声纹识别技术及其应用现状

郑方[1]*，李蓝天[1]，张慧[2]，艾斯卡尔·肉孜[1]

[1]清华信息科学技术国家实验室技术创新和开发部语音和语言技术中心
清华大学信息技术研究院语音和语言技术中心
清华大学计算机科学与技术系 北京市 100084
[2]贵州大学科技学院 贵阳市 550001

{lilt, askar}@cslt.riit.tsinghua.edu.cn, hebe.hui.zhang@gmail.com
*通讯作者：郑方，教授，E-mail：fzheng@tsinghua.edu.cn

摘要：随着信息技术的快速发展，如何准确认证一个人的身份、保护个人隐私和保障信息安全，成为当前亟需解决的问题。与传统身份认证方式相比，生物特征识别身份认证技术在使用过程中具有不会丢失、被盗或遗忘的特性；其不但快捷、方便，而且准确、可靠。声纹识别作为当前最热门的生物特征识别技术之一，在远程认证等应用领域中具有独特优势，受到了越来越多的关注。文章将以声纹识别技术及其应用现状为主线，依次介绍了声纹识别的基本概念、发展历程、应用现状及其行业标准化现状；综述了声纹识别所面临的各类问题及其解决方案；最后对声纹识别技术以及应用的发展前景进行展望。
关键字：生物特征识别、身份认证、声纹识别、发展历程、技术应用
中图法分类号：TP391.4

引言

在我国古代战乱时期，官兵进出城池通过对照预先设定的口令判断是敌是友；在现实生活中，我们每天通过钥匙或电子卡进出家门；在上网过程中，用户的账户和密码是登录某网站或某邮箱进行下一步操作的渠道；在一些电子支付中，通过发送验证码到用户手机让其输入进行支付确认……。上述提到的口令、钥匙、电子卡、账户密码以及支付验证码都代表了认证操作者的相关信息，但在科技和互联网迅速发展的今天，传统身份认证显然已不能满足用户对个人信息及财产保护的需求。口令易被泄露、钥匙或电子卡易丢失和被复制、账户和密码易被遗忘和攻击、验证码易被截取等一系列的安全隐患所带来的事故时有发生。因此，传统认证方式将逐渐成为历史，而生物特征认证方式将闪亮登场。

在“无处不账户、无处不密码”的时代，人们常常因为遗忘或丢失密码而感到烦恼，生物特征认证技术的出现无疑带来了更便捷、高效的服务方式。生物特征可“随身携带”，可随时随地使用人们身上的生物特征来对自己的身份进行认证，是“用自己来识别自己”的一门技术。显然，在使用生物特征的认证过程中，人们无需担心遗忘、丢失。此外，生物特征还有防伪性好、不易被改造和窃取等优点。有关新闻报道和预测指出，未来明文密码将成为历史，各种口令也将不复存在，生物特征认证技术将取而代之。

生物特征可分为两类，包括生理特征和行为特征[1]。常见的生理特征有指纹、掌纹、人脸、虹膜、视网膜等；常见的行为特征有声纹、签名、心跳等。除此之外，生物特征还有手形、步态、DNA、味纹、血管纹路等。生物特征代表着每个人所固有的特点，它具有普遍性、唯一性、稳定性、不易复制性等，但在实际应用中，这些生物特征的认证都有其

一定的局限性。例如手指、手掌蜕皮或磨损的情况下会使身份认证辨认度降低；不法分子通过戴指模躲过指纹认证系统掩盖其真实身份以逃避司法追究；虹膜识别技术需要昂贵的摄像头聚焦和较好的光源；视网膜识别技术要求激光照射眼球的背面以获取视网膜特征的唯一性，故可能会影响使用者的健康，并且视网膜身份认证技术的使用性不是很好，研究成本也高；而在 2015 年 10 月 21 日新闻报道中，“赵薇‘被卖房’”案件表明人脸识别系统也存在一定的风险性。

与其它生物特征相比，作为行为特征的声纹具有以下特点[2]：

- 蕴含声纹特征的语音获取方便、自然，在采集过程中涉及到的用户个人隐私信息较少，因此使用者更易接受；
- 语音采集装置成本低廉，使用简单，一个麦克风即可，在使用通讯设备(如电话、手机)时更无需额外的录音设备；
- 配合语音识别技术，可使声纹口令动态变化而无需担心密码遗忘、丢失和窃取问题，防止录音假冒，因此也尤为适合远程身份认证 。

为此，本文以声纹识别技术为主线，介绍声纹识别技术的基本概念、应用领域以及声纹识别在实际应用中所存在的问题与对应解决方案；最后，展望声纹识别技术在身份认证领域的发展前景。

1 声纹识别基本概念

在日常生活中，我们时时刻刻都在从外界接受和向外界传达着各种信息，语音信息则是其中重要的一部分。在语音领域中，人的语音通常被定义为人的发音器官所发出的、带有一定实际含义的声音，也常常被研究者认为是语言的发音符号。音频信号的处理在人工智能和机器学习领域研究中具有很重要的地位。人类语音中含有各类丰富的信息，既有丰富的说话人个性信息和发音的内容信息，也有录制环境的噪声信息、信道信息等等。

声纹其实就是对语音中所蕴含的、能表征和标识说话人的语音特征，以及基于这些特征（参数）所建立的语音模型的总称[3]，而声纹识别是根据待识别语音的声纹特征识别该段语音所对应的说话人的过程[3]。与指纹类似，每个人在说话过程中所蕴含的语音特征和发音习惯几乎是独一无二的，就算被模仿，也改变不了话者最本质的发音特性和声道特征。有相关科学研究表明，声纹具有特定性和稳定性等特点，尤其在成年之后，可以在相对长的时间里保持相对稳定不变。声纹是一种行为特征，由于每个人在讲话时使用的发声器官如舌头、牙齿、口腔、声带、肺、鼻腔等在尺寸和形态方面有所差异，以及年龄、性格、语言习惯等多种原因，加之发音容量大小和发音频率不同，在发音时千姿百态，因而导致这些器官发出的声音必然有着各自的特点。可以说，任何两个人的声纹图谱都不尽相同。

声纹识别技术又称说话人识别技术，就是基于这些信息来探索人类身份的一种生物特征识别技术。这种技术基于语音中所包含的说话人特有的个性信息，利用计算机以及现在的信息识别技术，自动的鉴别当前语音对应的说话人身份[4-5]。声纹识别与语音识别不同，声纹识别的过程是试图找到区别每个人的个性特征，而语音识别则是侧重于对话者所表述的内容进行区分。在实际应用中，往往把语音识别技术和声纹识别技术结合起来应用，以提高声纹身份认证系统的安全性能。

声纹识别是一类典型的模式识别问题，其主要包含了说话人模型训练和测试语音识别两个阶段，下图 1 是一个基本的说话人识别框架：

- 训练阶段：对使用系统的说话人预留充足的语音，并对不同说话人语音提取声学特征，然后根据每个说话人的语音特征训练得到对应的说话人模型，最终将全体说话人模型集合在一起组成系统的说话人模型库；

- 识别阶段：说话人进行识别认证时，系统对识别语音进行相同的特征提取过程，并将语音特征与说话人模型库进行比对，得到对应说话人模型的相似性打分，最终根据识别打分判别得到识别语音的说话人身份。

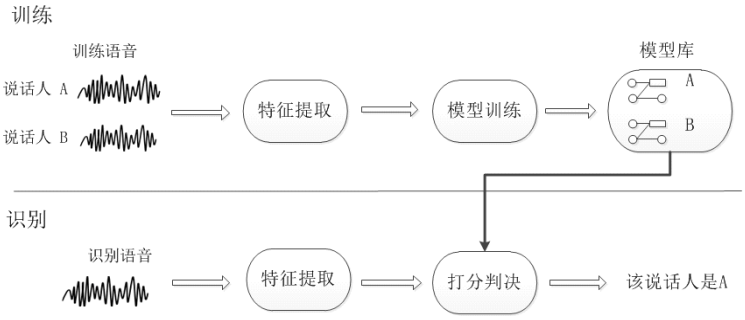

图 1 一个基本的说话人识别系统框架

1.1 声纹识别的分类

按照不同的分类角度，声纹识别可大致分为以下几类。

1.1.1 声纹辨认和声纹确认

声纹识别根据实际应用的范畴可分为声纹辨认和声纹确认[5]。这两类任务的识别目标略有不同。声纹辨认是指判定待测试语音属于目标说话人模型集合中哪一个人，是一个“多选一”的选择问题。而声纹确认是确定待识别的一段语音是否来自其所声明的目标说话人，是一个“一对一”的判决问题。

对于声纹识别辨认来说，根据测试识别来自说话人范围的不同，说话人辨认又可划分为闭集识别和开集识别[5]。闭集识别是指待测试语音必定属于目标说话人集合中的某一位，即待识别语音为集内说话人；所谓开集识别，是指待识别语音的发音者可能不属于目标说话人集合中的任何一位。

除此之外，根据实际应用场景，声纹识别还包括说话人检测(即检测目标说话人是否在某段语音中出现)和说话人追踪(即以时间为索引，实时检测每段语音所对应的说话人)[6]。

1.1.2 文本相关、文本无关和文本提示

按照待识别语音的文本内容，声纹识别可分为文本无关(text-independent)、文本相关(text-dependent)和文本提示(text-prompted)[5]三种。文本无关是指说话人识别系统对于语音文本内容无任何要求，说话人的发音内容不会被预先限定，在训练和识别阶段说话人只需要随意的录制达到一定长度的语音即可；而文本相关是指说话人识别系统要求用户必须按照事先指定的文本内容进行发音。对比这两类说话人识别，文本相关的说话人识别的语音内容匹配性优于文本无关的说话人识别，所以一般来说其系统性能也会相对好很多，但是对说话人预留和进行识别时语音的录制要求更高并且识别文本易于窃取；而文本无关的说话人识别使用更加方便灵活，具有更好的推广性和适应性。

为此，综合二者的优点，文本提示型的说话人识别应运而生，其是指识别系统从说话人的训练文本库中随机提取若干词汇组合后提示用户发音，这样不仅避免了文本相关的假冒录音闯入，并且实现相对简单、安全性高，是说话人识别技术的一大热点。

1.2 声纹识别的性能评价

根据声纹识别任务的不同，其系统性能的评价指标也略有不同。对于声纹确认系统，通常采用 DET 曲线、等错误率 EER 和检测代价函数 DCF；而声纹辨认系统则根据测试集

合的不同，选择不同的系统评价指标。

1.2.1 声纹确认系统性能指标

- DET 曲线及等错误率 EER[7]

声纹确认识别系统的性能评价主要看两个参量，分别是错误接受率(False Acceptation Rate, FAR)和错误拒绝率(False Rejection Rate, FRR)。FAR 是指将非目标说话人判别为目标说话人造成的错误。FRR 是指将目标说话人误识成非目标说话人造成的错误。二者的定义如下：

$$\text{错误接受率(FAR)} = \frac{\text{被接受的错误识别的语音样本数}}{\text{应被拒绝的语音样本总数}} \times 100\% \qquad \text{式 1}$$

$$\text{错误拒绝率(FRR)} = \frac{\text{被拒绝的正确识别的语音样本数}}{\text{应被接受的语音样本总数}} \times 100\% \qquad \text{式 2}$$

在声纹识别系统中，可通过设定不同的阈值对 FAR 和 FRR 进行权衡。系统所要求的安全性越高，则设定阈值应越高，此时接受条件就越严格，即 FAR 越低，但 FRR 越高；反之，如果系统追求较好的用户体验性（通过率高），则阈值应越低，此时接受条件就越宽松，FAR 就越高，但 FRR 越低。一般采用检测错误权衡曲线(Detection Error Trade-offs Curve, DET)来反映两个错误率之间的关系：对一个特定的声纹识别系统，以 FAR 为横坐标轴，以 FRR 为纵坐标轴，通过调整其参数得到的 FAR 与 FRR 之间关系的曲线图，就是 DET 曲线（参见图 2）。显然 DET 曲线离原点越近，系统性能越好。

在 DET 曲线上，第一象限角平分线与其的交点处，FAR 与 FRR 值相等，该错误率称为等错误率(Equal Error Rate, EER)。显然，等错误率 EER 值越小，系统性能应该越好，它代表了声纹识别系统的一个大约性能，是衡量系统性能的重要参数。

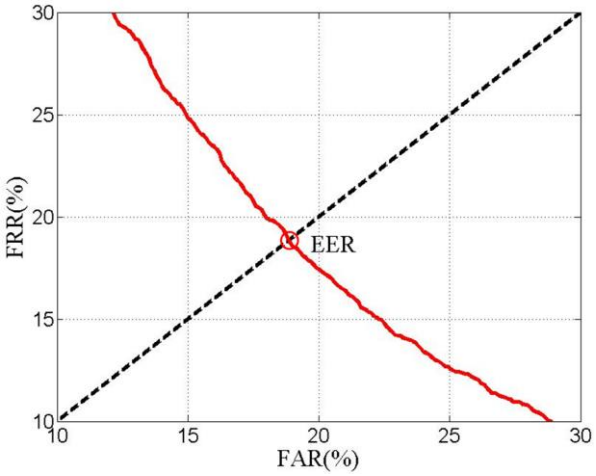

图 2 检测错误权衡曲线示例

- 检测代价函数 DCF (Detection Cost Function)[5]

在美国国家标准技术研究所(National Institute of Standards and Technology, NIST)的评测中，还定义了 FAR 和 FRR 的加权和函数，即检测代价函数（Detection Cost Function, DCF）作为系统性能的评价指标。DCF 的定义如下：

$$C_{DCF} = C_{Miss} \times FRR \times P_{target} + C_{FalseAlarm} \times FAR \times (1 - P_{target}) \qquad \text{式 3}$$

其中，C_{Miss}和$C_{FalseAlarm}$分别表示错误拒绝和错误接受的权重，表示目标说话人的先验概率。针对不同的应用场景，对 FAR 和 FRR 定义不同的权重，并用最小 DCF 即

$\min C_{DCF}$来表示系统能够取得的最优性能。

1.2.2 声纹辨认系统性能指标

通常情况下，在开集声纹辨认系统中仍可采用等错误率 EER 和检测代价函数 DCF 来评价系统性能指标；而在声纹闭集辨认系统中通常采用正确识别率(简称识别率)、错误识别率(简称为错误率)以及前 N 正确率(Top N Correctness)作为评价系统性能的指标。

识别率是指待识别语音能够从目标说话人集合中正确找到所对应说话人的概率。通常认定待识别语音与目标说话人集合中相似度最大的作为辨认说话人，其辨认正确的比率又可称为 Top-1 辨认正确率；若目标说话人集合中相似度最大的 N 个辨认说话人包含正确说话人时认为辨认正确，如此统计出来的辨认正确的比率称为 Top-N 辨认正确率。

1.3 声纹识别的发展历程

“闻其声而知其人”，通过人的听觉来判断说话人的声音具体来自哪一个人，古已有之。以语音作为身份认证的手段，最早可追溯到 17 世纪 60 年代英国查尔斯一世之死的案件审判中。对说话人识别的研究始于 20 世纪 30 年代[8]。自 1937 年的 C. A. Lindbergh 儿子被拐骗事件开始，人们针对语音中的说话人信息开展了科学的探索和研究。1945 年，Bell 实验室的 L. G. Kesta 等人借助肉眼观察，完成语谱图匹配，并首次提出了“声纹”的概念；且在 1962 年第一次介绍了采用此方法进行说话人识别的可能性。1966 年，美国法院的第一次采用“声纹”进行了取证。Bell 实验室的 S. Pruzanshy 提出的基于模板匹配(template matching)和统计方差分析的说话人识别方法[9]，引起信号处理领域许多学者的注意，兴起了说话人研究的高潮。1969 年 Luck JE 首先将倒谱技术用于说话人的识别，得到了较好的效果，BS Atal 将线性预测倒谱系数(Linear Predictive Cepstrum Coefficient, LPCC)[10]用于说话人识别，提高了识别系数的精度。Doddington 提出了利用共振峰进行说话人确认[11]，1972 年，Atal 用提出的基频轮廓进行说话人识别[12]。

从 20 世纪 70 年代末至 80 年代末，说话人识别的研究重点转向对声学特征参数的处理以及新的模式匹配方法上。研究者相继提出了 LPC 谱系数[13]、LSP 谱系数、感知线性预测系数(Perceptual Linear Predictive, PLP)[14]、梅尔倒谱系数(Mel-frequency Cepstrum Coefficient, MFCC)[15]等说话人识别特征参数。此时，动态时间规整法(Dynamic Time Warping, DTW)[16]、矢量量化法(Vector Quantization, VQ)[17]、隐马尔科夫模型(Hidden Markov Model, HMM)[18]、人工神经网络法(Artificial Neural Network, ANN)[19]等技术在语音识别领域得到了广泛的运用，也成为说话人识别的核心技术。

20 世纪 90 年代以后，尤其是 D. Reynolds 对高斯混合模型(Gaussian mixture model, GMM)[20]做了详细介绍后，GMM 以其简单、灵活、有效以及较好的鲁棒性，迅速成了目前与文本无关的说话人识别中的主流技术，将说话人识别研究带入一个新的阶段。2000 年，D. Reynolds 在说话人确认任务中提出了高斯混合模型-通用背景模型 GMM-UBM (Gaussian mixture model-Universal background model)[21]结构，为说话人识别从实验室走向实用作出了重要贡献。

进入 21 世纪，在传统 GMM-UBM 的方法上，P. Kenny、N. Dehak 等人先后提出了联合因子分析(Joint factor analysis, JFA)[22]和 i-vector 模型[23]，将说话人模型映射到低维子空间中，克服了 GMM-UBM 系统中高斯分量互相独立的局限性，提高了系统性能。为进一步提高模型的区分性能力，相关的区分性训练方法也应运而生。此外，随着深度机器学习在语音识别、图像处理等领域的快速发展和成功应用，近年来，基于深度学习的相关方法也逐渐应用到说话人识别中[24-25]，并取得了不俗的成效。

2 声纹识别的应用

声纹识别技术早已在西方许多国家开始应用，如：1998 年欧洲电信联盟应用声纹识别技术在电信与金融结合领域，完成了 cAvE 计划；2004 年美国最大的银行自动出纳机制造商 NCR 分部，开始试验自动出纳机的声纹核实效果。同年 5 月美国加利福尼亚州 Beep card 公司发明了一种带有特殊安全功能的信用卡，这种信用卡只有在识别出主人的声音后确认身份后才能正常操作。2006 年，荷兰的 ABN AMRO 银行率先使用了美国 Voice Vault 的声纹识别系统，借助预先录制的个人私密问题进行身份验证。目前在国外，声纹识别技术已经广泛应用到军事、国防、政府、金融等多个领域。

国内对声纹识别技术的研究起步稍晚于国外，但经过国内研究人员的共同努力，声纹识别技术在国内已经得到了较好的发展与应用。2011 年中国建设银行与北京得意公司合作，构建了基于说话人识别技术的声纹电话银行系统；2013 年 11 月，厦门天聪公司与厦门公安局指挥中心合作，搭建厦门“110”报警声纹采集与辅警系统。2013 年 12 月，北京得意公司与中大信通合作的社区矫正项目，利用声纹识别技术为深圳司法局提供服务。

根据实际应用范畴，本章从声纹辨认和确认等方面详细介绍声纹识别技术的应用，并总结相关的行业及国家标准。

2.1 声纹辨认技术领域

声纹辨认技术通常广泛应用于公安司法、军队国防领域中，如：刑侦破案、罪犯跟踪、国防监听等[5,26,27]；

- 监听跟踪

  恐怖分子在作案前后通常会与组织、同伙保持联系，通讯中可能会包含关键内容。因此，在通信系统或安全监测系统中预先安装声纹辨认系统，可通过通讯跟踪和声纹辨别技术对罪犯进行预防和侦查追捕。据悉，拉登的落网正是美国情报部门充分利用了声纹鉴别技术。此外，声纹辨认技术还用于对满刑释放的犯罪嫌疑人进行监听和跟踪，可有效阻止犯罪嫌疑人再次犯科，也利于对其进行及时逮捕。

- 国防安全

  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，继而对交谈内容进行跟踪(战场环境监听)；当通过电话发出军事指令时，可以对发出命令者进行身份辨认(敌我指战员鉴别)。目前该技术在国外军事方面已经有所应用。据报道，2001 年 4 月 1 日迫降在我国海南机场的美军 EP-3 侦察机就载有类似的声纹识别侦听模块。

- 公安技侦

  犯罪嫌疑人通过非法渠道到获取受害者的个人信息，通过电话勒索、绑架等刑事犯罪案件时有发生。如：2015 年 09 月 21 日，中国警察网新闻报道了一起电话“勒索‘消灾费’每天恐吓数百名学生家长”的案件；2015 年 11 月 19 日报道了富豪被绑架勒索的案件等。对于此类的刑事犯罪案件，公安司法人员可利用声纹辨认技术，从通话语音中锁定嫌疑犯人、减小刑侦范围。在车站、飞机、码头等公共安检点装入声纹辨认系统，可以有效对危险人物进行鉴别和提示，降低肉眼识别所带来的错误，提高人们生命财产的安全性。

2.2 声纹确认技术领域

随着互联网的快速发展，便捷的网上交易越来越受人们的亲睐，因而远程身份认证的安全性亟待加强。声纹确认技术可以满足网上交易、支付、远程身份认证的安全性需要，并已逐渐广泛应用于证券交易、银行交易、个人设备声控锁、汽车声控锁、公安取证、信

用卡识别等[5,26,27]。

- 电子支付

2014 年中国互联网支付用户调研报告显示，网上支付、手机支付、第三方支付已成为现代人购物付款的主流方式。显然，网络支付的安全性应当重视起来，网络支付的身份认证也愈发重要。近年来，有相关媒体接二连三地报道支付宝被盗刷、网银被转出等案件。为了防止这类案件的再次发生，将声纹确认技术加入到交易支付中，通过动态密码口令等方式进行个人身份认证，有效地提高了个人资金和交易支付的安全性。例如，荷兰 ABN AMRO 银行、澳大利亚国家银行 National 借助声纹识别系统实现用户身份认证；全球互联网支付系统的领导者 Voice Commerce Group 也于 2008 年推出了基于声纹识别的 Voice Pay 服务。目前在国内，由北京得意公司提供的声纹认证技术正在中国建设银行等领域推广使用。

- 声纹锁控

据媒体报道，近几年数以万计的腾讯 QQ 用户出现了账号被盗取的情况。盗号者通过联系用户的亲朋好友进行金钱诈骗，给用户及其亲友带来了严重的损失。为了避免这类事件再次发生，有必要将声纹认证代替明文密码认证。例如，微信已上线使用基于声纹动态口令的登录方式，极大提高了使用者账号的安全性。随着声纹认证技术的成熟，相信声纹控锁技术将被广泛地应用在各类账户声控密码锁、电脑声控锁、汽车声控锁等领域中。

- 社区矫正

有关资料显示，全国每年都有上万人甚至更多的人冒领社保达数亿元之多。为了防止养老金被冒领，进一步完善对养老保险金的管理和监督，社保局可通过预装声纹身份认证系统，再结合人工辅助手段，对领养老金者进行现场身份认证或当本人无法亲临现场时可通过电话进行远程身份确认，有效地阻止国家社保养老金的流失，提高社保服务机构工作的准确性和安全性。与其它生物认证技术相比，声纹认证技术具有更强的远程操控性，可快捷灵活的应用于远程身份认证中。

2.3 其它应用领域

除了上述相关应用领域，说话人检测和追踪技术也有着广泛的应用。在含有多说话人的语音段中，如何高效准确地把目标说话人检测标识出来有着十分重要的意义。例如，在现有音频/视频会议系统中，通常设有多麦克风阵列用以实时记录会议中每一个说话人的讲话。通过将说话人追踪技术嵌入该会议系统，可实时标识每段语音所对应的说话人，实时追踪“who spoke when”。该技术广泛应用于远程会议中，方便会议纪要总结，有利于提高公司的工作效率。

2.4 声纹识别的行业及国家标准

为了使生物特征识别技术得到更好地发展，国际标准化组织(International Organization for Standardization, ISO)对生物特征识别的相关术语及其产业技术制订了标准和规范，其中涵括了声纹识别技术。我国国家标准和相关行业权威部门也针对声纹识别技术制定了一系列的标准及规范，如[27]：

- SJ/T 11380-2008

由北京得意公司、清华大学智能技术与系统国家重点实验室（语音与语言技术中心）和中国电子技术便准化研究所共同起草的《自动声纹识别（说话人识别）技术规范》(SJ/T 11380-2008)于 2008 年 3 月 11 日正式颁布实施，该标准的内容主要包括声纹识别（说话人识别）的术语与定义、数据交换格式和应用编程接口，适用于各种计算机、网

络和智能设备的声纹识别系统。该标准是我国第一个关于声纹识别(说话人识别)的标准，其颁布很好地推动和规范了我国的声纹识别产业的发展。

- GA/T 893-2010

由清华大学、中国科学院自动化研究所、中国科学院计算技术研究所等机构单位共同起草的《安防生物特征识别应用术语》(GA/T 893-2010)标准于 2010 年 12 月 1 日起实施，该标准规范化了生物特征识别技术通用术语，其中包括了声纹识别专用术语的定义规范。该标准的颁布实施给生物特征识别技术的研究带来了方便，同时也避免了研究人员因滥用自定义术语而对技术研究造成不良影响。

- GA/T 1179-2014

2014 年 9 月 19 日，由全国安防标委会人体生物特征识别应用分技术委员会正式发出公告，《安防声纹确认应用算法技术要求和测试方法》(GA/T 1179-2014)标准已通过审核批准予以颁布，并于 2014 年 10 月 1 日开始实施。该标准是由清华大学语音和技术中心和北京得意公司为主要单位共同起草的。该标准首次提出声纹识别安全分级的概念。它的颁布在一定程度上促进了国内声纹技术在安防行业的发展应用。

此外，全国信息标准化委员会生物特征识别分技术委员会(SAC/TC28/SC37)也设有了生物特征识别标准委员会，其生物特征识别标准委员会也对生物特征识别在其应用领域提供了一些标准。这将对生物特征识别技术的发展起到推动性的作用。然而，目前这些标准对于生物特征识别行业的发展还是远远不够的，更多更精细的标准有待制定，以此满足生物特征识别技术和产业的发展。

3 声纹识别所面临技术问题

近年来，声纹识别技术发展迅速，并已在许多领域得以应用。然后，实际应用中的复杂环境对声纹识别系统的鲁棒性提出了巨大的挑战。为此，许多高校、科研机构和公司企业针对声纹识别面临的各类问题开展了一系列探究，并已提出实施了相应解决方案。本章将罗列声纹识别技术所面临问题和对应解决方案[28]。

3.1 环境噪音

实际应用中，人们所处的录音环境总是包含着各种类型的噪音，如白噪音、音乐播放、开关门的声音等等。这些噪音不仅在一定程度上淹没了语音信号中所蕴含的说话人信息，并且使得声纹系统无法获取准确的说话人声纹特征。此外，环境噪音通常是不可预知的，这使得声纹识别性能具有极大的不确定性。为了解决环境噪音对声纹识别系统的影响，关于噪音鲁棒性的研究陆续展开；其主要包括两个方向，一方面是提高声纹特征的噪音鲁棒性；另一方面是建立噪音鲁棒的声纹模型。针对声纹特征，研究着提出了频谱减法(Spectral Subtraction)用于解决固定环境噪音[29-30]；RASTA 滤波法用以消除信道缓慢变化的噪音[31]；也有许多算法用以提高特征鲁棒性，如主成分分析法[32]、线性判别法[33]和异方差线性判别分析[34]等。针对模型的噪音鲁棒性，其通常采用模型补偿算法[35-36]来减少测试和训练之间的噪音失配程度。

3.2 信道失配

信道失配是影响声纹识别性能的另一大因素。在实际应用中，语音信号可通过各式各样的录音设备录制得到，如不同麦克风、手机、固定电话、采访录音笔等等。而录音设备的不同会直接导致语音信号传输信道的变化，使得语音信号发生频谱畸变，进而严重影响语音声学特征和说话人模型对说话人特性的表征能力，造成测试语音声学特征与说话人模

型在声学空间分布上的失配。这种失配在很大程度上降低了说话人识别系统的性能。现有解决信道失配问题的方法也有很多，如传统 GMM-UBM 框架下的特征变换（feature transformation）[37,38,39]、模型补偿（model compensation）[40,41]、分数归一化（score normalization）[21,42]；JFA/i-vector 模型与区分性方法（如 WCCN[43]，LDA[23,44]，NAP[45]，PLDA[46,47]等）的结合；在基于神经网络的说话人识别系统中通过消除一些网络隐藏节点[48]等。

3.3 多说话人

多说话人是指在同一时刻有两个或两个以上的人同时说话，从而形成了多说话人的混合语音，其声纹识别的复杂性远大于单个说话人。在实际应用中，说话人在声纹语音录制时往往会掺杂其他说话人，如果系统不能实现多说话人的语音分离，将直接影响到声纹识别系统的性能。为此，研究者提出了说话人分割技术，对多说话人混合语音进行分割和聚类处理，捕捉获取语音信号各时间点所对应的说话人信息[49]。根据分割聚类过程的不同，可分为同步语音分割聚类和异步语音分割聚类。前者指声纹系统在分割语音片段的同时判别语音片段所对应的说话人类别；后者是将多说话人的混合语音分割成若干个独立的说话人语音片段，而后再将同一说话人的语音片段聚在一起进行每个说话人身份认证[50]。

当前常用的同步说话人语音分割聚类的算法有：基于 E-HMM 算法自上而下[51]、自下而上[52,53]的方法；为解决自上而下初始参数设定的问题，Imseng 提出了具有鲁棒性的自适应方法[54]；由于同步分割聚类系统主要采用基于 HMM 的方法，为确定 HMM 状态数，Fox 等人先后又提出了基于 HDP-HMM 的分割聚类算法和粘性的 HDP-HMM（sticky HDP-HMM）算法[55]。

3.4 说话人自身

说话人自身的影响是指自身的一些因素对声纹识别性能带来的影响。同一个说话人的声音具有易变性，会受其身体状况、时变、情感、语速以及语言等各种因素的影响，这些因素的影响也是当前语音信号处理的重要难点[56]。针对以上提到的这些影响因素，本节将逐一介绍与剖析。

- **身体状况**：语音发出者可能由于身体不适，如感冒、喉炎、鼻塞及其它原因，引起声音变化，这种变化会使声纹特征发生畸变，导致声纹识别的准确度降低。早在 1996 年，Tull 等人已经对感冒在说话人识别中所带来的影响展开了相关的研究[57,58]。该研究分别在感冒情况下和正常情况下的声道、基频和梅尔倒谱系数、共振峰等信号参数进行了分析。研究发现，感冒情况下的语音信号中含有由于嗓子嘶哑和咳嗽所产生的噪音，而这种噪音在正常语音中是不存在的。
- **时变**：人的声道会随着年龄的增长而变化，因此同一个人在不同年龄阶段所发出的声音也是有所不同的[59,60]。在实际应用中发现，声纹预录和声纹验证通常不在同一个时间段甚至相隔了很长时间，而这种时间间隔会对声纹识别系统性能造成明显的衰减[61,62]。为提高声纹识别系统的时变鲁棒性，研究者提出了一系列方法。如定期要求用户进行声纹模型更新，或者借助最新用户数据完成用户声纹模型自适应更新[63,64,65,66]。从特征域角度，有研究者提出了基于 F-ratio 准则的频带区分性特征算法和基于性能驱动的频带弯折算法[67,68]，其强调说话人个性信息的同时弱化时变信息，提取了时变鲁棒的声纹特征。
- **情感**：情感属于语音中的一种信息，同一个人在不同情感下所发出的语音是有所不同的。在实际应用中，用户情绪不可能是一层不变的，其通常会受各种因素的影响而产生不同情绪化的语音，其对说话人的音量、语速、语调均会产生一定的影响。有研究表明，在不同的情绪状态（喜、怒、哀、乐）下，每种状态的频谱分布有所不同[69,70,71]。为此，针

对情感对声纹识别的影响，研究者提出了附加情感的模型训练方法[72]，以此获取情感相关的声纹模型；此外，还有研究者提出了基于特征和模型联合优化的方法[73,74]，将情感特征投影到中性特征空间，进而弱化情感信息的影响。

- **语速**：语速是对一个人说话快慢的度量，其是一种高层的说话人信息。对同一个人而言，当其在重复同一段话时，几乎不可能实现语音的时间同步，而这就是由语速快慢而造成的[75]。有研究表明，语速对声纹识别系统有较大的影响，语速过快或者过慢均会使系统性能降低[76]。针对语速的研究尚少，目前是在文本相关的声纹识别系统上采用时间对准的方法降低语速对系统性能的影响[77]；而对于文本无关的声纹识别，由于无法预知时间对准信息，因此对语速的研究较为棘手。
- **跨语言**：跨语言是指说话人在进行声纹预留和声纹验证时使用不同的语言，如：说话人使用汉语进行模型训练，而用英语进行声纹验证。实验表明，与同语言识别结果相比，跨语言声纹识别准确率大幅降低[78]。在这个多民族、多国家、多文化的时代，单一的语言已经不能满足人们工作和学习交流的需要，因此，跨语言声纹识别是必须要解决的问题。针对跨语言声纹识别已经有了一些成果，如在声纹建模时采用多种语言的语音，训练一个多语言说话人模型，提高模型的语言鲁棒性[78]；提取更加鲁棒的声纹特征，削弱语言信息的影响[79]；提出语言因子补偿算法[80]，试图将语音中的语言因子信息消除，进而降低跨语言对系统的影响。

3.5 假冒闯入

早在上个世纪 90 年代，研究者已关注到生物特征识别中的假冒闯入问题，并针对指纹、人脸识别提出了一系列防假冒闯入的方法和对策。随着声纹识别技术的快速发展与广泛应用，针对声纹识别的防假冒闯入研究也逐渐兴起[81,82,83]。声纹识别的防假冒闯入场景主要分为声音模仿、语音合成、声音转换和录音重放四个方面。

- **声音模仿**：模仿是最早的研究方向[84]，研究者认为模仿更多体现地是对韵律和讲话风格的模仿，而未从根本上改变声道特性。因此，模仿更多是对人耳的欺骗，而对声纹识别系统影响不大[85]。
- **语音合成**：近年来语音合成技术发展迅速，其可借助少量语音实现特定说话人的模型自适应，而后将合成得到的特定说话人语音进行声纹系统假冒闯入[86,87,88,89]。通过探究正常语音与合成语音在声学特性之间的差异性（如：频谱/相位谱动态特性[88,90]、高阶梅尔倒谱系数的离散度[91]、F0 统计[92,93]等），现已有许多方法对策实现针对语音合成的闯入检测。
- **声音转换**：其通常分为离线训练和在线转换两个过程[94]，如何建立训练闯入语音和目标语音之间的转换函数决定了声音转换假冒闯入的效果。现有的声音转换检测方法，如余弦相位谱[90]、MGDF 相位谱分析[90]，基于句子层和短时的特征离散度统计[95,96]等。
- **录音重放**：与前三者相比，录音重放在实际应用中更易出现[97]。闯入者无需任何语音学知识，仅借助简单的录音放音设备即可实现录音重放闯入[98]。与此同时，实验表明录音重放闯入率也较高。为此，研究者开展了一系列的研究。如：基于语谱图，构建“语音指纹”模型[99]；针对录音重放带来的信道变化，提出了信道检测算法[100,101]等。

针对实际应用中的假冒闯入问题，除了上述提到的各类解决方法外，活体检测也是一种有效地防闯入机制。活体检测通俗地讲就是判断系统输入是预先处理得到的语音（如合成语音、转换语音、录音重放语音）还是真实的活体人声。如何将活体检测技术合理应用到声纹识别系统中，对防假冒闯入和提高系统鲁棒性具有十分重要的意义。近年来许多研究机构和公司开展了一系列研究，并提出了相应的技术方案[102,103]。

3.6 短语音

对于实际应用中的声纹识别系统，其用户体验性的好坏已成为一项重要的评价指标，显然较短的测试语音时长会带来更好的用户体验；此外，在很多声纹识别的应用领域，实际使用时无法获取足够长度的测试语音（如刑侦安防等领域）。因此，研究较短的测试语音时长下的声纹识别具有很强的现实意义。

早在 1983 年，研究者就注意到语音时长对说话人系统的性能有着直接的影响[104]。然而，对于目前主流的几种声纹识别系统（GMM-UBM，i-vector，JFA），在较短测试语音条件下的系统性能变化均十分剧烈，且都不能取得令人满意的效果。其原因在于短语音测试条件下，测试语音中所包含的说话人信息不均衡，进而导致训练与识别的匹配性严重下降；此外，短语音条件下测试语音中的信息量太少，不足以提供充足的区分性信息，使得识别混淆度变大[2]。

近些年来，针对短语音声纹识别，研究者们也提出了一系列方法与对策。例如，从语音中筛选更具有区分性的数据[105]；融合不同声学特征获得更鲁棒的特征参数[106]；结合语音识别的先验知识构建更精细的声纹模型[107]；更合理准确的双边似然分计算方式[108]等。

4 总结与展望

本文对生物特征技术中的声纹识别作了详细介绍。综述了声纹识别的基本概念、应用领域、行业标准、声纹识别所面临问题及其解决方案，同时对现有和未来的潜在应用进行了介绍。

声纹识别技术发展到今天，对所面临问题的解决方法并未完全成熟，与实际情景中的运用还存在一定的距离，但声纹识别技术在未来方方面面的潜在应用是有目共睹的。该技术有望应用于金融安全、公共安全、社保生存认证、社区矫正系统、移动互联网安全、车联网安全等各个领域。

目前，在实际应用中，可通过声纹识别融合其它的生物认证方式如人脸、指纹、虹膜等技术的优势提高系统识别认证的安全性；在远程身份认证中，可以声纹认证为主，人工为辅对操作者进行双重认证；在手机支付及声纹锁认证中，可借助动态随机码和语音识别以防止录音假冒的闯入。总之，在不同的应用场合下，可根据实际的需求，灵活的使用声纹识别认证技术。

聆听美好声音，科技不仅在你的身边，还在你的声音中。用你的声音探索身边的科技，用你的声音掌控你身边的智能设备，声纹识别技术会让科技更人性化，让人们的生活更愉快、更轻松。

参考文献

[1] Wikipedia, https://en.wikipedia.org/wiki/Biometrics

[2] 张陈昊. 短语音说话人识别研究. 清华大学计算机科学与技术, 博士论文. 2014.

[3] 中华人民共和国电子行业标准, “自动声纹识别(说话人识别)技术规范”. SJ/T 11380-2008.

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

[5] Campbell Jr J P. Speaker recognition: A tutorial. Proceedings of the IEEE, 85 (9): 1437-1462, 1997.

[6] Wikipedia, https://en.wikipedia.org/wiki/Speaker_recognition.

[7] Martin A, Doddington G, Kamm T, et al. The DET curve in assessment of detection task performance. European Conference on Speech Communication and Technology (Eurospeech 1997), Rhodes, Greece, September 1997. (4):

1895－1898

[8] 吴玺宏. 声纹识别听声辨认. 计算机世界. 2001. (8).

[9] Furui S. 50 years of progress in speech and speaker recognition [J]. Speech Communication 2005, Patras, 2005: 1-9.

[10] Atal B S, Hanauer S L. Speech analysis and synthesis by linear prediction of the speech wave [J]. The Journal of the Acoustical Society of America, 1971, 50(2B): 637-655.

[11] Doddington G R, Flanagan J L, Lummis R C. Automatic speaker verification by non-linear time alignment of acoustic parameters: U.S. Patent 3,700,815[P]. 1972-10-24.

[12] Atal B S. Automatic speaker recognition based on pitch contours [J]. The Journal of the Acoustical Society of America, 1972, 52(6B): 1687-1697.

[13] Makhoul J, Cosell L. LPCW: An LPC vocoder with linear predictive spectral warping [C]. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 1976, 1: 466-469.

[14] Hermansky H. Perceptual linear predictive (PLP) analysis of speech [J]. the Journal of the Acoustical Society of America, 1990, 87(4): 1738-1752.

[15] Vergin R, O'shaughnessy D, Farhat A. Generalized mel frequency cepstral coefficients for large-vocabulary speaker-independent continuous-speech recognition [J]. IEEE Transactions on Speech and Audio Processing, 1999, 7(5): 525-532.

[16] 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.

[17] Burton D K, Shore J E, Buck J T. A generalization of isolated word recognition using vector quantization [C]. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 1983, 8: 1021-1024.

[18] Rabiner L R, Juang B H. An introduction to hidden Markov models [J]. ASSP Magazine, IEEE, 1986, 3(1): 4-16.

[19] Jain A K, Mao J, Mohiuddin K M. Artificial neural networks: A tutorial [J]. Computer, 1996 (3): 31-44.

[20] Reynolds D. Gaussian mixture models [M]. Encyclopedia of Biometrics. Springer US, 2009: 659-663.

[21] 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.

[22] Dehak N, Dumouchel P, Kenny P. Modeling prosodic features with joint factor analysis for speaker verification [J]. IEEE Transactions on Audio, Speech, and Language Processing, 2007, 15(7): 2095-2103.

[23] 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.

[24] Variani E, Lei X, McDermott E, et al. Deep neural networks for small footprint text-dependent speaker verification [C]. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2014: 4052-4056.

[25] Kenny P, Gupta V, Stafylakis T, et al. Deep neural networks for extracting Baum-Welch statistics for speaker recognition [C]. IEEE Odyssey - the Speaker and Language Recognition Workshop. 2014.

[26] Furui S. Recent advances in speaker recognition [C]. Audio-and Video-based Biometric Person Authentication. Springer Berlin Heidelberg, 1997: 235-252.

[27] Zheng T F. Prove yourself by yourself with the use of speaker recognition technology. NCMMSC'15, 2015.

[28] Zheng T F, Jin Q, Li L T, et al. An overview of robustness related issues in speaker recognition [C]. Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC 2014), 2014: 1-10.

[29] Boll S F. Suppression of acoustic noise in speech using spectral subtraction [J]. IEEE Transactions on

Acoustics, Speech and Signal Processing, 1979, 27(2): 113-120.

[30] Berouti M, Schwartz R, Makhoul J. Enhancement of speech corrupted by acoustic noise [C]. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 1979, 4: 208-211.

[31] Hermansky H, Morgan N. RASTA processing of speech [J]. IEEE Transactions on Speech and Audio Processing, 1994, 2(4): 578-589.

[32] Kocsor A, Tóth L, Kuba A, et al. A comparative study of several feature transformation and learning methods for phoneme classification [J]. International Journal of Speech Technology, 2000, 3(3-4): 263-276.

[33] Lomax R G, Hahs-Vaughn D L. Statistical concepts: a second course [M]. Routledge, 2013.

[34] Saon G, Padmanabhan M, Gopinath R, et al. Maximum likelihood discriminant feature spaces [C]. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2000, 2: II1129-II1132 vol. 2.

[35] Gales M J F, Young S J. Robust continuous speech recognition using parallel model combination [J]. IEEE Transactions on Speech and Audio Processing, 1996, 4(5): 352-359.

[36] Renevey P, Drygajlo A. Statistical estimation of unreliable features for robust speech recognition [C]. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2000, 3: 1731-1734.

[37] Reynolds D. Channel robust speaker verification via feature mapping [C]. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2003, 2: II-53-6 vol.2.

[38] Zhu D, Ma B, Li H, et al. A generalized feature transformation approach for channel robust speaker verification [C]. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2007, 4: IV-61-IV-64.

[39] Vair C, Colibro D, Castaldo F, et al. Channel factors compensation in model and feature domain for speaker recognition [C]. IEEE Odyssey - the Speaker and Language Recognition Workshop, 2006: 1-6.

[40] Heck L P, Weintraub M. Handset-dependent background models for robust text-independent speaker recognition [C]. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 1997, 2: 1071-1074.

[41] Teunen R, Shahshahani B, Heck L P. A model-based transformational approach to robust speaker recognition [C]. INTERSPEECH. 2000: 495-498.

[42] Auckenthaler R, Carey M, Lloyd-Thomas H. Score normalization for text-independent speaker verification systems [J]. Digital Signal Processing, 2000, 10(1): 42-54.

[43] Hatch A O, Kajarekar S S, Stolcke A. Within-class covariance normalization for SVM-based speaker recognition [C]. INTERSPEECH. 2006.

[44] McLaren M, Van Leeuwen D. Source-normalised-and-weighted LDA for robust speaker recognition using i-vectors [C]. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2011: 5456-5459.

[45] Solomonoff A, Quillen C, Campbell W M. Channel compensation for SVM speaker recognition [C]. IEEE Odyssey - the Speaker and Language Recognition Workshop. 2004, 4: 219-226.

[46] Ioffe S. Probabilistic linear discriminant analysis [M]. Computer Vision - ECCV 2006. Springer Berlin Heidelberg, 2006: 531-542.

[47] Prince S J D, Elder J H. Probabilistic linear discriminant analysis for inferences about identity [C]. 11th International Conference on Computer Vision (ICCV), IEEE, 2007: 1-8.

[48] Kishore S P, Yegnanarayana B. Speaker verification: Minimizing the channel effects using autoassociative neural network models [C]. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2000, 2: II1101-II1104 vol. 2.

[49] Tranter S E, Reynolds D. An overview of automatic speaker diarization systems [J]. IEEE Transactions on

Audio, Speech, and Language Processing, 2006, 14(5): 1557-1565.

[50] Kotti M, Moschou V, Kotropoulos C. Speaker segmentation and clustering [J]. Signal processing, 2008, 88(5): 1091-1124.

[51] Meignier S, Bonastre J F, Fredouille C, et al. Evolutive HMM for multi-speaker tracking system [C]. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2000, 2: II1201-II1204 vol. 2.

[52] Ajmera J, Wooters C. A robust speaker clustering algorithm [C]. IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), 2003: 411-416.

[53] Wooters C, Huijbregts M. The ICSI RT07s speaker diarization system [M]. Multimodal Technologies for Perception of Humans. Springer Berlin Heidelberg, 2008: 509-519.

[54] Imseng D, Friedland G. Tuning-robust initialization methods for speaker diarization [J]. IEEE Transactions on Audio, Speech, and Language Processing, 2010, 18(8): 2028-2037.

[55] Fox E B, Sudderth E B, Jordan M I, et al. A sticky HDP-HMM with application to speaker diarization [J]. The Annals of Applied Statistics, 2011: 1020-1056.

[56] Huang C, Chen T, Li S Z, et al. Analysis of speaker variability [C]. INTERSPEECH. 2001: 1377-1380.

[57] Tull R G, Rutledge J C. Analysis of “cold‐affected”speech for inclusion in speaker recognition systems [J]. The Journal of the Acoustical Society of America, 1996, 99(4): 2549-2574.

[58] Tull R G, Rutledge J C. ‘Cold Speech’ for Automatic Speaker Recognition. Acoustical Society of America 131st Meeting Lay Language Papers, May, 1996.

[59] Kersta L G. Voiceprint Recognition. Nature, No. 4861, pp. 1253-1257, December 1962.

[60] Bonastre J F, Bimbot F, Boë L J, et al. Person authentication by voice: a need for caution [C]. INTERSPEECH. 2003.

[61] Kato T, Shimizu T. Improved speaker, verification over the cellular phone network using phoneme-balanced and digit-sequence-preserving connected digit patterns [C]. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2003, 2: II-57-60 vol. 2.

[62] Hébert M. Text-dependent speaker recognition [M]. Springer handbook of speech processing. Springer Berlin Heidelberg, 2008: 743-762.

[63] Bimbot F, Bonastre J F, Fredouille C, et al. A tutorial on text-independent speaker verification [J]. EURASIP journal on applied signal processing, 2004, 2004: 430-451.

[64] Beigi H. Effects of time lapse on speaker recognition results [C]. 16th IEEE International Conference on Digital Signal Processing, 2009: 1-6.

[65] Beigi H. Fundamentals of speaker recognition [M]. Springer Science & Business Media, 2011.

[66] Lamel L F, Gauvain J L. Speaker verification over the telephone [J]. Speech Communication, 2000, 31(2): 141-154.

[67] Wang L-L, Wu X-J, Zheng T F, et al. An investigation into better frequency warping for time-varying speaker recognition [C]. Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC 2012), 2012: 1-4.

[68] Wang L-L, Zheng T F. Creation of time-varying voiceprint database [J]. Proc. of O-COCOSDA 2010, 2010.

[69] Bie F-H, Wang D, Zheng T F, et al. Emotional speaker verification with linear adaptation [C]. IEEE China Summit & International Conference on Signal and Information Processing (ChinaSIP), 2013: 91-94.

[70] Zetterholm E. Prosody and voice quality in the expression of emotions [C]. ICSLP. 1998.

[71] Pereira C, Watson C I. Some acoustic characteristics of emotion [C]. ICSLP. 1998.

[72] Wu T, Yang Y, Wu Z. Improving speaker recognition by training on emotion-added models [M]. Affective Computing and Intelligent Interaction. Springer Berlin Heidelberg, 2005: 382-389.

[73] Shahin I. Speaker identification in emotional environments [J]. Iranian Journal of Electrical and Computer Engineering, 2009, 8(1): 41-46.

[74] Bie F-H, Wang D, Zheng T F, et al. Emotional adaptive training for speaker verification [C]. Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC 2013), IEEE, 2013: 1-4.

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

[76] Matsui T, Furui S. Comparison of text-independent speaker recognition methods using VQ-distortion and discrete/continuous HMM's [J]. IEEE Transactions on Speech and Audio Processing, 1994, 2(3): 456-459.

[77] Yasuda H, Kudo M. Speech rate change detection in martingale framework [C]. 12th IEEE International Conference on Intelligent Systems Design and Applications (ISDA), 2012: 859-864.

[78] Ma B, Meng H. English-Chinese bilingual text-independent speaker verification [C]. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2004, 5: V-293-6 vol. 5.

[79] Nagaraja B G, Jayanna H S. Combination of Features for Multilingual Speaker Identification with the Constraint of Limited Data [J]. International Journal of Computer Applications, 2013, 70(6): 1-6.

[80] Lu L, Dong Y, Zhao X, et al. The effect of language factors for robust speaker recognition [C]. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2009: 4217-4220.

[81] Lindberg J, Blomberg M. Vulnerability in speaker verification-a study of technical impostor techniques [C]. Eurospeech. 1999, 99: 1211-1214.

[82] Evans N, Kinnunen T, Yamagishi J. Spoofing and countermeasures for automatic speaker verification [C]. INTERSPEECH. 2013: 925-929.

[83] Wu Z, Evans N, Kinnunen T, et al. Spoofing and countermeasures for speaker verification: a survey [J]. Speech Communication, 2015, 66: 130-153.

[84] Lau Y W, Wagner M, Tran D. Vulnerability of speaker verification to voice mimicking [C]. 2004 IEEE International Symposium on Intelligent Multimedia, Video and Speech Processing, 2004: 145-148.

[85] Perrot P, Aversano G, Blouet R, et al. Voice Forgery Using ALISP: Indexation in a Client Memory [C]. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), (1). 2005: 17-20.

[86] Masuko T, Tokuda K, Kobayashi T, et al. Speech synthesis using HMMs with dynamic features [C]. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 1996, 1: 389-392.

[87] Masuko T, Tokuda K, Kobayashi T, et al. Voice characteristics conversion for HMM-based speech synthesis system [C]. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 1997, 3: 1611-1614.

[88] De Leon P L, Pucher M, Yamagishi J, et al. Evaluation of speaker verification security and detection of HMM-based synthetic speech [J]. IEEE Transactions on Audio, Speech, and Language Processing, 2012, 20(8): 2280-2290.

[89] Galou G, Chollet G. Synthetic voice forgery in the forensic context: a short tutorial [C]. Forensic speech and audio analysis working group (ENFSI-FSAAWG). 2011.

[90] Wu Z, Siong C E, Li H. Detecting Converted Speech and Natural Speech for anti-Spoofing Attack in Speaker Recognition [C]. INTERSPEECH. 2012.

[91] Chen L-W, Guo W, Dai L-R. Speaker verification against synthetic speech [C]. 7th International Symposium on Chinese Spoken Language Processing (ISCSLP), 2010: 309-312.

[92] Ogihara A, Hitoshi U, Shiozaki A. Discrimination method of synthetic speech using pitch frequency against synthetic speech falsification [J]. IEICE transactions on fundamentals of electronics, communications and computer sciences, 2005, 88(1): 280-286.

[93] De Leon P L, Stewart B, Yamagishi J. Synthetic Speech Discrimination using Pitch Pattern Statistics Derived from Image Analysis [C]. INTERSPEECH. 2012.

[94] Stylianou Y. Voice transformation: a survey [C]. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP). 2009: 3585-3588.

[95] Alegre F, Vipperla R, Evans N. Spoofing countermeasures for the protection of automatic speaker recognition systems against attacks with artificial signals [C]. INTERSPEECH 2012.

[96] Alegre F, Amehraye A, Evans N. Spoofing countermeasures to protect automatic speaker verification from voice conversion [C]. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2013: 3068-3072.

[97] Lindberg J, Blomberg M. Vulnerability in speaker verification-a study of technical impostor techniques [C]. Eurospeech. 1999, 99: 1211-1214.

[98] Alegre F, Amehraye A, Evans N. Spoofing countermeasures to protect automatic speaker verification from voice conversion [C]. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2013: 3068-3072.

[99] Wu Z, Gao S, Cling E S, et al. A study on replay attack and anti-spoofing for text-dependent speaker verification [C]. Asia-Pacific Signal and Information Processing Association, 2014 Annual Summit and Conference (APSIPA ASC), IEEE, 2014: 1-5.

[100] Villalba J, Lleida E. Detecting replay attacks from far-field recordings on speaker verification systems [M]. Biometrics and ID Management. Springer Berlin Heidelberg, 2011: 274-285.

[101] Wang Z-F, Wei G, He Q-H. Channel pattern noise based playback attack detection algorithm for speaker recognition [C]. IEEE International Conference on Machine Learning and Cybernetics (ICMLC), 2011, 4: 1708-1713.

[102] Shiota S, Villavicencio F, Yamagishi J, et al. Voice liveness detection algorithms based on pop noise caused by human breath for automatic speaker verification [C]. Sixteenth Annual Conference of the International Speech Communication Association. 2015.

[103] 郑方, 邬晓钧, 张陈昊, 王军, 瞿世才, 熊音. 基于动态密码语音的身份确认系统及方法. 专利号：ZL201310123555.0.

[104] Li K P, Wrench Jr E H. An approach to text-independent speaker recognition with short utterances [C]. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 1983, 8: 555-558.

[105] Kwon S, Narayanan S. Robust speaker identification based on selective use of feature vectors [J]. Pattern Recognition Letters, 2007, 28(1): 85-89.

[106] Zhang C-H, Zheng T F. A fishervoice based feature fusion method for short utterance speaker recognition [C]. IEEE China Summit & International Conference on Signal and Information Processing (ChinaSIP), 2013: 165-169.

[107] Zhang C-H, Wu X-J, Zheng T F, et al. A K-phoneme-class based multi-model method for short utterance speaker recognition [C]. Asia-Pacific Signal & Information Processing Association Annual Summit and Conference (APSIPA ASC 2012), 2012: 1-4.

[108] Malegaonkar A, Ariyaeeinia A, Sivakumaran P, et al. On the enhancement of speaker identification accuracy using weighted bilateral scoring [C]. 42nd IEEE Annual International Carnahan Conference on Security Technology (ICCST), 2008: 254-258.

基金资助

本项研究工作得到了国家自然科学基金(编号. 61271389/61371136)和国家重点基础研究发展计划(973 计划，编号. 2013CB329302)的支持。

作者信息

郑方　　清华大学信息技术研究院 语音和语言技术中心主任　教授　博士生导师

李蓝天　清华大学信息技术研究院　语音和语言技术中心　博士研究生

张慧　　贵州大学科技学院　工学部　本科生

艾斯卡尔·肉孜　清华大学信息技术研究院　语音和语言技术中心　博士研究生

Overview of Voiceprint Recognition Technology and Applications

Zheng Thomas Fang[1*], Li Lantian[1], Zhang Hui[2], Asker Rozi[1]

[1]Center for Speech and Language Technologies
Division of Technical Innovation and Development, Tsinghua National Laboratory for Information Science and Technology; Research Institute of Information Technology;
Department of Computer Science and Technology
Tsinghua University, Beijing, 100084, China
{lilt, askar}@cslt.riit.tsinghua.edu.cn
*Corresponding author: Zheng Thomas Fang, E-mail：fzheng@tsinghua.edu.cn

[2]The College of Science and Technology of Guizhou University
hebe.hui.zhang@gmail.com

Abstract: With the rapid development of information technology, how to identify a person to protect his/her personal privacy as well as information security has become a hot issue. Comparing with the traditional identity authentication, the biometrics features are not easy to get lost, to be stolen or forgotten when being used. The use of them is not only fast and convenient, but also accurate and reliable. Being one of the most popular biometric authentication technologies, the voiceprint recognition technology has its unique advantages in the field of remote authentication and other areas, and has attracted more and more attention. In this paper, the voiceprint recognition technology and its applications will be mainly introduced, including the fundamental concept, development history, technology applications and industrial standardizations. Various kinds of problems and corresponding solutions are overviewed, and the prospects are pointed out finally.

Keywords: Biometric Recognition, Identity Authentication, Voiceprint Recognition, Development History, Technology Applications