基于 i-vector 声纹识别上课点名系统的设计与实现

王 伟，韩纪庆，郑铁然，郑贵滨，周星宇，金 声
（哈尔滨工业大学 计算机科学与技术学院，哈尔滨 150001）

摘 要： 当今课堂教学过程中存在着学生旷课、老师需要经常检查课堂出勤情况等问题。本系统着眼于课堂教学的实际需要，旨在方便任课老师了解学生上课出席情况，以及防止冒名顶替等不公正现象的发生，开发了基于 i-vector 声纹识别技术的上课点名系统，不仅在说话人识别的研究领域有探索意义，而且在方便老师课堂管理方面有着重要的实践意义。

关键词： 上课点名系统；声纹识别；跨平台交互
中图分类号： TP391 **文献标志码：** A **文章编号：** 2095-2163(2016)06-0108-03

Design and realization of class attendance system based on i-vector speaker recognition

WANG Wei, HAN Jiqing, ZHENG Tieran, ZHENG Guibin, ZHOU Xingyu, JIN Sheng
(School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China)

Abstract: Nowadays, there exists student absenteeism so that teachers need to check the classroom attendance regularly in classroom teaching process. The system focused on actual needs of teaching, aimed at helping teacher to know the classroom attendance, therefore preventing unfair phenomenon like imposter. Developing class attendance system based on an i-vector speaker recognition is not only of great exploring significance in research field of speaker recognition, but also has great practical significance for class management.
Keywords: class attendance system; speaker recognition; interacting across heterogeneous platform

0 引 言

声纹识别(voice print recognition)也称为说话人识别[1](speaker recognition)，是通过对说话人语音信号特征的分析处理，识别说话人身份的过程。与语音识别不同，说话人识别侧重于说话人的身份而非说话的内容。按照说话内容的类型，可分为文本有关和文本无关。前者要求说话人在训练与识别阶段说相同的内容，而后者无此要求。所以，文本无关说话人识别应用将更为广泛，但识别难度也必然更大。

对说话人识别的研究始于 20 世纪 30 年代，早期主要进行有关人耳听辨方面的研究。而对说话人自动识别的研究则需上溯自 60 年代。在语音特征提取方面，1962 年 Kesta 提出使用语谱图进行说话人识别的方法[2]，1969 年 Luck 等人将倒谱技术首度应用于说话人识别[3]，1976 年 Atal 等人进一步提出线性预测倒谱系数[4]。而在说话人识别模型方面，60 及 70 年代初期，主要采用的是模板匹配方法。70 年代后期，动态时间规整和矢量量化技术相应地已然成为研究和发展重点[5-7]。

80 年代后，Davis 等人提出了 Mel 频率倒谱系数(Mel Frequency Cepstrum Coefficient，简称 MFCC)[8]。由于 MFCC 考虑了人耳的听觉感知机理，表现出良好的识别效果和噪声鲁棒性，从而成为说话人识别中使用的基础评判参数。与此同时，人工神经网络和隐马尔可夫模型[9]也在语音识别领域获得了成功与广泛的应用，由此晋升为说话人识别中的核心技术。90 年代后，高斯混合模型凭借其简单、灵活、有效以及出众的鲁棒性，迅速演进成为目前与文本无关的说话人识别中的主流技术[10]。进入 21 世纪以后，Reynolds 等人提出 GMM-UBM(Gaussian Mixture Model with Universal Background Model)模型用于文本无关说话人识别，使得训练 GMM 的样本数量要求减少，文本无关的说话人识别开始从实验室走向了广阔现实应用[11]。

2006 年，Campbell 等人在 GMM-UBM 基础上提出高斯超向量概念，并与支持向量机融合为高斯混合超向量-支持向量机模型用于文本无关说话人识别，由此突破性地发展成为目前国内外文本无关说话人识别的热点关键实用技术[12]。近年来，学者们又在高斯超向量基础上，提出了联合因子分析[13]、鉴别性向量(i-vector)[14]等模型，使得文本无关说话人识别系统的性能取得了显著的改进与提升。本系统是基于 i-vector 构建的声纹识别系统。i-vector 说话人建模技术[15]的基本思想可大致描述为：信道和会话的影响均包含在总体变化子空间中，通过对包含说话人信息和信道信息的 GMM 均值超矢量在低秩的总体变化子空间上进行投影，得到只包含说话人信息的低维矢量。基于声纹识别的上课点名系统为教师课堂出勤管理提供了一整套行之有效的解决方案，从而大大提高了上课点名系统的性能。

1 系统的整体架构

上课点名系统总体架构分为 PC 端和移动端，两者通过

收稿日期：2016-02-02
基金项目：高等学校博士学科点专项科研基金博导类资助课题(20112302110042)。
作者简介：王 伟(1977-)，女，博士研究生，讲师，主要研究方向：语音信号处理、音频信息处理；韩纪庆(1964-)，男，博士，教授，博士生导师，主要研究方向：语音信号处理、音频信息处理；郑铁然(1974-)，男，博士，副教授，主要研究方向：语音信号处理、音频信息处理。

无线网络相连。系统呈现 C/S 结构，在整体架构上可分为 4 层：最底层由无线网络、PC 和手机来提供支撑；第二层包含网络 TCP/IP 协议和 Socket；第三层由声纹识别算法、数据库和 Android 系统的 API 构成；最顶层即由系统各功能模块组成。综上可知，以上 4 层构成了完整的上课点名系统。

上课点名系统分为 5 大模块，具体可表述为：PC 端的训练模块、点名模块、统计模块、移动端的点名模块和本地录音模块。PC 端承载了大运算量的识别任务，而移动端只负责语声的录音、简单转换和发送工作。PC 端点名模块设计为主控整体的点名过程，移动端的点名模块则实施完成每个学生的点名工作。

1.1　训练模块

训练模块是上课点名系统的核心模块之一。训练过程一般由管理员控制并完成，模型训练后即可分发给教师使用。训练模块有如下的功能：

1) 训练参数设置。训练参数设置包括 MFCC 参数设置、UBM 和 TV 参数设置和默认路径设置。其中，MFCC 参数设置项包括帧长、帧移、Mel 滤波器个数、MFCC 维数、是否加入 Δ、是否加入 ΔΔ、预加重系数和提升系数。UBM 和 TV 参数设置项包括 UBM 混合数和 i-vector 维数。

2) 模型训练。模型训练包括 MFCC 特征提取、UBM 模型训练、TV 模型训练、i-vector 提取和模型存储等过程。

3) 模型载入。模型载入包括 UBM 模型载入和 TV 模型载入。

1.2　点名模块

点名模块是上课点名系统的关键模块，主要由教师负责上课的点名工作。教师利用此模块可进行以下操作：

1) 教师在自己教授课程中选择当前课程。

2) 查看当前课程的所有学生。

3) 灵活设定点名策略。例如：点名策略可采用学号尾号策略、学号倍数策略和限定人数策略等。

4) 可以选择自由点名（无顺序，学生可在任何情况下答到）或顺序点名（学生只能在被点名时答到）。

5) 可以控制点名进度，例如暂停或继续本次点名。

6) 可以设置判定阈值。

7) 在系统自动化判断的基础上，可手动更改结果。

8) 保存点名结果。

如果从移动端与 PC 端交互来看，将可展开如下步骤：

1) PC 端开启 Server，等待移动端连接。

2) 移动端成功连接 PC，显示已连接状态。

3) PC 端发出点名请求，移动端准备录音。

4) 移动端录音完毕，并发送到 PC 端。

5) PC 端接收录音，显示已接受状态。

6) PC 端调用模型测试，并显示判断结果。

7)（可选）教师修正判断结果。

如果从学生点名来看，则可表述为如下 4 种状态：

1) 未登录。此时表示该学生并没有登陆主机。可能没有来教室或已来教室但手机还未登录。

2) 已登陆。此时该学生手机已登录上主机，准备答到。

3) 已答到。此时该学生已通过登陆手机进行答到。

4) 已判定。系统已经通过内部算法给出识别结果。

1.3　统计模块

统计模块是上课点名系统的必要模块。具体来说，就是在教师登陆后可选择进入统计界面，在设计上每个教师只能看到自己所教课程学生的统计信息，教师利用该模块可方便了解学生的出勤情况，或者依此来计算平时成绩中对应的出勤部分结果。现实运作时，模块将通过条件查找和筛选，快捷优势获得想要的信息，这就极大减轻了教师们的实际日常负担。统计模块可完成如下的功能：

1) 能查看自己所教课程的点名记录。

2) 能对点名记录进行筛选工作（比如只看被点到过的）。

3) 能查看某一学生的个人详细记录。

4) 能手动更改某一学生的某一记录。

1.4　移动端点名模块

移动端点名模块也是支撑该系统的关键模块之一。PC 端的点名模块需要移动端点名模块的配合才能完成整个实时点名工作。移动端点名模块是学生用于答到的功能途径，因为需要与主机处于同一内网中，就会使得在教室范围外无法进行答到，这样也避免了“远程答到”现象的发生。同时，待答到的学生也必须身处该教室中才能知道自己是否需要答到，或者知道自己正在被点名。综上的这种双重设计使得作弊现象势将完全根绝。

移动端点名模块有如下的功能：

1) 与 PC 端主机连接的功能，即登陆主机端功能。

2) 传送语音数据的功能。

3) 对语音进行初步处理的功能，比如转换成 wav 文件。

移动端点名模块功能实现的操作步骤具体如下：

1) 登陆主机。包括建立连接和发送学生信息等。用户进入登录界面后，输入主机 IP 和学号即可登录，登录的同时系统即会自动将信息发送到 PC 端。登录成功后转至点名界面。

2) 录音传送。包括语音录制、语音转换和语音传送等。在点名界面下，按住语音键可进行录音，语音录制的同时系统则自动将其转换成 wav 格式（该系统目前只支持 wav 格式的录音），松开键后语音就会通过网络传送到 PC 端。

1.5　移动端本地录音模块

移动端本地录音模块是上课点名系统的辅助模块，主要是为了方便学生自己采集语音，通过组织层次（如按班收集），最终提交给管理员进行总体训练。另外，由于学生训练和测试均采用同一信道，使得因信道差异而造成的对识别率的干扰也降至最低，从而提高了点名准确率。移动端本地录音模块中最为重要的功能就是本地录音，具体则包括语音录制、语音转换和保存录音等。

2　声纹识别算法的设计与实现

说话人识别技术[10]是上课点名系统研究开发中选用的专业核心技术，同时这也是影响和决定该系统各项性能以及可用性的重点实现部分。从 PC 端成功接收移动端发送的语音信息开始，直到 PC 端显示结果前，由说话人识别/确认模块进行智能处理和模式识别，从而判断出话者的身份（是否为同一人）。一个完整的说话人识别过程通常可分为 2 个阶段：训练阶段与识别阶段。从功能上，则可以分为 3 个功能模

块:特征提取、模型训练和模型打分。

在宏观思维上,文本无关说话人识别可分为说话人辨识和说话人确认。其中,说话人辨识是判定说话人是一群说话人中哪一个,而说话人确认则是确定说话人是不是某一个说话人。具体推断可知,上课点名系统可明确归至说话人确认。在训练阶段,首先在PC端由特征提取模块实现对输入语音的参数提取,然后再经由模型训练模块得到相应的说话人模型,模型是对说话人特征的进一步抽象。而在识别阶段,对移动端传来的测试语音进行特征提取后,将利用模型给测试语音来完成打分,并经过不同的判决准则,由此得到最终的判决结果。

3 系统关键任务实现

3.1 跨平台网络交互实现

由于该系统横跨两大平台Windows桌面平台和Android移动平台,即使得两者间通过网络的交互和数据的准确传送就显得尤为重要。首先,需要重点考虑是选用HTTP协议还是较为底层的TCP/IP协议。就现实方便而言,HTTP协议相应地要更为占优,对于HTTP消息的解析工具和API均已臻至完善。但是由于HTTP协议直接针对应用层部分,每一个数据包中都含带了许多冗余的参数信息,在效率上将略差于采用底层的TCP协议。而且该系统对实时性要求很高,尤其是声音文件的传送更需要做到尽量地快捷、快速,虽然在编程工作量方面TCP比HTTP要显现劣势,但是综合考虑性能,本研究决定选用较为底层的TCP协议进行编程。

其次,是PC端与移动端两大平台间消息格式的设定,其设定的好坏将直接关系到数据能否实现稳定快速的传送。研究中,系统采用了自定义TLV(Tag-Length-Value)编码,该方式以标签头+内容长度+内容的方式进行编码,如此既保证了数据发送接收的一致性,又提高了整体运行效率。在此,给出这一信息编码格式如图1所示。

2字节tag	4字节长度	不限字节内容

图1 交换编码格式图

Fig. 1 Exchange encoding format chart

当TAG=1时,表示内容为JSON格式编码。当TAG=2时,表示对应内容为二进制编码(该系统一般为语音文件)。由于平台和编程语言的不同(PC端为C++、Android端为JAVA),在实现方式上又将各有差异。

3.2 计算繁重任务的多线程实现

在训练和测试中,由于数据量大,运行时间长,若单线程运行会出现"假死"的现象,所以多线程操作也就成了必然。QT中的多线程采用继承QThread类,通过实现run函数来构建完成。多线程运行的同时,还需要并行与主线程消息的通讯,如主线程需要获知模型训练的进度等。这个时候就可以采用QT独有的Signal-Slot机制,利用异步性质,使得多线程的消息传递简单可靠。

4 结束语

基于i-vector声纹识别技术的上课点名系统为教师课堂出勤管理提供了一整套行之有效的解决方案,减轻了教师上课点名的负担,减少了冒名顶替等不诚信、不公正现象的发生。系统在PC端使用QT5和C++实现,在移动端使用JAVA和Android SDK实现,完成了从PC端到移动端完整清晰、连贯一体的系统;该系统具备简洁实用的特性,训练时既可采用默认设置一键训练,也可自己调整参数手动训练;点名时既可采取无管控、无顺序的自由点名,也可采用模拟传统方式的顺序点名。与此同时,声纹识别的最新技术研究必然对该领域的创新研究有启发性作用。

参考文献:

[1] CAMPBELL J P. Speaker recognition: a tutorial [J]. Proceedings of the IEEE, 1997, 85(9): 1437-1462.

[2] KERSTA L G. Voiceprint-Identification infallibility [J]. The Journal of the Acoustical Society of America, 1962, 34(12): 1978.

[3] LUCK J E. Automatic speaker verification using cepstral measurements [J]. The Journal of the Acoustical Society of America, 1969, 46(4B): 1026-1032.

[4] ATAL B S. Automatic recognition of speakers from their voices [J]. Proceedings of the IEEE, 1976, 64(4): 460-475.

[5] SOONG F K, ROSENBERG A E, JUANG B H, et al. A vector quantization approach to speaker recognition [J]. AT&T technical journal, 1987, 66(2): 14-26.

[6] RABINER L R, PAN K C, SOONG F K. On the performance of isolated word speech recognizers using vector quantization and temporal energy contours [J]. AT & T Bell Laboratories Technical Journal, 1984, 63(7): 1245-1260.

[7] SAKOE H, CHIBA S. Dynamic programming algorithm optimization for spoken word recognition [J]. IEEE Transactions on Acoustics, Speech and Signal Processing, 1978, 26(1): 43-49.

[8] DAVIS S, MERMELSTEIN P. Comparison of parametric represent-ations for monosyllabic word recognition in continuously spoken sentences [J]. IEEE Transactions on Acoustics, Speech and Signal Processing, 1980, 28(4): 357-366.

[9] RABINER L. A tutorial on hidden Markov models and selected applications in speech recognition [J]. Proceedings of the IEEE, 1989, 77(2): 257-286.

[10] REYNOLDS D A, ROSE R C. Robust text-independent speaker identification using Gaussian mixture speaker models [J]. IEEE Transactions on Speech and Audio Processing, 1995, 3(1): 72-83.

[11] REYNOLDS D A, QUATIERI T F, DUNN R B. Speaker verification using adapted Gaussian mixture models [J]. Digital signal processing, 2000, 10(1): 19-41.

[12] CAMPBELL W M, STURIM D E, REYNOLDS D A, et al. SVM based speaker verification using a GMM supervector kernel and NAP variability compensation [C] // IEEE International Conference on Acoustics, Speech and Signal Processing. Toulouse: IEEE, 2006: 97-100.

[13] KENNY P, BOULIANNE G, OUELLET P, et al. Speaker and session variability in GMM-based speaker verification [J]. IEEE Transactions on Audio, Speech, and Language Processing, 2007, 15(4): 1448-1460.

[14] DEHAK N, KENNY P, DEHAK R, et al. Front-end factor analysis for speaker verification [J]. IEEE Transactions on Audio, Speech, and Language Processing, 2011, 19(4): 788-798.

[15] 栗志意, 何亮, 张卫强. 基于鉴别性i-vector局部距离保持映射的说话人识别 [J]. 清华大学学报(自然科学版), 2012, 52(5): 598-601.

[16] 鄂勇, 李宓, 李子明. 文本无关的说话人识别研究 [J]. 数字通信, 2013, 40(4): 48-52.