面向数字版权管理的声纹辅助认证系统

陈龙[1]，吴志勇[1]，袁春[1]，蒙美玲[1,2]，蔡莲红[1]

（1. 清华大学深圳研究生院，清华大学-香港中文大学媒体科学、技术与系统联合研究中心，深圳 518055；
2. 香港中文大学系统工程与工程管理学系，香港）

摘　要：本文针对传统数字版权管理系统中存在的由于密钥容易丢失和遗忘而造成用户使用不便、非法用户对密钥的窃取和伪造、以及合法用户主动泄漏密钥造成信息提供商在知识版权方面权益的丧失等问题，构建了一种针对对等网络（Peer-to-Peer，P2P）数字版权管理的声纹辅助认证系统。该系统采用随机数字文本提示的方式，进行说话人确认以及基于语音内容的信息确认，并采用 SVM 模型进行融合判决；针对说话人语音内容不匹配时存在的误识率较高的问题，进一步提出了 SVM 模型加单字阈值级联融合判决的方法。实验表明，SVM 模型加单字阈值级联融合判决能有效降低误识别率，同时误拒绝率也有所降低。将该方法用于数字版权管理的声纹辅助认证中，一方面保证了系统的安全性（误识别率极低），同时也保证了系统的便利性（误拒绝率较低），有效地提高了系统的方便性、可靠性和适用性。

关键词：数字版权管理（DRM）；说话人确认（SV）；语音信息确认（VIV）；支持向量机（SVM）

中图分类号： TP 391；TN 912.34

A speaker and verbal information verification system for digital rights management

CHEN Long[1]， WU Zhiyong[1]， YUAN Chun[1]， MENG Helen[1,2]， CAI Lianhong[1]

**(1. Tsinghua-CUHK Joint Research Center for Media Sciences, Technologies and Systems, Graduate School at Shenzhen, Tsinghua University, Shenzhen 518055;
2. Human-Computer Communications Laboratory, The Chinese University of Hong Kong, Hong Kong)**

Abstract: In the conventional digital rights management (DRM) system, there may exist several issues related to identity identification: the authentication key might be lost, forgotten or stolen leading to the inconvenience or profit loss of the user; the legitimate user may share his/her own key to the public leading to the loss in intellectual property rights of the content providers. This paper proposes a voiceprint based biometric system to enhance the authentication strategy for DRM system by combining speaker verification (SV) and verbal information verification (VIV) technologies. In the system, text prompted speaker verification (SV) is utilized to check the identity of the speaker; while verbal information verification (VIV) is used to ensure the correctness of the spoken content; and support vector machine (SVM) model is adopted as the fusion method in making the final decision. Experimental results indicate that the proposed method can greatly reduce the false acceptance rate (FAR) while lower the false rejection rate (FRR) simultaneously. By adopting this method to identity authentication, the DRM system can not only ensure the security of the system (speech based voiceprint biometric is speaker dependent) but also provide the convenience to the user (speech is not easily lost, forgotten or stolen).

Keywords: Digital rights management (DRM); Speaker verification (SV); Verbal information verification (VIV); Support vector machine (SVM)

信息技术的发展给人类的信息活动，例如查找、复制、转载、传送、分发等提供了极大的方便；但与此同时，信息技术的进步也为非法的信息获取活动提供了可利用的工具。这就对信息提供商提出了更高的要求，即信息提供商不仅要能够提供丰富的媒体内容，同时也需要提供有效、便利的服务方式方便客户保护自己的隐私和利益，以及使用有效的信息技术对自己提供的信息进行知识版权保护。

数字版权管理（Digital Rights Management，DRM）是现阶段在媒体信息的知识版权保护方面运用最成功和最广泛的一种方法。DRM 方法对媒体信息进行加密，在没有获得解码证书的情况下，非法用户即使下载到媒体数据，媒体播放设备也无法进行媒体数据的解码，非法用户也因此不能正常收听和观看媒体内容。只有当合法用户的媒体终端得到授权许可，并运用许可证书进行媒体内容解码的情况下，用户才可以正常收听和观看媒体内容。因此，数字版权管理为知识产权和客户利益的保护提供了一种有效的手段。

基金项目：深圳市科技计划—深港创新圈项目，国家自然科学基金(60805008，60928005，61003094)，教育部博士点基金(200800031015)

作者简介：陈龙（1977-），男（汉），湖北。

通讯作者：吴志勇，副研究员，E-mail：zywu@sz.tsinghua.edu.cn

在 DRM 中，授权许可的获取是通过用户身份确认的方法来进行的。正确确认用户身份，是 DRM 系统至关重要的一环。但传统的 DRM 系统使用用户密钥的方法进行身份确认，即信息提供商提供给用户一系列的数字或文字密钥，用户在需要获取媒体内容时输入密钥，即可获得授权许可。但使用用户密钥的方法存在一些严重的缺陷，一方面是用户密钥容易丢失和遗忘，造成用户使用的不便；二是非法用户对密钥的窃取和伪造，可以造成用户隐私泄漏及利益丧失；三是合法用户人为地主动泄漏密钥，会造成信息提供商在知识版权方面权益的丧失。

提供一种较高可信度的身份认证方法是解决这些问题的根本途径。生物特征识别（Biometrics）是利用人体所固有的生理特征或行为特征进行个人身份辨认和确认的技术，例如指纹识别、脸像识别、虹膜识别、声纹识别等。应用生物特征具有的唯一性、不可复制性、且不会被遗忘和丢失、不易伪造或被盗、随身“携带”以及随时随地可用等优点，提供一种高安全级别的身份认证方法，可以针对 DRM 系统的用户在身份合法性认证方面出现的密钥存在容易丢失、伪造、遗忘，且无法区分合法的真正用户和非法取得密钥的冒充者等缺点，提供一种有效的解决方法。

在生物特征识别技术中，声纹识别运用人类语音作为判决特征。语音是人类最为自然的一种信息传送方式，并且语音输入设备造价低廉、最为常用，避免了其他生物识别技术需专用的、且造价昂贵的输入设备的弊端。由于其在通用性上的这些优越性，声纹识别成为一种运用最为自然且最为经济实用的生物特征认证方法。

本文运用声纹识别系统的这些优点，结合 P2P 数字版权管理系统，实现了一种基于声纹辅助认证的 P2P 网络数字版权管理系统。该系统不仅具有区分说话人与冒认人的功能；而且采用文本提示的语音信息确认技术，避免了非法用户录音回放的欺骗冒认问题；并基于支持向量机对两者进行融合判决，提高声纹认证系统的准确性。

本文安排如下，第一节介绍了声纹识别模块的系统功能需求分析；第二节介绍了声纹识别模块的系统实现，其中第一部分介绍了语音内容信息确认模块的系统实现，第二部分介绍了说话人确认模块，第三部分是将两者进行综合考虑的融合判决模块，最后给出了系统的实验过程及结果。

1 声纹辅助认证模块系统分析

P2P数字版权管理系统的特性要求系统在进行用户身份认证时，具有以下特点：

1. 对用户身份鉴别的唯一性：即不仅能够准确识别出合法的用户，而且能够准确避免非法用户的冒认；
2. 对于用户语音具有实时鉴别的要求：在系统实现上表现为对于语音内容具有区分性，用于避免非法用户采用录音回放来进行欺骗冒认问题。

根据以上功能要求，系统需要同时实现语音信息确认和说话人确认的功能，并在判决阶段综合考虑语音信息确认和说话人确认的结果，实时地保证说话人身份认证的准确性。

本文在进行声纹辅助认证的系统实现时，采用文本提示（Text Prompted）的说话人确认技术，并基于隐含马尔可夫模型（Hidden Markov Model，HMM），采用统一的HMM模型框架同时进行语音信息确认（Verbal Information Verification）以及说话人身份确认（Speaker Verification），并利用支持向量机（Support Vector Machine，SVM）将两者的判决结果进行融合判决（Decision Fusion），以提高声纹确认系统的准确性。

与传统的文本无关（Text Independent）的基于高斯混合模型（Gaussian Mixture Model，GMM）的说话人确认技术相比，本系统具有如下优势：

1）在说话人确认方面，相对于文本无关的非特定内容的语音系统，本系统采用文本提示的方式缩小了用户语音模型匹配的搜索范围，使系统具有更高更精确的说话人统计特征；

2）在语音信息确认方面，系统随机产生的文本内容提示用户录入语音，并采用语音内容确认技术，保证验证语音内容的实时性和准确性；同时，针对用户本人的语音模型进行语音内容的识别，相比于通用语音识别系统，能显著提高语音内容识别及确认的性能。

3）在识别模型方面，采用统一的基于隐含马尔可夫模型（HMM）的识别框架同时进行语音信息的确认和说话人身份的确认，保证了系统的计算复杂度性能。

2 声纹辅助认证模块系统实现

根据系统的要求，在进行系统实现时，主要围绕着说话人确认功能、语音内容确认功能、以及融合判决功能等对系统进行实现，本文在下文中分别对这三个模块进行了介绍。在实现这三个模块功能的同时，系统也考虑了如下需求：

1. 在实现说话人确认功能和语音内容确认功能时，文本提示内容是必要的输入信息。系统在实现时，提供文本提示的输出功能，提示用户根据文本信息进行录音。另外，为保

证语音内容信息的时效性，在验证阶段采用随机产生数字序列的方式产生提示文本。

2. 结合 DRM 系统的网络部署，系统采用客户端与服务器端的架构。为降低系统对网络传输的依赖性，客户端不直接传送 PCM 语音数据到服务器端进行模型训练，而是在进行特征提取后传送语音特征，以减少网络传输量，降低网络开销。综合以上考虑，系统的实现架构如图 1 所示。

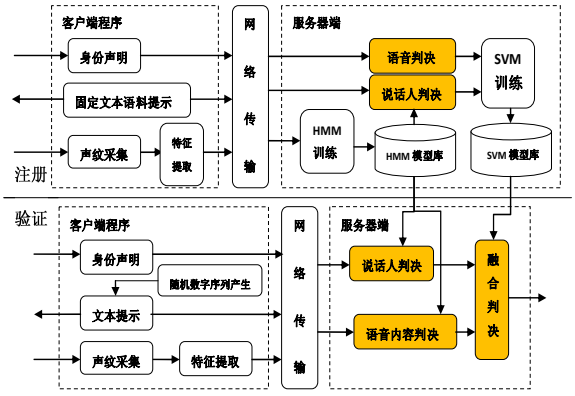

图 1　声纹认证模块系统结构图

2.1　语音信息确认

语音信息确认模块要求系统在进行每一次验证时，能够根据验证时间的变化，使用不同的语音内容进行身份确认。为了保证这项功能的实现，系统采用了

1）产生随机数字序列，提示用户根据此数字序列进行录音；

2）基于HMM模型进行语音信息的内容确认；

3）根据用户申明的身份使用与用户相关的HMM模型进行语言内容的确认，相比于通用的语音模型，能提高语音内容确认的准确性。

语言内容确认模块包括注册和确认两个阶段。在注册阶段，系统首先要求用户进行身份的声明。在对用户身份进行登记之后，客户端应用程序要求用户根据客户端界面上输出的文本进行语音录制。为了模型训练的需要，注册阶段所采用的录音文本，是综合考虑各单字语音训练次数的需要、以及相互之间连读的影响而预先设计好的。声纹采集模块对输入语音进行采样量化（采样频率44kHz，量化精度16bit，单声道），并存储为PCM格式的语音数据。特征提取模块从语音数据中提取语音特征（本系统采用Mel倒谱系数MFCC[2]及其1阶、2阶差分作为声纹特征）。通过网络传输模块，客户端应用程序将提取出来的语音特征传输到认证服务器。服务器端的模型训练模块对传送来的语音特征利用Baum-Welch算法[3]，根据文本提示的内容进行隐含马尔可夫模型HMM的嵌入式训练，进而对单字模型进行统计训练，得到跟当前所申明用户身份对应的、说话人相关的、单字HMM模型。最后，系统将训练好的说话人语音模型存储到说话人的语音模型库中。

在验证阶段，为了语音内容确认的实时性，客户端应用程序首先生成随时间改变的、随机的数字验证码文本，要求用户按照此提示文本进行录音。经过声纹采集模块以及特征提取模块对用户录制的语音数据进行处理后，客户端通过网络传输模块将语音特征及用户声明的身份信息传送到服务器端进行身份验证。服务器端的判决模块根据接收到的用户声明身份，从模型库中调出用户的语音模型，对需要验证的语音特征基于HMM模型运用Viterbi算法[4]进行解码，得到语音特征对该说话人模型的单个数字的先验概率$P(\boldsymbol{O}_i|S_i)$，并用公式(1)[5]得到语音特征序列的各单个数字后验概率$P(S_i|\boldsymbol{O}_i)$，作为最后判决的依据。

$$\log P(S_i \mid \boldsymbol{O}_i) = \log[P(\boldsymbol{O}_i \mid S_i)] - \log\{\frac{1}{n}\sum_{i=1}^{n}\exp[\gamma \log(P(\boldsymbol{O}_i \mid S_{in}))]\}^{1/\gamma} \quad (1)$$

式中，$\log P(S_i|\boldsymbol{O}_i)$代表观察向量$\boldsymbol{O}_i$针对于模型$S_i$的后验概率；$\log P(\boldsymbol{O}_i|S_i)$代表观察向量$\boldsymbol{O}_i$针对于模型$S_i$的先验概率；$\log P(\boldsymbol{O}_i|S_{in})$代表观察向量$\boldsymbol{O}_i$针对于$S_i$模型的反模型$S_{in}$的先验概率，这里的反模型选择了非语音字之外的其他数字；n代表反模型的个数；参数γ用来调整各个反模型对后验概率的影响。

2.2　说话人确认

说话人确认模块需要实现说话人辨认区分功能，系统在实现时采用HMM作为说话人语音的统计模型。这样的好处是：

1）针对说话人建立的单字语言模型，可以降低语音向量统计范围，使统计数据更加精确，相比于通用的GMM方法，能够得到更精确的概率得分；

2）与语音信息确认模块所需的模型保持一致，能够有效降低系统训练和识别时的计算复杂度。说话人识别模块也包括注册和验证两个子模块，其基本操作和语音信息确认模块的操作相同，其不同点是后验概率的计算。

在得到语音特征对说话人模型的先验概率后，与语音信息确认时与其他非当前语音内容的模型进行比较以计算后验概率不同，说话人确认的比较对象是非当前说话人的模型。一般的方法是训练一个通用背景模型（Universal Background Model，UBM），用于所有说话人后验概率的计算。训练通用背景模型时，假设系统共有N个非说话人数据，每个人具有语音段每个人具有语音段$\boldsymbol{O}_i$, i=1…M，全局模型M_{UBM}[6]就是根据这M×N段说话人的语音

数据$\boldsymbol{O}_i$分别训练各数字的UBM，最大化下面公式似然概率得到最后的UBM。

$$\prod_{i=1}^{N} P(\boldsymbol{O}_i \mid M_{UBM}) \tag{2}$$

在计算后验概率时，首先计算语音特征针对说话人模型和UBM背景模型的先验概率，再运用贝叶斯公式，将先验概率转换为后验概率作为判决的依据，下面的公式(3)[6]对语音特征向量$\boldsymbol{O}_i$进行的后验概率的计算。#

$$\begin{aligned}&\log P(M_{spea\ker} \mid \boldsymbol{O}) = \\ &\log P(\boldsymbol{O} \mid M_{spea\ker}) - \log P(\boldsymbol{O} \mid M_{UBM})\end{aligned} \tag{3}$$

式中，$\log P(M_{speaker}|\boldsymbol{O})$代表观察向量$\boldsymbol{O}$针对于模型$M_{speaker}$的后验概率；$\log P(\boldsymbol{O}|M_{speaker})$代表观察向量$\boldsymbol{O}$针对于模型$M_{speaker}$的先验概率；$\log P(\boldsymbol{O}|M_{UBM})$代表观察向量$\boldsymbol{O}$针对于$M_{UBM}$模型的先验概率。

2.3 系统的融合判决

概率统计方法中的判决一般情况下采用最小错误率的方法选择出错误率最小的阈值。当测试语音的概率得分低于此阈值时，判决输入的语音特征不属于此类，而大于该阈值时，则判断输入的语音特征属于此类。基于这个阈值设定原则，考虑到声纹辅助系统在进行判决时，需要同时考虑说话人确认和语音内容的确认两项指标。所以系统首先设定说话人概率得分的最小错误率阈值θ_{SV}，以及语音概率得分的最小错误率阈值θ_{VIV}，并根据这两个阈值进行级联判断，作为最终的判决准则，即使用公式(4)进行判决。

$$\begin{cases}\boldsymbol{O} \subseteq \text{说话人说指定语音内容}, \\ \quad \text{if } P(S/O) > \theta_{SV} \text{ and } P(M/O) > \theta_{VIV} \\ \boldsymbol{O} \not\subset \text{说话人说指定语音内容}, \text{ if else}\end{cases} \tag{4}$$

但基于阈值级联判断的方法存在着不足：

1. 阈值设置鲁棒性不强，容易受到训练数据与测试数据概率波动的影响；
2. 没有充分考虑到两个概率得分之间的相关性，例如正例样本中会存在说话人确认得分小于θ_{SV}而语音信息确认得分大于θ_{VIV}（反之亦然）的情况。

支持向量机（SVM）[7]是基于结构风险最小化原理实现的，能够同时优化经验风险和模型复杂度，在解决有限样本学习问题具有优异的性能。但 SVM 也存在着对训练数据集的数量以及正反例的数量均衡有较高要求的弊端。

系统在实现时，首先运用 SVM 对冒认人语音以及语音内容不匹配的说话人语音进行识别，然后运用阈值设定方法对语音中的单字语音进行读音正确与否的判断。这样的级联方法不仅有效的利用了 SVM 的鲁棒性来区分冒认人语音以及语音内容完全不匹配的说话人语音；使用阈值判决方法对单字语音进行判断也能够避免训练数据量较小的问题，而且该判决方法也能够减少错误语音的误识别率。

3 实验

3.1 实验数据

3.1.1 实验数据集

实验数据为10个录音人的共400句连续的数字串语料，每个数字串的长度为 10 个数字单字。在语料设计上，每个录音人的 40 句语音语料，包括如下 3 个子集：

1) 子集 1：数字单字语料，每一句话以不同次序包含 0-9 共 10 个数字，共 10 句；
2) 子集 2：连读数字语料，考虑数字连读（如相同数字重复、协同发音影响）的情况，共 10 句；
3) 子集 3：随机数字语料，随机产生的数字排列，共 20 句。

3.1.2 实验数据规划

将每个说话人语音录音数据划分为 HMM 训练数据、SVM 训练数据、和测试数据 3 部分进行实验，数据的使用和划分如下：

1) HMM 训练：使用说话人本人语音数据子集 1 和子集 2 训练说话人的 HMM 模型；
2) SVM 训练：使用子集 3 中的 10 句语料，建立 SVM 训练所需的正例集合和反例集合。正例集合、和反例集合中的所有数据，通过训练好的说话人 HMM 模型，计算出说话人确认和语音信息确认的二维概率得分，作为训练 SVM 的正例和反例。定义如下：

- 正例集：说话人语音通过语音内容匹配的说话人 HMM 计算的概率得分；
- 反例集 1：冒认人语音通过语音内容匹配的说话人 HMM 计算的概率得分；
- 反例集 2：说话人语音通过部分内容不匹配的说话人 HMM 计算的概率得分；
- 反例集 3：冒认人语音通过部分内容不匹配的说话人 HMM 计算的概率得分；
- 反例集 4：说话人语音通过内容完全不匹配的说话人 HMM 计算的概率得分；
- 反例集 5：冒认人语音通过内容完全不匹配的说话人 HMM 计算的概率得分。

其中，反例集的非说话人语音为其他非说话人数据子集 3 中的前 10 句。部分内容不匹配语音数据考虑了语音与模型出现 1 个字到 9 个字不匹配的情况，实验中采取替换 1-9 个与实际语音内

容不同的 HMM 模型来达到数据不匹配的目的。实验数据组合分配如表 1 所示。

表 1 实验数据集合

	正确语音模型	部分不匹配语音模型	完全不匹配语音模型
说话人语音	正例集	反例集 2	反例集 4
冒认人语音	反例集 1	反例集 3	反例集 5

实验从考察语句中单字的数量对系统性能的影响这个角度出发，并不将单字的概率得分输入到 SVM，而是从每一句语音的 10 个单字中依次选取若干个单字求均值，作为实验的数据。求均值的方法使用公式(5)，通过设置 S 不同值，用循环求均值的方法得到含有不同长度单字的语音均值。

$$Mean_{si} = \frac{W_i + W_{(i+1) \bmod 10} + \ldots + W_{(i+S-1) \bmod 10}}{S} \quad (5)$$

$$i = 0 \cdots 9, S = 2 \cdots 9$$

式中 W_i 代表语句中的单字，S 单字的长度。本次实验选择了 S=2…9 字长循环求均值，以及整句语音的 10 个字长求均值，共 9 种字长的组合作为考察数据。

3) 测试：使用的正例和反例采取与训练 SVM 时相同的数据划分及处理方法，而测试的语音是数据语料子集 3 中剩下的 10 句。

3.2 实验评测标准

实验采用在统计学习方法中常用误拒绝率和误识别率[8]来衡量系统的性能。

误拒绝率（False Rejection Rate）定义为公式（6）

$$FRR = \sum_n Rf \Big/ \sum_n PE \quad (6)$$

误识别率（False Acceptance Rate）定义为公式（7）

$$FAR = \sum_n Af \Big/ \sum_n NE \quad (7)$$

式中，n 代表进行测试的说话人的总数，本次实验是 10 个；$\Sigma_n PE$ 代表正例样本的总个数；$\Sigma_n NE$ 代表反例样本的总个数；$\Sigma_n Rf$ 代表正例样本中误拒绝的总个数；$\Sigma_n Af$ 代表反例样本中误识别的总个数。

3.3 实验结果

3.3.1 实验 1

实验 1 考察了字长对系统性能的影响。实验时，考虑到训练集的影响，共选用了 2 个不同的训练集训练 SVM，针对 2-9 字长求均值的正反例测试数据集进行实验， 2 种训练集是：

1) 训练集 1：对正反例样本 2 字长循环求均值；

2) 训练集 2：对正反例样本 9 字长循环求均值；

实验结果表 2 中显示了正例集误拒绝率（FRR），反例集 1 和反例集 2 的误识别率（FAR）随字长增加的情况。

表 2 SVM 实验结果

字长	训练集 1			训练集 2		
	正例集	反例集 1	反例集 2	正例集	反例集 1	反例集 2
	FRR	FAR	FAR	FRR	FAR	FAR
2	27.2%	18%	17.7%	69.8%	7.5%	6.4%
3	30.4%	12.5%	13.8%	57%	7%	5.1%
4	26%	11%	13.9%	48.5%	6%	4.9%
5	24.7%	9%	15.1%	42.3%	5.5%	5.9%
6	22.9%	6%	15.3%	36.4%	4.5%	6.3%
7	23.2%	4%	16%	33.6%	3.5%	6.4%
8	21.3%	2%	16.9%	30.25%	1%	6.7%
9	21.3%	0%	16.9%	28%	0.5%	7.1%

实验结果显示出，随着字长的增加，正例集的误拒绝率会逐步下降，反例集 1 误识别率趋近于 0，但反例数据集 2 的误识别率会略有增加。

3.3.2 实验 2

实验 2 考察了反例集 2 误识别率的分布情况，分别以内容不匹配的说话人语音中包含不匹配模型的个数为统计点，统计了误识别率分布情况。统计结果如图 2 所示。

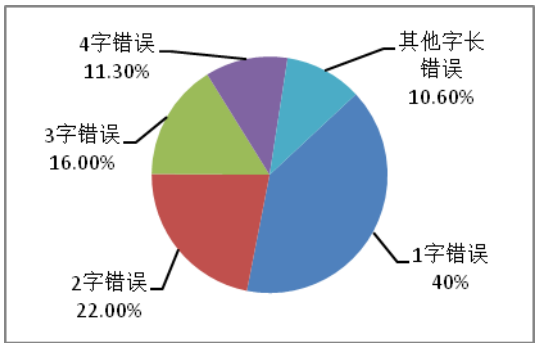

图 2 反例集 2 针对不匹配模型个数的误识别率分布

图 2 显示，随着测试语料中包含不匹配语音数量增加，误识别率逐步降低。其中包含 1 个不匹配模型的误识别率占总误识别率的 40%，而包含 5-9 字不匹配模型的误识别率仅占 10.6%。

针对上面的结果，我们继续考察了字长与所含不匹配模型的关系，图 3 分别统计了 2-9 个字长的测试语料中包含了 1 个不匹配语音模型的情况下误识别的概率。

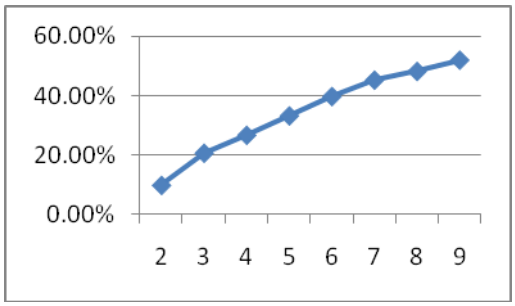

图 3　包含 1 个不匹配模型的
2-9 字长均值数据的误识别错误率分布

统计显示，当循环求均值的字长越长，包含不匹配模型的语料判决错误的比例越高。字长为 9 的情况下，判决错误的比例达到 52%。根据上面的两个统计结果，我们认为当求均值的字长较长，而不匹配的语音模型数量少时，测试语料容易出现错误接受的情况。其原因是因为当匹配模型的数量较多时，其匹配模型的概率得分会对最后的均值得分作用较大，而不匹配模型的得分对均值得分的作用就被掩盖掉，从而增加了误识别的情况。

3.3.3　实验 3

根据实验 1、2 的结果，随着字长的增加，以整句话的概率均值进行判决时，系统可以很好的区分冒认人语音以及语音内容完全不匹配的说话人语音；但对于部分内容不匹配的说话人语音，误识别率随着不匹配模型数量的减少而增加。这显示出，系统在单字语音判决方面的性能不足。

针对这一不足，实验 3 加入了对单字语音概率值的考虑，其目的是以文本无关的语音内容识别的结果来辅助进行语音信息确认。实验判决分两部分：

1. 以当前说话人的 HMM 模型进行文本无关的语音内容识别，判断语音识别结果是否与当前的提示文本相同，如果相同则判断输入语音为正确语音内容；
2. 如果不同，将语音识别结果作为反模型，按照公式（1）得到语音的后验概率。在训练阶段，根据后验概率设定每个单字的判决阈值；在测试阶段，当且仅当每个单字的后验概率均大于相应的阈值时，才判断输入语音为正确语音内容。

实验 3 的训练集及测试集的数据设置如下：

1. 训练集：说话人语音子集 3 中前 10 句组合为正例集，该 10 句中每一句依次替换 1 个不匹配模型组成反例集，对单字语音阈值进行训练；
2. 测试集：说话人语音子集 3 中后 10 句组合为正例集，该 10 句中每一句依次替换 1 个不匹配模型组成反例集进行测试。

表 3 显示了基于该方法进行语音内容判决（即语音信息确认）的实验结果，计算各单字的误拒绝率和误识别率。

表 3　单字判决实验结果

	正例集	反例集 2	
	FRR	FRR	FAR
阈值判决	1.4%	8%	16%

3.3.4　实验 4

实验 3 的结果显示出该方法能够较好的进行单字语音内容的判决。针对上述实验 1 和 2 中反例集 2 的误识率较高的问题，系统考虑使用 SVM 判断和单字阈值判决进行级联的方法进行融合判决。

1. 系统首先使用 SVM 对输入语音取 10 字均值概率进行文本提示的说话人确认和语音信息确认的融合判决。若判决结果为拒绝（即判断为内容不匹配或冒认人的语音），则接受该结果；
2. 否则，针对 SVM 判决为内容正确的说话人的语音，使用单字阈值判决的方法判决语句中错误语音单字的数量，如果错误单字的数量大于 1，即拒绝该语音；否则接受语音。

实验 4 使用说话人确认、语音信息确认、SVM 融合判决、以及 SVM 加单字阈值级联融合判决的方法分别进行了实验。其中说话人确认、语音信息确认的阈值设定使用最小错误率的方法；实验结果如表 4 所示。

表 4　SVM 实验结果与其他实验方法结果对比

判决模块	误拒绝率	误识别率		
	正例集	反例集 1	反例集 2	反例集 3、4、5
说话人确认	9%	0.7%	32%	0.1%
语音信息确认	6%	22.2%	23.8%	0.3%
SVM 融合	28%	0%	7.1%	0%
级联融合	5%	0%	0%	0%

实验结果显示，相比于说话人确认、语音信息确认、以及 SVM 融合判决，使用级联融合判决时，部分内容不匹配的说话人语音（反例集 2）的误识别率明显降低，在内容匹配说话人语音（正例集）的误拒绝率方面，性能也有所提升。

4　结论

本文针对传统的数字版权系统中的密钥使用方法，存在的容易遗忘、丢失、以及容易泄密等问题，利用生物特征识别中的语音及说话人识别方法，构建了一种面向P2P网络数字版权管理的声纹辅助认证系统。该系统采用随机数字文本提示的方式，进行说话人确认以及基于语音内容的信息确

认，并采用SVM模型进行融合判决；针对说话人语音内容不匹配时存在的误识率较高的问题，进一步提出了SVM模型加单字阈值级联融合判决的方法。实验表明，SVM模型加单字阈值级联融合判决能有效降低误识别率，同时在误拒绝率方面性能也有所提升。在实际应用中，该方法既保证了系统的安全性（误识别率极低），同时也保证了系统的便利性（误拒绝率较低）。

参 考 文 献

[1] L.R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition [C]. // Proc IEEE, 77(2), 1989: 257-286.

[2] P. Mermelstein. Distance measures for speech recognition, psychological and instrumental [C]. // Proc Pattern Recognition and Artificial Intelligence, New York: Academic Press, 1976: 374-388.

[3] L.E. Baum, T. Petrie, G. Soules, N. Weiss. A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains [J]. *Ann. Math. Statist.*, 1970, 41(1): 164-171.

[4] A.J. Viterbi. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm [J]. *IEEE Trans. Information Theory*, 1967, 13(2): 260-269.

[5] Q. Li, B. Juang. Automatic verbal information. verification for user authentication [J]. *IEEE Trans. Speech. Audio Processing*, 2000, 8(5): 585-596.

[6] D. Reynolds, R. Rose. Robust text-independent speaker identification using Gaussian mixture speaker models [J]. *IEEE Trans. Speech Audio Process*, 1995, 3(1): 72-83.

[7] C. Cortes, V. Vapnik. Support-vector networks [J]. *Machine Learning*, 1995, 20(3): 273-197.

[8] C.J. van Rigsbergen. Information Retrieval [M]. London: Butterworths, 1979.