

生产环境下声纹识别系统的设计与实现

张 舒¹, 王成强¹, 李 想², 李 慧²

(1. 淮海工学院商学院, 江苏 连云港 222005)

(2. 淮海工学院计算机工程学院, 江苏 连云港 222005)

[摘要] 声纹识别是一种能根据待识别语音的声纹特征识别说话人的技术. 本文阐述了声纹识别系统的原理知识, 介绍了声纹识别系统的体系架构, 本系统采取分层结构, 核心业务分解为业务层和实现层, 多个层的功能模块被设计成独立的服务, 从而提升了声纹识别系统的识别准确率, 最后给出了系统在实际生产环境下的产品设计方案.

[关键词] 声纹识别, 原理, 产品, 架构设计

[中图分类号] TP391 **[文献标志码]** A **[文章编号]** 1672-1292(2018)04-0086-07

The Design and Implement of Voiceprint Recognition System in the Production Environment

Zhang Shu¹, Wang Chengqiang¹, Li Xiang², Li Hui²

(1. Business School, Huaihai Institute of Technology, Lianyungang 222005, China)

(2. School of Computer Science, Huaihai Institute of Technology, Lianyungang 222005, China)

Abstract: The voiceprint recognition is a kind of technology that can recognize the speaker according to the voiceprint characteristics of the voice to be recognized. The paper illustrates the principal knowledge of voiceprint recognition system, and introduces its system architecture. This system adopts a hierarchical structure and the core business layer is also divided into a business layer and implementation layer. Several functional modules of multiple layers are designed as independent services, thus improving the recognition efficiency of the voiceprint system. Finally, the product design scheme of the system under the actual production environment is given.

Key words: voiceprint recognition, principle, product, architecture design

声纹识别也称为说话人识别,同指纹识别、人脸识别以及虹膜识别一样,是一种基于生物特征进行身份认证的技术. 通过计算和分析语音的波形以及波形中反映的说话人生理和行为特征的语音参数来判断说话人的身份. 声纹识别相比于其他认证技术有如下特点:成本低廉,只需要麦克风等录音设备,不需要增加额外的硬件设备;用户接受程度高,低隐私性,用户无须特别保护隐私,且无须记忆;安全可靠,通过动态码可以防止回放录音及录音拼接.

声纹识别系统是一个基于声纹识别引擎且集成鉴权、会话管理、数据存储等功能的面向用户的软件服务,声纹识别系统在移动支付、智能硬件等领域的应用广泛. 基于声纹识别服务,开发者可以开发出多种类型的产品. 例如,使用声纹识别来实现上下班的打卡考勤,使用声纹操控智能家居、进行移动支付等. 随着物联网的快速发展,以及人们对信息安全要求的不断提升,声纹服务的应用前景十分广阔.

声纹识别技术^[1]由于应用领域非常广泛而被人们高度关注,现有的研究方法主要有线性预测倒谱系数(linear prediction cepstrum coefficient, LPCC)^[2]、梅尔频率倒谱系数(Mel frequency cepstral coefficients, MFCC)^[3]. LPCC 是一种线性预测参数,基于人声道的个体性差异而被应用于声纹识别. MFCC 是一种非

收稿日期:2018-04-18.

基金项目:连云港市科技计划项目(JC1608、CG1611)、江苏省“青蓝工程”培养对象、江苏省第五期“333 高层次人才培养工程”、淮海工学院教学改革项目(XJG2017-2-5)、教育部协同育人项目(201702134005、201701028110)、连云港市“521 高层次人才培养工程”(ZKK201604).

通讯联系人:张舒,讲师,研究方向:智能信息处理,数据采集与数据挖掘. E-mail:shufanzs@126.com

线性的特征参数,能够有效模拟人耳耳蜗的听觉特性,更符合人的听觉机理^[2],其应用范围更广。声纹识别的另一个核心模块就是声纹模型。声纹模型应用比较广泛的有:动态时间规整模型动态时间归整算法(dynamic time warping, DTW)^[4]、矢量量化模型(vector quantization, VQ)^[2]、隐马尔可夫模型(hidden markov model, HMM)、高斯混合模型(gaussian mixture model, GMM)^[5]、支持向量机模型(support vector machine, SVM)^[6]等。

本文将系统阐述声纹识别技术的原理及理论知识,然后对声纹识别系统在实际生产应用中所需的各功能、业务模块进行讲解,最后给出面向服务的、分布式的声纹识别系统架构。

1 声纹识别背景知识

声纹识别研究的重点主要集中于特征参数提取和模式匹配两个方面。

1.1 特征参数提取

特征参数提取即提取语音中能够表征说话人信息的基本特征,这些特征须具有精确性和稳定性。一般地,语音信号会被看作为短时平稳的序列,在进行语音特征提取时,首先要对语音信号进行分帧处理,使用窗函数来降低由于 Gibbs 效应(在进行截断处理时导致)产生的影响,由于需要压缩语音的动态范围,我们还需要利用高频预加重技术来提升语音中高频信息。然后使用频谱技术对每一帧语音进行频谱分析,可得到各种不同、覆盖较全面的特征参数。目前常用的特征参数有 LPCC、MFCC 和感知线性预测参数(perceptual linear predictive, PLP)等。MFCC 在说话人识别实验中比 LPCC 和 PLP 表现出了更优秀的识别性能。MFCC 是目前应用最广的特征参数。为了提高识别性能,多种线性和非线性的变换方法也相继被提出。其中特征规整法和特征高斯化通过调整特征参数的统计分布,可以大幅提高特征参数的鲁棒性。

1.2 模式匹配方法

在目前的声纹识别领域,常用的模式匹配方法可以分为两类:模板匹配法、概率模型法。

1.2.1 模板匹配法

模板匹配法是指在进行声纹预留时,从说话人预留的语音中分析提取相应的特征参数,要求提取的特征参数要能够详尽地描述说话人的语音特性,之后将其作为参考模板。在进行测试时,采取同样的方法从说话人的测试语音中提取得到测试模板,将测试模板和参考模板进行比较,根据两者匹配度进行相应的判断。模板匹配法的不足,就是对声纹模型的存储需求较大,在参考说话人数规模较大时,识别性能较差。

1.2.2 概率模型法

概率模型法,相比于模板匹配法来说,概率似然得分也更有意义,且更加灵活。概率模型法是根据概率或者似然得分进行相关的匹配判别,并根据概率分布建模。与模板匹配法差别较大。因为模板匹配法是根据模板来建模,且根据与模板的距离来判断类别。目前常用的概率模型法有高斯模型、GMM 和 HMM 等。其中高斯混合模型和隐含马尔可夫是声纹识别中两种最常用的概率模型。高斯混合模型,即使用若干个线性组合(高斯分布)来模拟多维矢量的任意的、连续的概率分布。能有效地描述说话人的特性。在文本无关的说话人识别领域,高斯混合模型的识别性能最好。在文本相关的说话人识别中, HMM 用的较多,因为它可以描述语音随时间变化的情况,对已知信息的利用率较高。

1.3 通用背景模型

在声纹识别过程中,用于训练模型的预留语音越多越好,这样可以充分获得其声纹特征。但由于在实际操作中,很多用户并不愿意录制大量语音,所以通常用来训练声纹模型的语音很少,因此无法从语音中提取足够的特征参数,进行测试时测试语音与模型的匹配程度较差。为了解决此问题,引入了通用背景模型的概念。通用背景模型(universal background model, UBM)是一种与说话人无关的、高阶的 GMM:通常用数百人、男女声均衡的数小时语音训练得到,用于表示说话人无关的特征分布。UBM 通过语音自适应得到说话人模型。当引入了 UBM 后,预留语音中包含的特征参数,可用说话人自己的特征进行建模,预留语音中所没有覆盖到的特征,则使用 UBM 来近似模拟,从而使得模型的匹配精准度大大提高。

2 声纹识别过程

声纹识别的主要过程包括声纹预留和声纹验证。

2.1 声纹预留

在进行声纹预留时,系统首先会对用户录制的 N 段内容不同的短语音进行预处理,提取说话人特征参数,并采用相关的建模方法,生成声纹模型. 声纹模型又包括语音模型和说话人模型,说话人模型为 GMM 模型,由 UBM 自学习(自适应)而来,具体实现时,则需要自适应均值参数,即

$$\hat{\mu} = \beta E_i(X) + (1 - \beta) \mu_i. \quad (1)$$

式中, i 为背景模型中所包含的每个高斯函数对应的索引; $E_i(X)$ 为自适应数据的均值期望; μ_i 为原始 UBM 的均值; $\hat{\mu}$ 为自适应后得到的均值, β 为调节系数.

声纹预留过程如图 1 所示.

2.2 声纹验证

在声纹验证阶段,说话人所声明的身份信息和语音是必须的输入参数. 首先,系统会对传进的语音进行预处理操作,以提取声纹特征,并将其与声纹预留阶段产生的声纹模型进行匹配,最后根据识别打分判定该语音是否属于该说话人.

声纹验证过程是一个融合的过程. 在验证语音经过特征提取操作之后,还需要分别进行语音识别(基于 HMM)和声纹确认(基于 GMM). 基于 HMM 的语音识别,需要根据提示文本生成对应的受限语法,采用 Viterbi 解码算法,得到语音识别得分. 受限语法与提示文本具有依赖关系,这也就导致了用录音设备回放其它数字串时,正确识别的数字个数会很少,识别得分也会很低. 该方法用于内容鉴别,避免录音冒充.

系统融合得分计算,表达式为:

$$S_F = \frac{1}{1 + \exp(-(S_{ASR}/2 + \alpha \cdot S_{VPR}))}. \quad (2)$$

式中, S_F 为系统融合得分; S_{ASR} 为基于 HMM 的语音识别得分; S_{VPR} 为 GMM 的声纹确认得分; α 是调节系数,可根据实际应用调节. 系统融合得分将与预设阈值比对,超过阈值则表示接受通过,未超过则予以拒绝,阈值可根据实际应用进行调整.

声纹验证过程如图 2 所示.

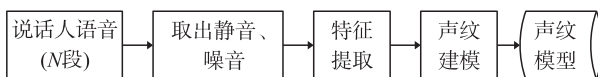


图 1 声纹预留过程

Fig. 1 Voiceprint reservation process

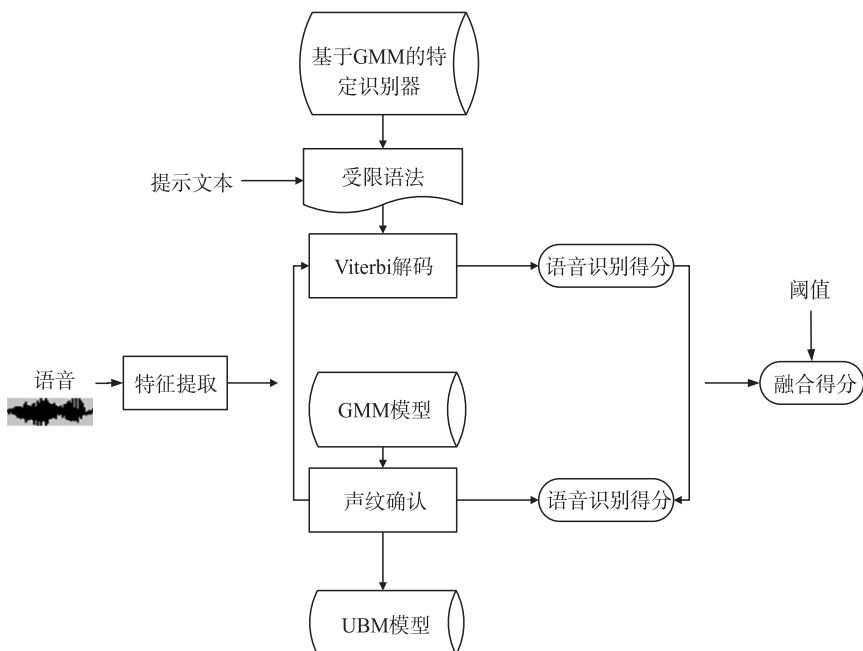


图 2 声纹验证过程

Fig. 2 Voiceprint verification process

3 声纹识别系统

3.1 架构设计

当声纹识别系统作为一个在生产环境中真实可用的产品,它就不只需要上述的声纹识别算法(引擎),还需要拥有独立的业务逻辑以及流程控制、存储等模块,以保证声纹识别系统能够更好地融入应用环境,且具有更好的可控性、可靠性. 本系统设计为了保证各模块功能的内聚、模块间的解耦、以及分布式部署的需求,将声纹识别系统划分了如下4层:

3.1.1 接口层

调用业务层逻辑,实现对外接口,供应用(手持设备等)进行请求调用,提供对多种请求协议的支持,使同一个业务逻辑能够处理不同的应用请求.

3.1.2 业务层

业务层主要负责提供执行声纹识别业务的规则,业务流程的实现,并且为应用的特定业务提供事务处理、安全控制. 包括以下功能模块:

- 鉴权:负责提供接口访问权限的注册、判断、管理等功能.
- 会话管理与信息共享:通过定义会话对象保持交易状态. 提供会话信息创建、记录、更新、查询功能.
- 声纹识别逻辑:调用声纹识别引擎,结合动态文本、声纹存储等功能实现声纹识别的业务逻辑. 此外,业务模块还包括交易信息记录、计费等业务功能、交易信息记录、客户端参数配置等模块.

3.1.3 实现层

包括业务层各模块功能的具体实现和声纹识别引擎(声纹识别的核心算法库,包括声纹预留、声纹验证、声纹特征检测及自学习等功能算法)的具体实现. 该层与业务层一起组成声纹识别系统的业务核心.

3.1.4 支持层

包括数据库服务、文件存储、日志服务、缓存服务等功能. 实现对声纹模型、语音数据、用户信息等持久化. 并为业务层与实现层提供基础支持.

系统架构如图 3 所示.

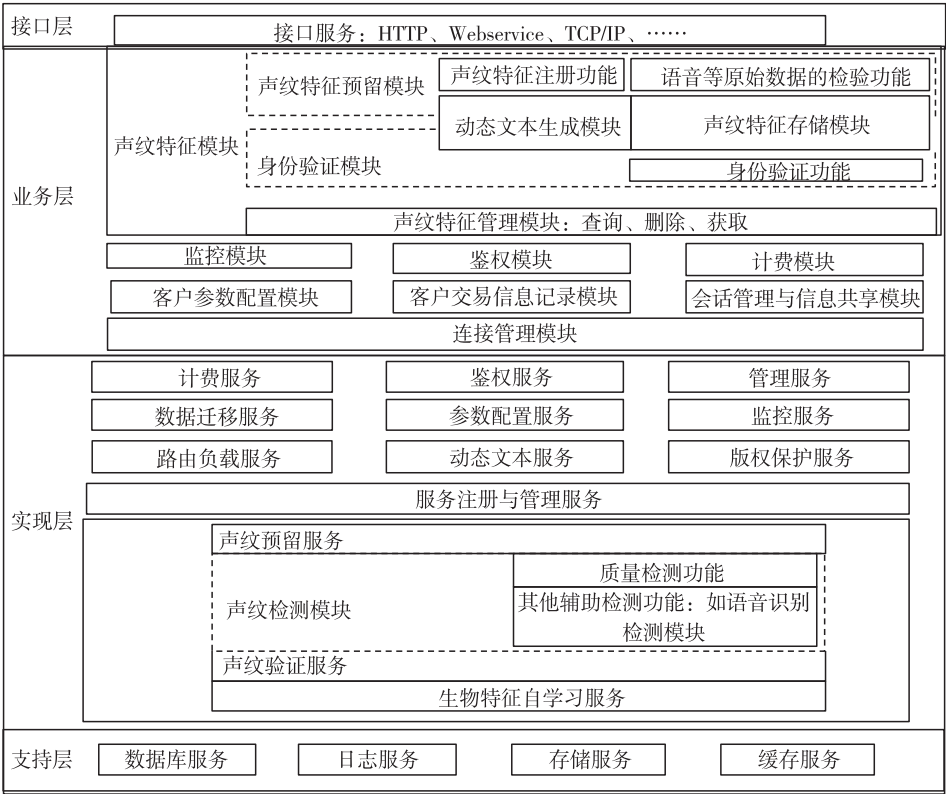


图 3 声纹识别系统架构图

Fig. 3 Voiceprint recognition system architecture

3.2 模块交互

由于声纹服务目前普遍用于移动登陆与支付模块,相对而言,系统的并发量较大,这就要求系统要能够支持分布式部署.因此系统采取了分层的结构,核心业务也被设计成了业务层和实现层,多层的多功能模块被设计成独立的服务.这样保证了系统的高内聚和低耦合,但是这样也带来了一个问题,如何设计系统的各模块的通信机制,以便提高系统的运行效率.本文采用远程过程调用(remote procedure call, RPC)和服务治理模式来解决这个问题,且采用发布-订阅的模式.首先,需要被调用的各服务在服务启动时就在服务协调组件中注册自己的基本信息,包括服务地址、端口、名称、状态等,当服务消费方需要调用服务时,首先到服务协调组件中,找到服务的信息,然后根据服务信息直接调用服务.服务协调组件可以根据一定的算法在集群部署服务中找到最适合的服务返回信息,实现负载均衡.在服务治理模块设置监控中心的角色,对服务的调用情况进行监控统计,以实现服务的有效管理.为了保证模块间通信的效率,数据采用压缩的二进制进行传输,服务端采用了非阻塞的 IO 模型(多路复用),对客户端与服务器间的连接资源进行池化管理,以实现连接的有效控制重用,保证服务不会因为资源占用过多而崩溃.在客户端的调用方面,则可以根据实际的应用场景选择相应的 IO 模型,例如在实时通讯的场景下,应该选择同步调用的方式,而当服务返回的结果并不会影响正常的业务流程时,则可以选择异步调用的方式,优先保证业务流程的高效率.具体的交互如下:实现层和支持层的各自功能模块都被设计为服务的形式.这些服务以 RPC 的形式提供其功能接口,供上层进行调用.这些服务在启动时都在服务协调组件(例如 zookeeper)中进行注册,服务协调组件中包含服务的完整信息,并且可以对服务的状态进行监控管理.当调用方需要使用到某一服务时,首先请求服务协调组件获得所需调用服务的信息,然后根据返回的信息,直接调用服务.当同一个服务实现了多机部署后,协调组件可以通过选择算法选择最适合调用的服务,保证了各服务的负载均衡.

4 生产环境下的声纹识别过程

生产环境下的声纹识别是一个比较复杂的过程,除了需要调用声纹识别引擎的相关模块外,还涉及到与系统各个模块的交互.

4.1 声纹预留

应用需要依次发送 3 个请求,即获取预留文本和会话信息、上传多条预留语音、进行声纹模型训练.每一次请求,声纹识别系统都会通过鉴权模块去校验其携带的授权码.当鉴权通过后,开始获取预留文本,系统首先会调用会话管理和信息共享模块,生成会话对象(包括会话 id、过期时间等),然后调用动态文本生成模块,生成动态文本,将会话和文本进行返回,期间会调用数据库服务,将相关的信息写入记录表.上传预留语音时,系统根据用户携带的会话信息,更新相关记录表,调用声纹引擎对上传的语音进行语音识别以及质量检测,之后调用存储服务,对声纹信息、语音等进行存储.最后是进行声纹训练,调用声纹识别引擎中的声纹训练模块对语音进行特征提取以及模型训练,并对训练生成的模型以及信息进行记录存储.在一般的业务流程中,通常会把声纹模型的训练与上传语音的步骤整合,即在上传最后一条语音的同时自动进行声纹训练.在声纹预留的中间过程中还会涉及到交易信息的记录管理、缓存服务的调用以提高数据库的操作效率、日志的打印收集、异常信息的处理以及与服务监控模块的交互等等.

声纹预留过程如图 4 所示.

4.2 声纹验证

与声纹预留类似,在声纹验证阶段,应用也需要依次发送 3 个请求,即获取验证文本和会话信息、上传一条验证语音、进行声纹验证.每一次请求,声纹识别系统都会通过鉴权模块去校验其携带的授权码(授权码为访问者通过特定 key 值访问鉴权服务所得,具有唯一性和时效性).当鉴权通过后,开始获取验证文本,系统首先会调用会话管理和信息共享模块,生成会话对象(包括会话 id、过期时间等),然后调用动态文本生成模块,生成动态文本,向客户端返回会话信息和验证文本,中间过程中,会调用数据库服务,将交易信息写入相关的记录表.上传验证语音时,系统根据用户携带的会话信息,更新相关记录表,调用声纹引擎对上传的语音进行语音识别以及质量检测,后调用存储服务,对声纹信息、语音等进行存储.最后是进行声纹验证,调用声纹识别引擎与之前训练的模型进行比对,给出验证得分,此时系统会根据当前生

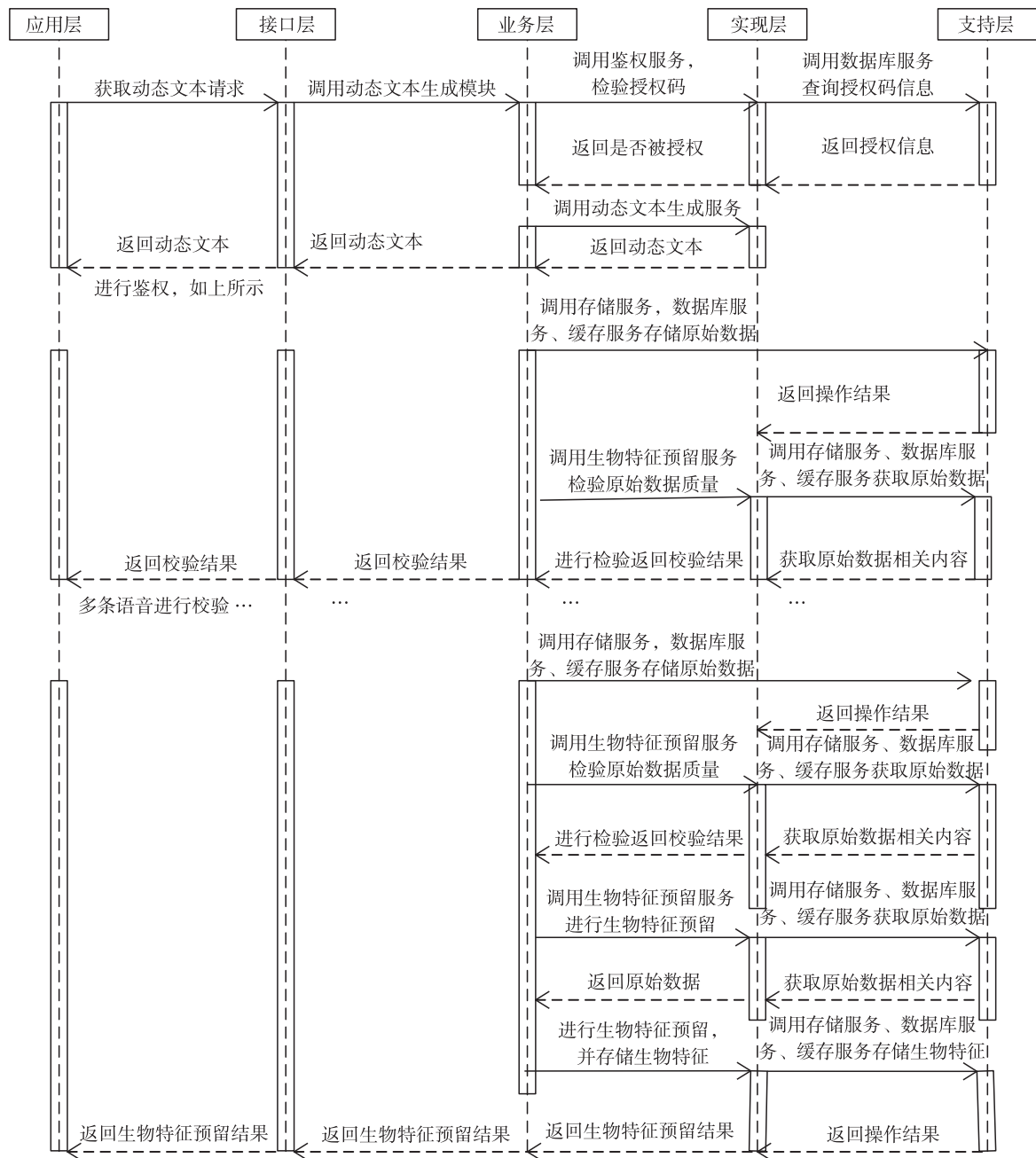


图 4 声纹预留过程序列图

Fig. 4 Sequence diagram of voiceprint recognition reservation process

产环境的安全等级(安全等级越高,要求说话人的声纹模型匹配度要更高)进行比对,得出当前验证是否通过的结果.在一般的业务流程中,通常会把声纹模型的验证打分与上传验证语音的步骤整合.中间过程中会涉及到交易信息的记录管理、缓存服务的调用以提高数据库的操作效率、日志的打印收集、异常信息的处理以及与服务监控模块的交互等等.可见,在实际的生产环境中,不管是声纹预留还是声纹验证,除了声纹识别引擎算法外,相关的业务、功能、控制模块也都是不可或缺的部分.概要流程图如图 5 所示.

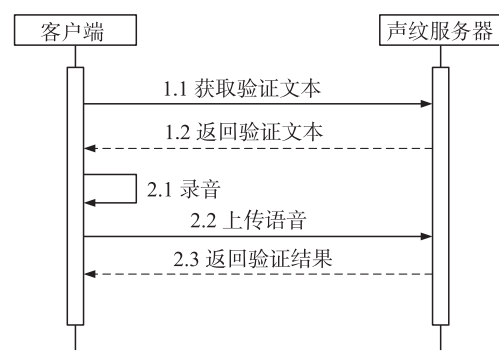


图 5 声纹验证过程流程图

Fig. 5 The flow chart voiceprint verification process

5 结语

本文系统地阐述了声纹识别所需的相关理论知识,特征提取和模式匹配,并且介绍了声纹预留和声纹验证的过程和原理.在此基础上,引出了在实际生产环境中,声纹识别系统应该包含的模块以及各模块的功能分析.最后给出了一个支持分布式的、高可用的声纹识别系统的架构设计.

[参考文献](References)

- [1] 郑凯鹏,周萍,张上鑫,等. 基于倒谱分量的融合参数应用于声纹识别[J]. 微电子学与计算机,2017,34(8):29-32.
ZHENG K P,ZHOU P,ZHANG S X,et al. Base on the fusion parameters of cepstrum components used in speaker recognition[J]. Microelectronics & computer,2017,34(8):29-32.(in Chinese)
- [2] OMER A E. Joint MFCC-and-vector quantization based text-independent speaker recognition system [C]//International Conference on Communication,Control,Computing and Electronics Engineering. IEEE,2017:1-6.
- [3] BOUSSAID L,HASSINE M. Arabic isolated word recognition system using hybrid feature extraction techniques and neural network[J]. International journal of speech technology,2018,21(1):29-37.
- [4] DIXIT A,VIDWANS A,SHARMA P. Improved MFCC and LPC algorithm for bundelkhandi isolated digit speech recognition[C]//International Conference on Electrical,Electronics,and Optimization Techniques. IEEE,2016:3755-3759.
- [5] AL-THAHAB O Q J. Speech recognition based Radon-Discrete Cosine Transforms by Delta Neural Network learning rule[C]//International Symposium on Fundamentals of Electrical Engineering. IEEE,2017:1-6.
- [6] BANERJEE A,DUBEY A,MENON A,et al. Speaker recognition using deep belief networks[EB/OL]. [2018-11-23]. <https://arxiv.org/ftp/arxiv/papers/1805/1805.08865.pdf>.

[责任编辑:陈 庆]