（MFCC）

人的听觉系统对声音频率的感知是非线性的，人耳相当于一个滤波器组，它能让人在各种变异环境下正常分辨声音，其滤波作用在对数频率尺度上进行，对

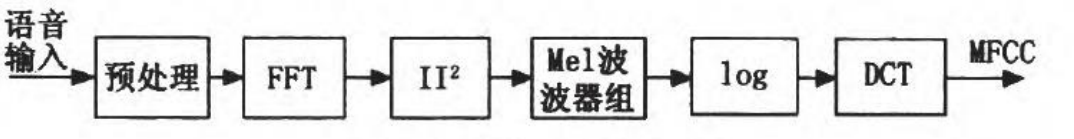


图片 5-7 Mel 频率倒谱系数示意图

1000Hz 以下的频率声音的感知呈近似线性关系；而对于 1000Hz 以上频率声音的感知遵循在对数频率坐标上的近似线性关系，使人耳对低频信号更敏感。

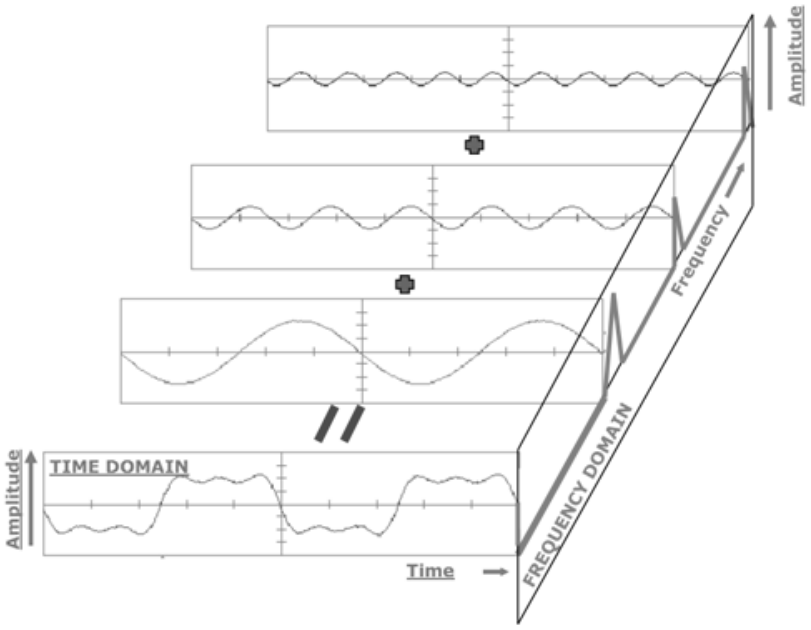
Mel 频率滤波器组就是由此实验得到的，其用于模拟人耳对不同频率语音的感知。Mel 频率倒谱系数是用一个在低频区域交叉重叠的三角形滤波器组-Mel 滤波器组对语音信号的能量谱进行滤波。Mel 滤波器组在信号的低频区域分布较密，中频区域分布较少，高频区域分布较为稀疏，单个滤波器的带通带宽较大，因此，高频区域的频率分辨率较低，频谱信息较弱，导致信息遗漏。

Mel 频率倒谱系数的提取过程如下（见图 14）：



图片 5-8 Mel 频率倒谱系数提取过程

（1）. 将每帧的数据进行快速傅里叶变换（FFT），将数据从时域转换为频



图片 5-9 傅里叶变换原理图

域。有些信号在时域上是很难看出什么特征的，但是如果变换到频域之后，就很容易看出特征了。这就是很多信号分析采用快速傅里叶变换变换的原因。

另外，快速傅里叶变换可以将一个信号的频谱提取出来，这在频谱分析方面也是经常用的。在此处就是利用快速傅里叶变换将信号的频谱提取

出来。

$X_K = \sum_{n=0}^{N-1} X_n \cdot e^{-i2\pi kn/N}$　　离散傅里叶变换（DFT）　　（5-4）

$X_n = \frac{1}{N}\sum_{k=0}^{N-1} X_k \cdot e^{i2\pi kn/N}$　　逆向离散傅里叶变换（IDFT）　　（5-5）

（2）. 求频谱的平方，得到谱线的能量：

$$\mathrm{E}(i \quad , \ k) = [x(i,k)]^2 \tag{5-6}$$

其中 i 表示第 i 帧，k 表示第 k 条谱线

（3）. 将谱线能量通过 Mel 滤波器组

1) 求最大/最小 Mel 频率，及滤波器的中心间距

$$\mathrm{melFremax} = 1125 \cdot \log\left(1 + \frac{SampleRate}{2 \cdot 700}\right) \tag{5-7}$$

$$\mathrm{melFremin} = 1125 \cdot \log\left(1 + \frac{0}{700}\right) \tag{5-8}$$

$$\Delta_{mel} = (melFremax - melFremin)/(filterNum + 1) \tag{5-9}$$

其中 melFremax 表示的是最大 Mel 频率，melFremin 表示的是最小 Mel 频率，SampleRate 表示的是信号的采样频率，filterNum 表示的是帧的个数。

2) 求 mel 滤波器组中每个滤波器的实际频率位

$$f[i] = floor\left(\frac{(N+1)\cdot 700 \cdot \left(exp\left(\frac{melFremin + \Delta_{mel} \cdot i}{1125}\right) - 1\right)}{SimpleRate}\right) \tag{5-10}$$

I 为当前滤波器　　melFremin 为最小 Mel 频率

N 为帧长　　Δ_{mel}为滤波器的中心间距

SimpleRate 为采样频率

3) 将谱线能量通过滤波器

$$\mathrm{S}(i \quad , \ m) = \sum_{k=0}^{N-1} E(i,k)\, H_m(k), \ 0 \le \mathrm{m} < \mathrm{M} \tag{5-11}$$

$$H_m(k) \begin{cases} \frac{k - f(m-1)}{f(m) - f(m-1)} & f(m-1) \le k \le f(m) \\ \frac{f(m+1) - k}{f(m+1) - f(m)} & f(m) \le k \le f(m+1) \\ 0 \quad \text{其他} & \end{cases} \tag{5-12}$$

I 表示第 i 帧　　m 表示第 m 个滤波器

N 表示帧长　　k 表示帧内第 k 个数据

M 表示滤波器的总个数

（4）. 求对数

$$\mathrm{mel}(i，m)=\ln\left(mel(i，m)\right) \quad (5\text{-}13)$$

（5）. 进行离散余弦变换

$$\mathrm{C}(i,n)=\sum_{m=0}^{M-1}S(i,m)\cos\frac{\pi n(m+0.5)}{M} \quad 0\le \mathrm{n}\le \mathrm{L}\ 且\ 0\le \mathrm{m}<\mathrm{M} \quad (5\text{-}14)$$

其中 i 为帧数，n 为阶数　　表示第 i 帧中的第 n 阶

M 为第 m 个滤波器　　M 为总滤波器的个数

$\mathrm{S}(i,m)$为[i][m]位置上的滤波能量

5.3.语音信号训练模型

5.3.1. 高斯混合模型(GMM)概述

在声纹识别系统中，模型的建立至关重要，不同的模型建立方法对识别的性能影响不同。在进行声纹识别时，将输入系统的待识别人的语音特征参数与对比说话人识别模型进行模型匹配度相似比分析，然后根据对比分析结果按照用户设置的辨认贬值对待识别的说话人身份做出相应的判断。因此，声纹识别模型的建立与特征参数提取同等重要，也是声纹识别的关键之一。传统的建模方法计算量较大，训练参数模型的时间较长，缺乏统计特性。但是高斯混合模型能较好地描述样本空间的状态，能够对任意形状的概率分布进行近似模拟。

近年来的研究调查发现，说话人的特征分布并没有严格遵循某一个特定分布，例如高斯分布等；但是几乎所有的分布都能够使用高斯分布的混合权值来近似接近，从而获得了高斯混合模型。[10]

基于高斯混合模型的声纹识别技术的基本理论为针对训练话者集合内的每一个说话人构建属于自己身份特征的概率分布模型，这种概率模型中所设计的参数值是由说话人自身的特征参数分布情况所决定，所以可以用以描述说话人的身份特性。[13]

高斯混合模型种类有单高斯模型（SGM）和高斯混合模型（GMM）两类。类似于聚类，根据高斯概率密度函数（PDF）参数不同，每一个高斯模型可以看作一种类别，输入一个样本 x，即可通过 PDF 计算其值，然后通过一个阈值来判断该样本是否属于高斯模型。

所以 SGM 适合于仅有两类别问题的划分，而 GMM 由于具有多个模型，划分更为精细，适用于多类别的划分，可以应用于复杂的语音信号概率建模。

5.3.2. 单高斯模型

单高斯模型的多维高斯（正态）分布概率密度函数定义如下：

$$N(x;\mu,\sum)=\frac{1}{\sqrt{(2\pi)|\sum|}}exp\left[-\frac{1}{2}(x-\mu)^{T}\sum^{-1}(x-\mu)\right] \quad (5\text{-}15)$$

X 是维数为 d 的样本向量（列向量），$\boldsymbol{\mu}$是模型期望，∑是模型方差。

假设训练样本属于类别 C，则上式变为：

$$N(x/C)=\frac{1}{\sqrt{(2\pi)|\sum|}}exp\left[-\frac{1}{2}(x-\mu)^T\sum^{-1}(x-\mu)\right] \quad （5\text{-}16）$$

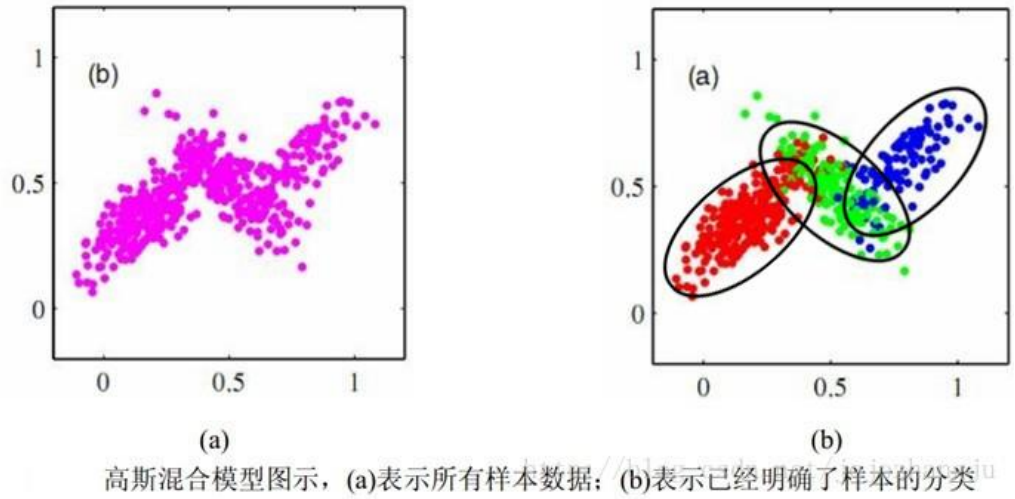
其表示了样本属于 C 的概率大小。若将任意测试样本x_i输入式（4-2），均可以得到一个标量$N(x;\mu,\sum)$。然后根据贬值 t 来确定该样本是否属于该类别。（贬值 t 的确定：可以为经验值，也可以由实验确定。另外可以先另 t=0.7，以 0.05 为步长一直减到 0.1 左右，选择使样本变化最小的哪个贬值作为最终 t 值）

单高斯分布模型的几何意义：单高斯分布模型在二维空间近似于椭圆，在三维空间上近似于椭球，因为在很多场合下，同一类别的样本点并不满足“椭圆”分布的特性，所以才引入了高斯混合模型。[10]

5.3.3. 高斯混合模型

高斯混合模型是单一高斯机率密度函数的延伸，由于 GMM 能够平滑地模拟任何形状的分布状态，因此近年来常被用在语音、图像识别等方面。

例如有一批观察数据$X=\{x_1,x_2\ldots x_n\}$，数据个数为 n，在 d 维空间中的分布



高斯混合模型图示，(a)表示所有样本数据；(b)表示已经明确了样本的分类

图片 5-10 高斯混合模型示意图

不是椭球状，那么就不适合以一个单一的高斯密度函数来描述这些数据点的机率密度函数。此时我们采用一个变通方案，假设每个点均由一个单高斯分布生成（具体参数 未知），而这一批数据共由 M（明确）个单高斯模型生成，具体某个数据 属于哪个单高斯模型未知，且每个单高斯模型在混合模型中占的比例未知，将所有来自不同分布的数据点混在一起，该分布称为高斯混合分布。

从数学上讲，我们认为这些数据的概率密度分布函数可以通过加权函数表示：

$$P(x_i)=\sum_{j=1}^{M}a_jN(x;\mu,\sum) \qquad \sum_{j=1}^{M}a_j=1 \quad （5\text{-}17）$$

$$N(x;\mu_j,\sum_j)=\frac{1}{\sqrt{(2\pi)^m|\sum_j|}}exp\left[-\frac{1}{2}(x-\mu_j)^T\sum_j{}^{-1}(x-\mu_j)\right] \quad (5\text{-}18)$$

a_j是权值因子，其中的任意一个单高斯分布$N(x;\mu,\sum)$叫做这个模型的一个聚类元，只要 j 取得足够大，这个 xx Mixture Model 就会变得足够复杂，就可以逼近任意连续的概率密度分布。

高斯混合模型是一种聚类算法，每个聚类元就是一个聚类中心。即在只有样本点，不知道样本分类的情况下，计算出模型参数$(x;\mu,\sum)$。其计算方法就是 EM 算法。

用训练好的模型去差别样本所属的分类，分为两个步骤：

（1）. 随机选择 k 个聚类元中的一个（被选中的概率是a_j）。

（2）. 把样本带入刚选好的聚类元中，判断是否属于这个类别，如果不属于则返回第一步。

现在假设我们由 N 个数据点，并假设他们服从某个分布，现在需要确定里面的一些参数的值。例如：在 GMM 中，我们就需要确定$(x;\mu,\sum)$这些参数，我们的想法是：找到这样的一组参数，它所确定的概率分布生成这些给定的数据点的概率最大，而这个概率的乘积则称为似然函数。

$\prod_{i=1}^{N}P_r(x_i;\theta)$　似然函数 （5-19）

通常单个点的概率很小，连乘之后数据会更小，容易造成浮点数下溢，所以一般取对数，公式（4-5）变成：

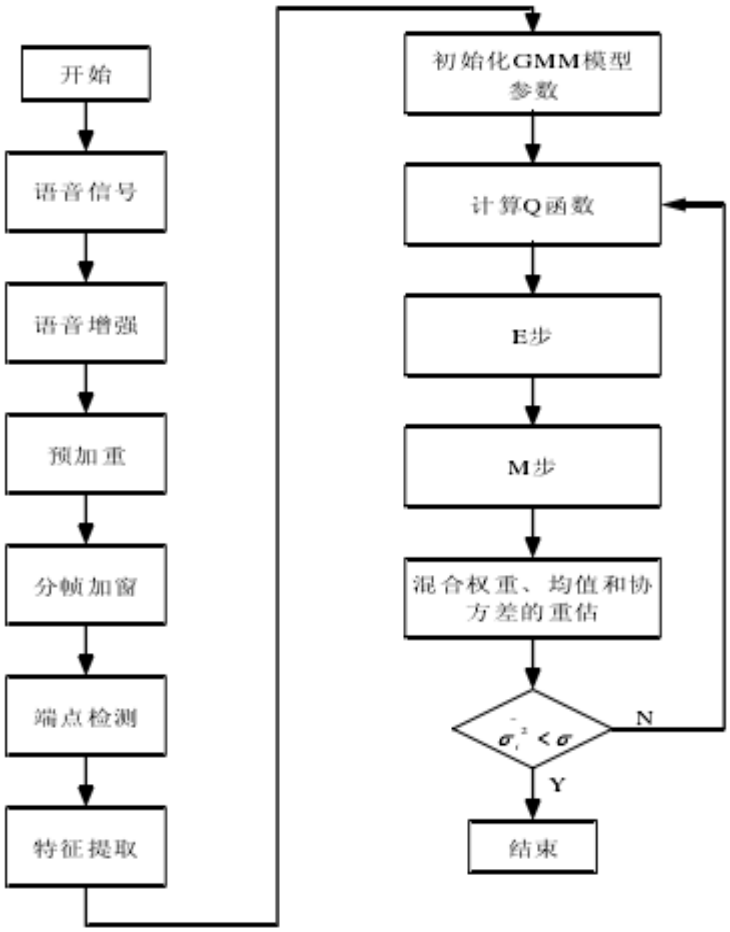
$\sum_{i=1}^{N}log\left(P_r(x_i;\theta)\right)$　　称为 log-likelihood function （5-20）

高斯混合模型的 log-likelihood function 为：

$$\sum_{i=1}^{N}\log\left[\sum_{k=1}^{k}a_kN(x_i;\mu_k,\sum_k)\right] \quad (5\text{-}21)$$

要找到最值的模型参数，使 GMM 的 log-likelihood function 的期望最大，使用 E-M 算法求解。

EM 算法的基本思路是：随机初始化一组参数θ，根据后验概率$P_r(Y|x;\theta)$来更



图片 5-11 高斯混合模型训练过程

新 Y 的期望$\mathrm{E}(Y)$，然后用$\mathrm{E}(Y)$代替 Y 求出新的模型参数θ，如此迭代直到θ趋于稳定。

- E-Step：假设模型参数已知的情况下隐含变量 Z 分别取Z_1, Z_2 ...的期望，亦即 Z 分别取Z_1, Z_2 ...的概率。在 GMM 中求数据点由各个聚类元生成的概率。

  $$\mathrm{r}(i,k) = b_k P_r(Z_k|x_i;a,\mu,\sum) \quad (5\text{-}22)$$

  权值因子b_k表示在训练集中数据点属于类别Z_k的频率，在 GMM 中它就是a_k。

  $$\mathrm{r}(i,k) = \frac{a_k N(x_i|\mu_k,\Sigma_k)}{\sum_{j=1}^{k} a_j N(x_i|\mu_j,\Sigma_j)} \quad (5\text{-}23)$$

- M-Step：用最大似然的方法求出模型参数，认为$\mathrm{r}(i.k)$就是数据点x_i由聚类元 k 生成的概率。

$$N_k = \sum_{j=1}^{N} r(i,k) \tag{5-24}$$

$$\mu_k = \frac{1}{N_k}\sum_{j=1}^{N} r(i,k)x_i \tag{5-25}$$

$$\sum_k = \frac{1}{N_k}\sum_{j=1}^{N} r(i,k)(x_i - \mu_k)\,(x_i - \mu_k)^T \tag{5-26}$$

$$a_k = \frac{N_k}{N} \tag{5-27}$$

5.4. 声纹识别方法

对于一个声纹识别系统，假设整个系统共有 N 个说话人，其对应的 M 阶的高斯混合模型参数分别为：$\gamma_1, \gamma_2 \ldots \gamma_N$。在辨识阶段，给定一个待识别的语音样本的特征矢量序列$\mathrm{X} = \{x_1, x_2, \ldots, x_T\}$，分别定义如下几个函数：$P(\gamma_n)$，$P(x)$，$P(x|\gamma_k)$，$P(\gamma_k|x)$。

$P(\gamma_n)$为第 n 个人的先验概率

$P(x)$为所有人的特征矢量集的概率密度

$P(x|\gamma_k)$为第 n 个人的特征矢量集的概率密度

$P(\gamma_k|x)$为后验概率，表示在特征矢量集为 x 的条件下，确定说话人为第 n 个人的概率。

则这段语音属于第 n 个说话人的最大后验概率为：

$$\mathrm{P}(\gamma_n|x) = \frac{P(x|\gamma_n)P(\gamma_n)}{P(x)} = \frac{P(x|\gamma_n)P(\gamma_n)}{\sum_{m=1}^{N} P(x|\gamma_\mathrm{n})P(\gamma_\mathrm{n})} \quad （5\text{-}28）$$

识别结果n^*可以由最大后验概率给出，即

$$n^* = \arg \underset{1 \le n \le N}{max} [P(\gamma_n|x)] \quad （5\text{-}29）$$

假定该语音信号出自封闭集合，则每个人的先验概率相同，则：

$$\mathrm{P}(\gamma_n) = \frac{1}{N} \quad 1 \le \mathrm{n} \le \mathrm{N} \quad （5\text{-}30）$$

即识别结果等价为：

$$n^* = \arg \underset{1 \le n \le N}{max} [P(x|\gamma_n)] \quad （5\text{-}31）$$

为了使计算更加简便，通常采用对数似然函数

$$\mathrm{L}(x|\gamma_n) = \ln P(x|\gamma_n) \quad （5\text{-}32）$$

得到判决准则为：

$$n^* = \arg \underset{1 \le n \le N}{max} \sum_{t=1}^{T} \ln P(x_t|\gamma_n) \quad （5\text{-}33）$$

其表示在 N 个人中能使$\sum_{t=1}^{T} \ln P(x_t|\gamma_n)$最大的第 n 个人。

其中：

$$P(x|\gamma) = \sum_{i=1}^{M} \beta_i \frac{1}{\sqrt{(2\pi)^S |\sum_i|}} exp\left[-\frac{1}{2}(x-\mu_i)^T \sum_i{}^{-1} (x-\mu_i)\right] \quad （5\text{-}34）$$

β_i为加权系数　　N 为库存人数

μ_i为均值向量　　M 为 SGM 的个数

$\sum_i$为协方差矩阵　　S 为 GMM 的维数

6. 桌面应用

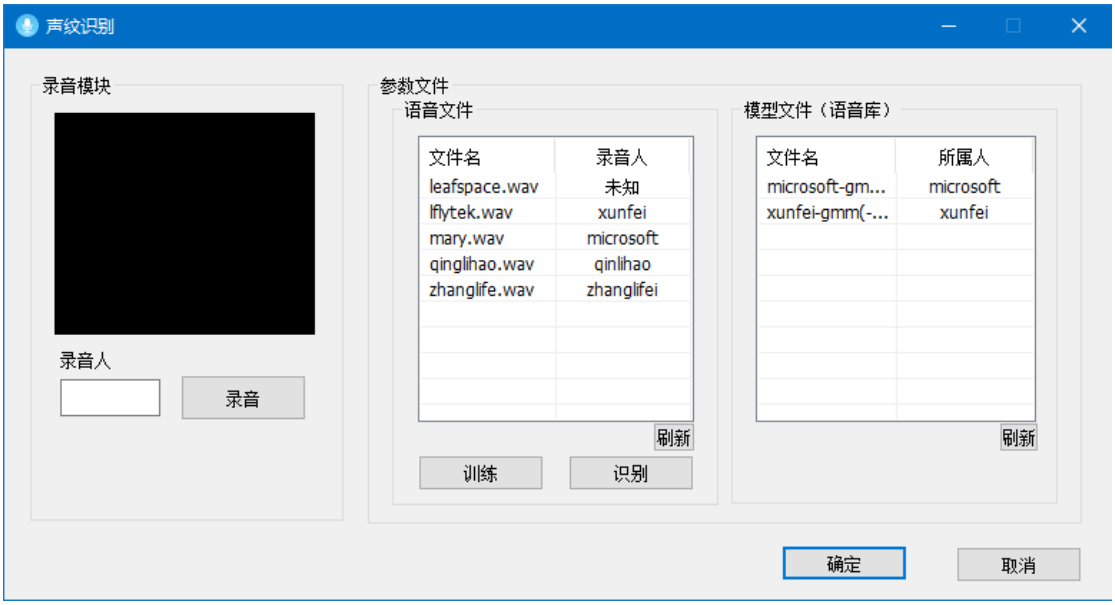
6.1. MFC 界面设计

本课题使用 MFC 进行界面上的设计，对录音 API 与高斯混合模型训练识别进行整合设计完成了一个完整的带有界面的声纹识别系统。下面介绍一些 MFC 常用的标准控件。

表格 6-1 使用的 Windows 标准控件

控件	MFC 类	描述
按钮	Cbutton	用来展示某种行为的按钮，以及复选框
编辑框	Cedit	用于键入文本
列表	ClistCtrl	显示文本及其图标列表的窗口
组合框	CcomboBox	编辑框与列表的组合
静态文本	Cstatic	常用语为其他控件提供标签

界面设计如图 6-1 所示。



图片 6-1 系统界面设计图

（1）. 通过录音模块中的录音按钮，可以即时录音，如果存在相应的录音设备软件会选择操作系统默认的录音设备开始录音，默认情况下，录音数据会采用采样位数 16bit，采样频率 1.Ghz 单声道的设置，并将结果保存在 wav 文件。录音前可在“录音人”的编辑框中填写录音人名。如果在录音完毕后录音人编辑框中有文字信息，则将录音人信息写入进文件。

（2）. 在语音文件与模型文件中用户能够看到库中的录音跟已经训练完成的数据。

（3）. 用户可以随机选择语音文件中的某一项进行训练，即将录音人的高斯混合模型参数写入文件存入库中。

（4）. 用户可以随机选择语音文件中的某一项进行识别，即将待识别人与语音库中的人进行人员识别。

6.2. 实验数据及结果

实验样本数为 5，每个样本中说话人所说的内容皆为“我们需要帮助”。训练每个人的语音到语音库中，每次选择其中个语音样本进行识别，看能否确认这个语音样本是否属于他本人。判决方式：对比每个数据大小，若数据集中绝对值最小的数据则为识别用户。例如：识别数据集为 1，0，-1，-2，5，那么 0 值锁对应的人则为识别结果。

（1）. 样本 wangzhe　文件大小 42144B　帧数 165

表格 6-2 wangzhe 实验结果数据

	Wangzhe	Zhanglifei	Zhaozouxiang	Liuchang	Zhaoquan
$P(x\|\gamma)$	-713.795	-1056.2	-1320.7	-1451.75	-2773.97

（2）. 样本 zhanglifei　文件大小 96044B　帧数：380

表格 6-3 zhanglifie 实验结果数据

	Wangzhe	Zhanglifei	Zhaozouxiang	Liuchang	Zhaoquan
$P(x\|\gamma)$	-3326.97	-2166.06	-4949.54	-3154.88	-8358.6

（3）. 样本 zhaozouxiang　文件大小 43574B　帧数 171

表格 6-4 zhaozouxiang 实验结果数据

	Wangzhe	Zhanglifei	Zhaozouxiang	Liuchang	Zhaoquan
$P(x\|\gamma)$	-1113.54	-3443.29	-724.993	-5008.76	-3609.02

（4）. 样本 liuchang　文件大小 47652B　帧数 181

表格 6-5 liuchang 实验结果数据

	Wangzhe	Zhanglifei	Zhaozouxiang	Liuchang	Zhaoquan
$P(x\|\gamma)$	-1080.08	-1085.65	-1275.73	-640.919	-3005.28

（5）. 样本 zhaoquan　文件大小 45630B　帧数 177

表格 6-6 zhaoquan 实验结果数据

	Wangzhe	Zhanglifei	Zhaozouxiang	Liuchang	Zhaoquan
$P(x\|\gamma)$	-1069.84	-1250.03	-1249.39	-949.363	-2898.08

在 5 个样本中样本 1、2、3、4 都成功识别，只有样本 5 识别结果为同是女生的 Liuchang。整个系统识别率达到 80%。

7. Windows Hello 应用

Windows Hello 是操作系统 Windows 10 中新提出的一种生物验证方法。它可以通过采集登陆者的数据进行后台验证，从而跳过输入密码的过程。

在本课题中，我们采用麦克风来采集登陆者信息，如果验证成功，则将密码组合成键盘消息发送给系统登陆程序（logonUI.exe）；如果验证失败，则弹出提示信息。

7.1. 在 Windows 登陆界面运行自定义程序

（1）. 在系统中运行“regedit”打开注册表，定位到注册表的 HKEY_LOCAL_MACHINE\SOFTWARE\Microsoft\WindowsNT\CurrentVersion\Image File Execution Options；

（2）. 在[Image File Execution Options]上点击右键，选择[新建]在新菜单中选择[项]并将其命名为 utilman.exe；

（3）. 点击[utilman.exe]，在右侧空白处单击右键，选择[新建]在新菜单中选择[字符串值]并命名为[Debugger]；

（4）. 双击打开 Debugger，将其数值数据框中输入要打开程序的路径，如[任务管理器]：C:\Windows\System32\taskmgr.exe；

（5）. 设置完成后，在锁定界面点击“轻松使用”将不会打开菜单，而是启动修改后的程序；从而在 Windows 登陆界面运行声纹识别程序。

7.2. 系统应用流程

在系统进入登陆界面时能够启动该应用。应用启动后会立即启动系统默认的录音设备开始采集说话人的信息。接着应用程序会休眠 5500ms，在此段时间之内，录音线程会自动将采集到的信息写入指定的缓冲区内。当结休眠结束，应用调用停止录音操作，录音线程将所有采集到的信息写入文件中供后面使用。（录音文件长度约 5s，其中 500ms 用于启动并设置录音设备用。）

获取到录音文件后，应用开始读取文件数据，将数据提取出来进行信号初始化并计算信号的特征参数。然后将库存中的数据提取出来，创建高斯混合模型，将预识别人的特征参数放入库存中每个人的高斯混合模型中，计算最大概率。如

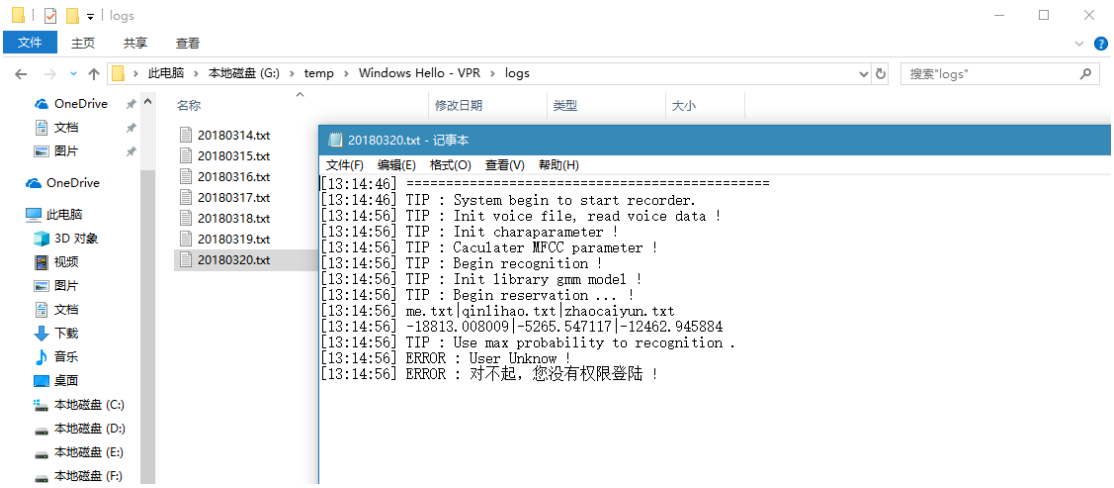
果最大概率是预先设置的设备使用者的数据集，那么系统会向 LogonUI.exe 发送登陆密码的组合消息。否则，则弹出提示。从而实现快速登陆。

8. Web 实时监控日志系统

为了监控系统的状态，捕获异常性的错误信息，web 实时监控日志系统记录了所以程序运行时的状态并上传至服务器。该 web 实时监控系统分为两个部分：基于本地的日志系统。基于远程的日志系统。

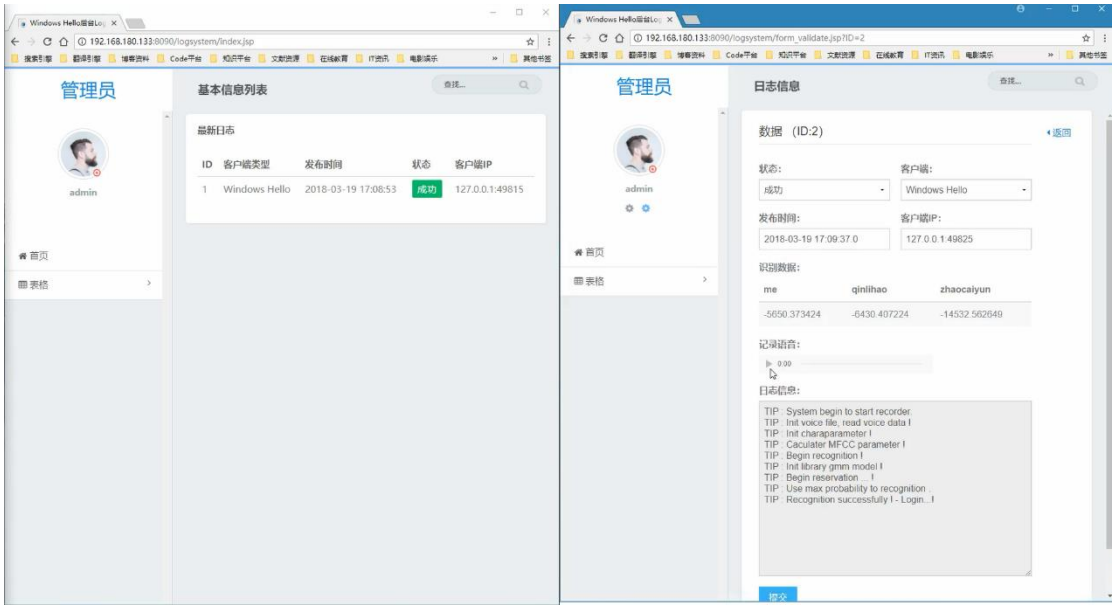
8.1. 基于本地的日志系统

当 Windows Hello 应用程序执行时，会自动向程序存放的位置中新建名为“logs”的文件夹（如果文件夹不存在），并以当前日期的形式新建文本文件。在程序的执行过程中，关键模块的衔接部分或异常性的信息会标明时间同时写入至文件当中。当用户需要时可以随时打开目录文件夹查看应用的执行过程跟执行情况。



图片 8-1 本地日志系统

8.2. 基于远程的 web 日志系统



图片 8-2 远程日志系统

本地日志系统只适用于用户能够访问验证机器的情况。但是远程 web 日志系统能够让已经获得访问许可的人查看日志信息。如果将验证机器通过内网穿透映射至公网内，即使在千里之外也能够查看信息。

将 web 日志系统发布到 web 服务器上，通过连接地址打开网站输入登陆账户跟密码即可访问到日志系统。用户可以查看或删除日志信息，从而实现远程监控系统状态。

结语

本文研究的内容是基于声纹识别的门禁系统，主要通过 MFCC 倒谱系数和高斯混合模型 GMM 来进行说话人信息处理。其中对说话人识别所采用的特征向量进行了筛选与组织，给出了 MFCC 倒谱系数的具体提取过程和算法。还有高斯混合模型的研究，最终将高斯混合模型与实际应用相结合起来，完成了这样一个 Windows Hello 的声纹识别系统。

起初想到这个题目是从朋友那里听来的“语音签名”这个词汇。原本是准备制作医疗信息系统（因为原先的指导老师做了医疗信息系统公司的高级技术顾问，准备利用现有条件开发产值的高新技术系统），但我个人的意思是将“语音签名”的本质声纹识别结合身边的应用做出一个发展。所以我就想到了 Windows Hello 这样一个概念。

经过一个学期的准备与三个月的制作，浏览了 44 篇论文，参考了 68 篇网络技术博客，终于完成了本课题。这个系统最终只是一个实验系统，期待未来能够真正的与微软公司合作，加入声纹识别这样一个认证方式。

最后，希望有一天，说话人识别技术能够真正成熟起来，与其他技术结合起来在现实生活中大放异彩。

参考文献

[1] 杨胜跃，周宴宇. 语音信号端点检测方法与展望. [J].通信技术，2005，卷号（07）：0005-04.

[2] 刘波.基于短时能量和过零率分析的语音端点检测方法研究[R].湖北武汉：武汉理工大学信息工程学院，2004.

[3] 李爱平，党幼云. VQ 声纹识别算法和实验. [J].西安工程科技学院学报，2007，卷号（21）：0848-04.

[4] 周跃海，童峰. 采用 DTW 算法和语音增强的嵌入式声纹识别系统.[J].厦门大学学报，2012，卷号（51）：0174-05.

[5] 于娴,贺松. 基于 GMM 模型的声纹识别模式匹配研究.[J].通信技术,2015,卷号(48):0097-05.

[6] 王正创. 基于 MFCC 的声纹识别系统研究[D]. 无锡：江南大学，2014.

[7] 鲁晓倩，关胜晓. 基于 VQ 和 GMM 的实时声纹识别研究. [J].，2014，卷号（23）

[8] 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，卷号（37）：0008-07.

[9] 臧晓笠. 基于基于高斯混合模型 GMM 的说话人识别方法. [J].科技信息，2006

[10] 辽宁工业大学.基于高斯混合模型的声纹识别方法及系统[P].CN：102324232 A，2011.

[11] 张超琼，苗夺谦. 基于高斯混合模型的语音性别识别.[J].计算机应用，2007，卷号（28）：0360-03.

[12] 茅剑,林奇. 基于声纹识别的嵌入式防盗系统.[J].计算机与现代化，2009,卷号(11):0163-03.

[13] 周雷.基于 MFCC 的声纹识别的说话人身份确认方法的研究[D].上海:上海师范大学，2016.

[14] 杨阳，陈永明. 声纹识别及其应用. [J].语音技术，2007，卷号（02）：0045-02.

[15] 裴鑫. 声纹识别系统关键技术研究[D]. 哈尔滨：哈尔滨理工大学，2016.

[16] 朱浩冰,郭东辉. 声纹识别系统原理及其关键技术.[J].网络安全，2007,卷号(09).

[17] 古今，郭立. 一种基于感知特征的鲁棒性语音认证算法.[J].中国科学院研究生院学报，2009，卷号（26）：474-482.

[18] 古今. 语音感知认证的关键技术研究[D]. 合肥：中国科学技术大学，2009.

[19] 周萍，李晓盼. 混合 MFCC 特征参数应用于语音情感识别. [J].计算机测量与控制，2013，卷号（21）：1996-03.

[20] 郭春霞，裘雪红. 基于 MFCC 的说话人识别系统.[J].电子科技，2005，卷号（11）：11-012.

[21] 韩一,王国胤. 基于 MFCC 的语音情感识别.[J].重庆邮电大学学报，2008,卷号(20):

0597-06.

[22] 丁爱明. 基于 MFCC 和 GMM 的说话人识别系统研究[D]. 南京：河海大学，2006.

[23] 王恩泽，何东健. 基于 MFCC 和双重 GMM 的鸟类识别方法. [J].计算机工程与设计，2014，卷号（35）：1868-04.

[24] 王萌，王福龙. 基于端点检测和高斯滤波器组的 MFCC 说话人识别. [J].计算机系统应用，2016，卷号（25）

[25] 何朝霞，潘平. 说话人识别中改进的 MFCC 参数提取方法. [J].计算机技术与工程，2011，卷号（11）

[26] 向昌盛. 高斯混合模型心音信号自动识别. [J].南京理工大学学报，2016,卷号(40):0560-06.

[27] 蔡桂林. 高斯混合模型用于语音情感识别研究[D]. 广西：广西师范大学，2016.

[28] 翟玉杰. 基于 GMM-SVM 说话人识别的信道算法研究[D]. 吉林：吉林大学，2015.

[29] 陈银燕. 基于 HMM 和 GMM 天然地震与人工爆破识别算法研究[D]. 广西：广西师范大学，2011.

[30] 张奇,苏鸿根. 基于高斯混合模型的乐器识别方法. [J].计算机工程，2004,卷号(30):0133-02.

[31] 杨澄宇，赵文. 基于高斯混合模型的说话人确认系统. [J].计算机应用，2001，卷号（21）：0007-02.

[32] 余清清，李应. 基于高斯混合模型的自然环境声音的识别. [J].计算机工程和应用，2011，卷号（25）：152-155.

[33] 任民宏，鲁秋菊. 基于高斯混合模型自适应肤色识别算法. [J].陕西理工学院学报，2016，卷号（32）：0053-04.

[34] 许百林. 基于矢量量化（VQ）和高斯混合模型（GMM）的说话人识别的研究[D]. 南京：东南大学，2005.

[35] 刘恒,吴迪. 运用高斯混合模型识别动物声音情绪. [J].应用天地，2016,卷号(35):520-20.

[36] 肖汉光，何为. 基于 MFCC 和 SVM 的说话人性别识别. [J].重庆大学学报，2009，卷号（32）：0770-05.

[37] 旁程，李晓飞. 基于 MFCC 与基频特征贡献度识别说话人性别. [J].华中科技大学学报，2013，卷号（41）：0108-04.

[38] 高原. 基于性别分类的说话人识别研究[D]. 徐州：江苏师范大学，2012.

致谢

请允许我致谢，为这四年一路走来的成长，为我经历和得到的所有。

首先我要感谢带我走进语音信号处理的刘永俊老师，一个很巧妙的契机进入了实验室，也正是刘永俊老师带领着我进入了机器视觉这个领域。正是在实验室的两年里，我的能力才渐渐的被发掘，原来本科生也能够完成硕士能够做出的实验 Demo。原来学历并不是本科生不能够学习高技术领域的借口。

我还要感谢我的同窗秦立浩，我做语音，他做图像。他其实比我更加适合做研究性的内容，我适合做一个项目经理。虽然他常常幻想担任我的角色，但其实我更喜欢他这样的人才。毕业了，他考研去了，我即将进入公司，我不清楚我的未来是不是还能遇见这样的朋友，但如果再见，我一定让他担任我们公司的研发组组长。

感谢我的指导老师赵彩云，赵老师是我的操作系统课程的授业老师，在授业过程中时刻都认真负责，她教授的多线程控制的技术、死锁解锁控制到现在我都收益匪浅，包括在本系统中，我也使用了多线程的技术。虽然我在写这篇致谢的时候我才大三，但我知道，未来的你一定是一个认真负责的老师。一定是你，也只有你才能指导我的论文到这么完美的地步。

感谢 14 软件单招班的所有同学，虽然你们都比较贪玩，不明白未来有多少困难等着你们，但我相信，未来一定有属于你们的出路。请不要畏惧，向着你们所向往的目标去飞吧，你们是我这 24 年来的骄傲与辉煌的见证者。

感谢所有教导过我们的老师，在我心里，你们各有特色，感谢你们的教导，感谢你们！