ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 30,No. 2,Mar. 2015,pp. 266－274
DOI:10.16337/j.1004-9037.2015.02.004

http://sjcj.nuaa.edu.cn
E-mail:sjcj@nuaa.edu.cn
Tel/Fax:＋86-025-84892742

说话人认证录音回放检测方法综述

贺前华[1]　潘伟锵[2]　胡永健[1]　朱铮宇[1]　李艳雄[1]　奉小慧[1]

(1.华南理工大学电子与信息学院,广州,510641;2.华南理工大学信息网络工程研究中心,广州,510641)

摘　要: 基于生物特征的身份认证已得到学术界及企业的高度重视,指纹、人脸识别应用已非常普遍,但对于非现场身份认证,语音相对其他生物特征,具有用户接受程度高、拾音设备简单、随时随地可用、数据量小、计算复杂度低等优势,因此基于声纹的身份认证系统应用越来越广泛。另一方面,由于录音回放攻击简单易行,不需要任何专业知识,且随着廉价、高质量的录音/播放装置的日益增多,回放录音与原始音的相似度越来越高,已成为声纹认证系统最主要的攻击手段之一,因此如何识别录音回放等攻击成为说话人认证系统必须面对的问题。本文对录音回放检测方法进入了全面的介绍,通过对各种方法的分析,表明其研究尚处于起步阶段,但需求日益旺盛。

关键词: 说话人认证;认证语音真实性;录音回放攻击

中图分类号: TN912.3　　**文献标志码:** A

Review on Playback Detection Methods in Speaker Authentication System

He Qianhua[1], Pan Weiqiang[2], Hu Yongjian[1], Zhu Zhengyu[1], Li Yanxiong[1], Feng Xiaohui[1]

(1. School of Electronic and Information Engineering, South China University of Technology, Guangzhou, 510641, China;
2. Information Network Engineering and Research Center, South China University of Technology, Guangzhou, 510641, China)

Abstract: Biometric authentication system has been widely used today, including fingerprint and face recognition systems. As for non-spot user authentication, compared with other biometric traits, speech has some advantages, such as high acceptability, low demand on equipments, flexible access anywhere and anytime, low computation complexity and suitability for remote authentication, which promotes the applications of speech-based authentication system. However, playback becomes a general risk because of its easy carring out without any training, and the availability of cheap-high quality audio recorder and speaker. This paper gives a through review to the methods for playback detection. It shows that the research is at the start point, but the demands are increasing.

Key words: speaker authentication; authenticity of forensic speech; record & playback attack

引　　言

基于知识或信物(What you know/have)的传统认证方式[1]已与人们的生活息息相关,比如电子邮箱、QQ账户、银行账户管理、各种“宝宝”管理等,但安全问题日益严重。为了提高身份认证安全,应用

基金项目: 国家自然科学基金(60972132)资助项目。
收稿日期: 2014-12-18;**修订日期:** 2015-01-28

系统往往采用多认证方式组合,典型组合方式是“know”与“have”的组合,如网银支付过程,需要网银账号、银行账号、U 盾及 U 盾支付密码。但无论是密码还是智能卡,存在易忘记、丢失、被人盗取等不足。而基于生物特征的身份认证[2-6]从“你是谁(Who you are)”的原则出发,利用人体的各种生理(指纹、虹膜、声纹、掌纹、人脸等)和行为特征(步态、签字、击键特征等)对用户进行身份认证,理论分析和实际应用都表明该认证方式比传统密码学认证具有更高的安全性和方便性。因此已广泛应用于出入境管理、电子商务、门禁系统、司法取证、网络游戏和养老金领取身份认证等。

2014 年一系列安全事件,如携程漏洞门[7]、虚拟信用卡和线下二维码支付被叫停、“HeartBleed”,把互联网安全问题推到了风口浪尖。公众缺乏安全感、网络攻击和诈骗手段的花样翻新、登录密码短信验证密码的屡次失手,所有现象都指向同一个需求:安全认证手段的更新换代。因此动态口令令牌、指纹认证、声纹识别[8-10]等第三方辅助认证形式,也再一次持续升温。

声纹是非常重要的生物认证依据之一,具有以下独特优势[3-4]:(1)语音获取方便、自然,使用者的接受程度高;(2)获取语音的设备成本低廉,使用简单,一个麦克风即可,在使用通讯设备时更无需额外的录音设备;(3)远程人机交互方便,且说话人辨认和确认的算法复杂度低;(4)配合一些其他措施,如通过语音识别进行内容鉴别等,准确率可以得到很大提高,并可提升安全性。上述特征使得声纹特别适合远程身份认证,若利用电话系统进行身份认证,语音是唯一可用的生物特征。另一方面,声纹识别(或者说话人识别)技术与语音识别一样,经历了 60 多年的发展,取得了很大的进步,逐渐进入商用阶段。美国国家标准与技术研究院(NIST)从 1996 年开始为说话人识别技术提供测试平台,已成为世界最有影响力的说话人识别技术评价手段,2012 年 NIST 的说话人识别评估报告[11]表明,噪声环境下的电话语音,等错误率达到 1.5%以下。因此国内基于声纹的身份认证产业发展迅速,成长了一批专业型企业,如科大讯飞、快商通声纹等。阿里巴巴也秘密部署声纹认证产品,并完成了第 2 轮内测[12],以应对其快速发展的电商业务安全问题。

生物认证系统也会受到各种攻击,图 1 是 ITU X.1086 标准建议[13]给出的生物认证系统的功能模块模型及可能受到的攻击。其中威胁 $T1$ 和 $T2$ 是数据输入端的攻击,$T1$,$T2$ 攻击可能是各种伪造的生物特征样本,也可能是合法用户已有生物特征样本的回(重)放,$T1$ 称为输入设备前端攻击,$T2$ 为输入设备后端攻击(很多时候称之为信道回放攻击)。在声纹身份认证系统中,$T1$,$T2$ 表现为录音回放(重放)、模仿、合成等攻击形式。

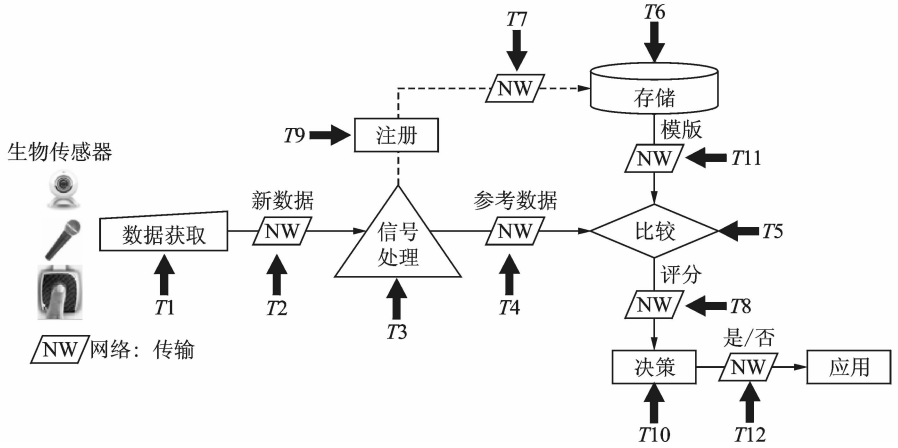

图 1 远程生物认证系统的可能受到的攻击

Fig. 1 Vulnerabilities on the telebiometric functional model

1 音源端攻击方式及危险性分析

针对声纹身份认证系统的攻击，即有生物认证系统攻击的共性，又有其特殊性。声纹认证系统的音源端攻击主要有录音回放攻击[14-19]、说话人仿冒攻击[20-21]、伪造认证语音攻击[22-24]3 种。

录音回放攻击是指攻击者采用高保真的录音设备录制合法用户进入认证系统时的语音，或通过其他各种手段获得用户的语音样本，然后在声纹身份认证系统的拾音器端通过高保真功放回放，从而达到对声纹身份认证系统实施攻击的目的[14-17]。说话人认证系统分为文本相关和文本无关两种，其对应的偷录语音方式也不同。对文本相关的声纹身份认证系统，回放语音是偷录用户进入系统时的语音，而对于文本无关的声纹身份认证系统只需获得用户语音即可实施录音回放攻击。随着科技的发展，高质量的廉价录音/播放设备日益增多，而且体积小，实施偷录容易，因此录音回放攻击已成为声纹身份认证系统中最容易实施的攻击。文献[25]对录音回放攻击的严重性进行了初步的实验探索，从说话人辨识及说话人识别两方面表明录音回放攻击成功率比较高。在普通实验室中，10 个人对 50 个真实语音和 50 个回放语音的辨识准确性只有 58.7%；采用 GMM(Gaussian mixture model)和 HMM(Hidden Markov model)的说话人识别系统对录音回放的认同率比真实语音只约低 6%，如表 1 所示，这表明录音回放攻击的危害很大。

表 1 真实语音和回放语音的说话人识别性能[25]

Table 1 Speaker recognition performance under genuine speech and playback speech[25] %

说话人识别系统	真实语音	回放语音
基于 HMM	87.28	81.10
基于 GMM	97.73	93.70

说话人仿冒是指一些善于模仿他人语音的入侵者通过模仿某合法用户的说话方式以及发音特点，达到非常相似的程度，借此实现对声纹身份认证系统的攻击目的[20]。

伪造语音攻击可分为合成[23]、转换[22,26]和拼接语音[27]攻击 3 种方式。合成语音攻击是指采用语音合成技术生成被攻击对象的语音，对说话人认证系统而言，只有合成语音与被冒充对象语音高度相似时才会有效；而转换语音是指将说话人 A 的语音，通过技术处理转换成具有说话人 B 语音特征的语音。拼接语音是指利用被冒充人的真实语音片段，拼接成表达某特殊语句的语音(例如对文本相关的说话人识别系统进行攻击)，拼接语音的特性类似于回放语音，但由于是不同音段连接而成，音段之间的不自然性更容易检测，所以将其归到伪造语音类。

上述 3 种音源攻击的实施难度有很大差异，对系统的危害也不同。说话人仿冒攻击需要攻击者具有很好的模仿能力，这种人很少；制造高质量的伪造语音往往需要较高的专业技能，能成功实施这一类攻击的人也不多；另一方面，无论是模仿音还是伪造音，终究不是真实音，现有的说话人识别技术足以辨认双胞胎的声纹特征[5]，即现有技术能够应对这两类攻击。文献[22]的研究验证了这一认识，其研究结果表明现有的语音转换技术对基于 GMM 的说话人识别系统攻击是存在的，但对基于音素的说话人识别系统，这种攻击是无效的。录音回放攻击只需要好的设备，不需要任何专业技能，非常容易实施。而且回放音可以在合法人与实施人的串通情况下获得，可以设想这样一种场景：社保金的发放采用声纹进行生存检测，而录音可以在合法人与子女的配合下获得，因此录音质量可以做到很好，这种情形下的回放音更能欺骗系统的认证能力。

综上所述，录音回放攻击易实施，发生概率高，成功率高，是最难检测的攻击方式之一[28]，随着声纹身份认证系统的不断推广应用，其危害性日益凸显，因此录音回放检测研究变得越来越迫切。

2 录音回放检测方法

录音回放检测的目的是检测认证语音的真实性,认证语音的真实性可分为内容真实性,来源真实性,说话人真实性和说话时间真实性等多个方面。基于真实性检测的侧重点不同,回放检测方法主要分为4类:基于挑战-响应的回放检测方法、利用多模态生物特征相关性的检测方法、利用语音随机性的检测方法以及基于语音生成信道特征分析的检测方法

2.1 基于挑战-响应的检测方法

挑战-响应(Challenge-response, CR)是身份认证的常用方法,比如谍报人员接头时使用暗语,目前账户认证的密码、动态验证码,其核心是基于知识(What you know)。在声纹身份认证系统中,挑战-响应是指认证系统以对话的方式向申请认证者提出问题,要求认证者回应[29],侧重认证语音内容真实性,由于机器以随机方式提出问题,录音难以预测这种高随机问题,因此是一种有效的防录音回放攻击的方法,也是ITU X.1086[13]推荐的方法。挑战-响应可分为文本无关与文本相关两种,在文本相关情形中,用户回答内容需根据系统提示。因此基于提示的说话人识别[6,30]尽管最初不是从安全角度考虑的,但很显然可以起到防止录音回放攻击的作用。文献[31]结合文本相关和文本无关的说话人识别,利用文本相关解决"你是谁"和"你知道什么",而利用文本无关挑战响应解决录音回放。Baloul等[32]也采用相同的方法来应对回放攻击。

在声纹认证系统中,使用挑战-响应存在不少问题:(1)合格的问题库建设难,为了保障认证系统提出的问题具有很高的随机性,需要事先准备丰富问题库,问题的选择涉及到语言学、心理学、人们日常生活习惯等多个方面;(2)用户需要对系统提出的一连串随机问题做答,而且很大概率会遇到无法回答的问题,不符合人们的会话习惯,因此用户接受度很低;(3)在人机会话不合拍的情况下,勉强的应对将影响发音质量,提升系统拒识概率,从而进一步提高人们的抗拒心理;(4)实际应用中,该方法只能应用在文本相关的声纹身份认证系统中,在做说话人识别的同时还要进行语音的文本识别,计算量大;(5)该方法会牺牲掉声纹身份认证系统对于特定用户的特定文本的安全保护性,会产生一些其他的安全问题[16]。

基于以上不足,挑战-响应在声纹认证中应用并不多,中外文文献库中能检索到的文献比较少。而在中国知网学术总库中用关键词"说话人、挑战-响应"检索,没有匹配结果。

2.2 基于多模态的检测方法

由于相对单一生物特征而言,多模态系统具有诸多优势,比如人群覆盖率、抗噪声能力、应对生物特征的不可靠性等,研究表明没有任何单生物特征可以为身份认证系统提供完善的抗攻击方案[33],多模态生物认证也是提高生物认证准确性和安全性的有效手段[34]。因此结合人脸识别,唇动语音一致分析或指纹等其他生物特征的说话人识别得到了广泛的重视。

基于多模态进行录音回放检测是依据多个生物特征做假比单一生物特征做假难度高很多的原理。若两个生物特征有具有强相关性,作假则更难,因此骗过系统的可能性就低很多。比如基于会话脸(talking-face)的身份认证综合利用了人脸、唇动和语音信息[35]。人们对偷录行为是防范的,因此要同时偷录到高质量人脸视频和语音比较难。

由于语音、人像获取自然度高,因此认证实体的接受度高,因此人脸与语音的结合得到了更多的关注,大多数研究是以提升身份认证准确性为目的,同时也为防录音回放提供思路。音视频信息的利用方式可分为信息融合[36-40]和相关性分析[41-47]两种。2004年Cetingul, H.E等[36]专门研究了适合于说话人识别的唇动特征,而文献[37]采用语音评分与人脸唇动信息评分相结合的方式进行说话人识别;文献[38]在文本提示的认证模式下,将唇动信息和语音信息相结合,利用Kernel Fisher判决分析作为分类

器,得到了比线性判决要好的说话人识别结果。2007 年,瑞典的学者 Maycel-Isaac Faraj 和 Josef Bigun[39]将声纹和人脸识别结合起来进行生物认证,并且通过唇动和语音结合的挑战-响应模式来进行活体检测。同年,Niall A. Fox 等人[40]通过融合人脸、唇动和语音 3 个模态系统的得分进行说话人身份认证,并用实验证明融合系统的鲁棒性明显优于单一模态系统。

在利用音视频相关性方面,Slaney 等[41]提出了一种运用典型相关分析法(Canonical correlation analysis, CCA)通过计算脸部正面运动与语音的关联性来检测两者是否一致的方法,该方法后来被应用于多模态系统的活体检测中。2006 年 Girija Chetty 等[42]则从语音与说话人脸的静态和动态相关性分析入手,提出了双模态特征融合、交叉模态融合和 3D 多模态融合的三层活体检测框架,可有效地防止动静态视频和语音回放攻击,并于 3 年后提出了一个音视频交互融合方案[47]。Eveno 等人[45]利用发声过程中音频与唇动的同步变化关系进行多模态身份认证的活体检测,并设计出一种发音唇动一致性的检测评分机制,并以此判断语音是否由视频中的人物实时说出。Shogo Kumagai 等[43]根据语音与唇动协同性区分新闻视频中的旁白场景和演讲场景,利用嘴唇形状和张开度描述说话人的嘴唇动作,利用音量和音素(12 维倒谱系数特征及其一阶差分)描述语音特征,然后在此基础上计算出 104 维互相关系数,利用互相关系数训练支持向量机(SVM)分类器,这种方法可以检测"假唱"类的录音回放攻击。该方法对旁白场景的检测效果比较理想,对演讲场景的检测效果不理想(准确率 53.3%)。以录音回放检测为目的,朱铮宇等[44]提出了基于时空相关度融合的语音唇动一致性检测算法。该方法构建了一个发音唇动时空模型,并以此为基础度量唇动时域特征、空域特性与语音的相关度,最后将时空上的相关度评分进行融合以判断语音唇动是否一致。对于 4 种不一致音视频数据,该方法比常用的协惯量方法降低了8.2%的等错误率(Equal error rate,EER)。

结合人脸等图像信息的说话人认证方法的有效性受制于光照条件和认证人员的配合及熟练程度,人脸转动对认证性能有很大的影响。另外,这类方法不适用于电话网络,目前的电话还难以实现音视频信息的同时采集。随着电信增值业务的发展,电话系统对声纹身份认证具有最大的需求。如养老金发放的身份认证,由于退休人员的身体状况和生活流动性等因素,采用声纹进行身份认证是最为合适的。因此多模态说话人认证方法并不是一个普适方法,其应用将受到一定环境等条件的约束。

另一类多模态说话人识别是采用音频麦克风和非音频麦克风两种途径获取音频信息,比如文献[48]结合骨导麦克风获得语音和常规语音进行说话人认证研究,降低了等错误概率。

2.3 基于语音随机性的检测方法

语音产生是一个随机过程,因此不可能获得两个完全相同的语音样本。录音是合法认证语音的复制品,即使考虑到环境噪声的影响,与原声的相似度非常高,特别是时间长度。研究人员利用这一性质发展出了一些回放攻击检测方法。Wei Shang 和 M. Stevenson[14-15]在分析影响回放攻击检测的各种因素的基础上,提出了一种基于 Peakmap 音频特征的针对远程电话交互式系统防回放攻击的算法,该方法计算语音与注册语音的 Peakmap 特征相似性,若相似度高于某一阈值,则判定测试语音为回放攻击;采用相对相似度[16]后,录音回放检测性能得到了显著提升,这一效果对受信道干扰的录音攻击检测更为突出。Wei Shang 和 M. Stevenson 方法的主要缺点在于:

(1)只能够应用于文本相关的声纹认证系统,对于文本无关认证系统无效;

(2)用户每次认证语音样本都要存下来,随着使用时间的增长,需要的存储空间越来越大;

(3)攻击检测时间长,每一次用户进入系统的语音样本都要和所有的存贮的样本进行相似性比较,随着系统时间增长,存储的样本会越来越多,相似性比较的计算时间也越来越长;

(4)如果回放语音不是在用户进入系统时录制的(也有可能是拼接得到的),即使对应的文本一样,该方法无效。

对于文本相关的说话人认证,2011 年 Jesus Villalba 等[18]研究了电话信道上录音回放攻击和音节拼接攻击的检测方法,其采用动态时间规正(Dynamic time warping, DTW)算法比较测试语音和注册语音的基频和 MFCC (Mel-frequency cepstral coefficients)特征的包络,依据相似性判断测试语音是否为拼接语音,而采用 SVM 判断测试语音是否为回放语音。该方法的本质相当于为发音状态建模,通过对注册的合法语音的基音和 MFCC 特征建立模板,然后将测试语音特征模板与存储模板进行比较。对于录音回放攻击,则采用 SVM 分别给原始语音和回放语音建立分类判决模型,对每一认证语音样本进行评分。该方法的困难在于实际中很难获得大量的回放语音来建立统计模型,而且也不能事先对未知的回放攻击信道进行预测。

2.4 基于信道特征模式分析的检测方法

最近几年,有学者开始研究直接从语音信号本身来寻找录音回放留下的一些痕迹,并将其直接用来进行回放检测。该方法对声纹身份认证系统并没有额外的生物特征采集设备要求,也不需要做多模态的特征融合,具有适应范围广、简单易行的特点。2008 年张利鹏等通过高保真录音设备转录过的数据与原始数据在信道上的差异来进行回放攻击检测[17],认为静音不受语音信号的影响能更好地体现信道信息,从静音中提取信道信息。以通用背景模型为基础,利用语音数据中的静音段对信道进行建模,检测待认证语音与训练语音的信道是否相同。其不足有 3 点:(1)静音段幅度小,比语音段更容易受到噪声污染;(2)以语音段的通用背景模型为基础并不一定能够训练出精确的信道模型;(3)认证语音样本中的音段时长难以保障,从而检测的可靠性也难以保证。

为了克服上述方法的不足,王志锋等[25,49]提出了信道模式噪声的概念,采用去噪滤波器和统计帧分析的方法从整个认证语音样本中提取信道模式噪声,在去噪滤波器设计时考虑了信道噪声主要分布在低频部分且频带比较窄的特点[50],并用 6 阶 Legendre 系数及 6 个统计特征表征信道模式噪声,建立基于支持向量机的信道模型判决模型,实验结果表明,回放攻击检测的准确度达到 97%以上。为了更好地获得信道模式噪声特征,文献[51]尝试建立认证语音产生过程模型,分析回放语音和原始认证语音的差异性,为回放检测提供理论依据;进一步考虑到信道噪声为时变信号,采用固定结构的去噪滤波器难以准确提取出信道模式噪声,而经验模态分解滤波(Empirical mode decomposition (EMD)-based filtering)[52]能够自适应选择本征模函数(Intrinsic mode function,IMF),从而有效地去除低频信道噪声的影响,实验结果表明,录音回放攻击取证的等错误率比采用常规去噪滤波器算法下降了 4.23%。

基于认证语音的信道特征分析的回放检测可以看成一个设备源识别问题[48],为了使得该方法更实用,合法信道可以通过注册语音的获取过程获得其特征表示,并建立相应的模型,而认证过程中的回放检测问题可简化为“认证语音仅来自注册信道吗”,或者说,检测认证语音的信道特征是否被污染,原理上讲,该思路就可应对各种不同的录音回放设备。

3 结束语

以上录音回放检测方法各有优缺点,应用场景也有所不同,在实际应用中,为了更好地检测各种可能的攻击,往往是多种应对措施结合,比如文献[38,44]将唇动信息和语音信息相结合,利用二者的内在相关性进行回放检测。录音回放检测方法的发展应该是从认证语音的真实性角度考虑,由于认证语音的真实性包括内容真实性、说话人真实性、时间真实性以及来源真实性等多个方面,根据检测真实的重点不同,所利用的方法也不相同。

本文对录音回放危害性及已发展的检测方法进行了较全面的介绍,该问题的研究还处于起步阶段[54],有许多亟待解决的问题。例如,攻击检测算法的性能不高,离实际应用有一定的差距;现今的攻击检测算法只能针对单一攻击,对组合攻击还未开展研究;在提高系统安全性的同时,如何保证身份认

证系统的准确度以及效率,即安全与性能的关联性问题。

参考文献:

[1] Shim K A. On the security of a certificateless aggregate signature scheme[J]. Communications Letters, IEEE, 2011, 15(10): 1136-1138.

[2] Prabhakar S, Ivanisov A, Jain A K. Biometric recognition: Sensor characteristics and image quality[J]. Instrumentation & Measurement Magazine, IEEE, 2011, 14(3): 10-16.

[3] Zhu Donglai, Ma Bin, Li Haizhou. Speaker verification with feature-space MAPLR parameters[J]. Audio, Speech, and Language Processing, IEEE Transactions on, 2011, 19(3): 505-515.

[4] You Changhuai, Lee Kongaik, Li Haizhou. GMM-SVM kernel with a Bhattacharyya-based distance for speaker recognition [J]. Audio, Speech, and Language Processing, IEEE Transactions on, 2010, 18(6): 1300-1312.

[5] Campbell J P, Shen W, Campbell W M, et al. Forensic speaker recognition[J]. Signal Processing Magazine, IEEE, 2009, 26(2): 95-103.

[6] Che Chiwei, Lin Qiguang, Yu Dongsu. An HMM approach to text-prompted speaker verification[C]// ICASSP. Atlanta: IEEE, 1996: 673- 676.

[7] 王晓庆,覃敏. 携程漏洞门引发公众担忧,快捷支付隐患显露[EB/OL]. http://finance.qq.com/a/20140331/005104.htm, 2014-3-31.

[8] Yun Lei, Scheffer N, Ferrer L, et al. A novel scheme for speaker recognition using a phonetically-aware deep neural network [C]// ICASSP. Florence: IEEE, 2014:1695-1699.

[9] Gerazov B, Pop-Dimitrijoska V, Ivanovski Z, et al. Use of gaussian mixture models in macedonian forensic speaker identification[C]// 20th Telecommunications Forum on Forensic Speaker Rrecognition Case Pre-assessment. TELFOR, 2012: 724-727.

[10] Cao Honglin, Kong Jiangping. Speech length threshold in forensic speaker comparison by using long-term cumulative formant (LTCF) analysis[C]// Second International Conference on Instrumentation, Measurement, Computer, Communication and Control. Hangzhou: IEEE, 2012: 418-421.

[11] NIST. 2012 NIST SRE report[EB/OL]. http://www.nist.gov/itl/iad/mig/sre12results.cfm. 2012-05-20.

[12] Mathilta. 阿里巴巴的下一个大动作在身份识别[EB/OL]. http://www.huxiu.com/article/33451/1.html,2014-05-27.

[13] ITU. ITU-T X 1086,Telebiometrics protection procedures-part 1: A guideline to technical and managerial countermeasures for biometric data security[S]. Geneva:[s.n.],2007.

[14] Wei Shang, Stevenson M. A playback attack detector for speaker verification systems[C]// Communications, Control and Signal Processing, 3rd International Symposium on. Australia: IEEE, 2008: 1144-1149.

[15] Wei Shang, Stevenson M. A preliminary study of factors affecting the performance of a playback attack detector[C]// CCECE. Canada: IEEE, 2008:459-464.

[16] Wei Shang, Stevenson M. Score normalization in playback attack detection[C]// ICASSP. Dallas: IEEE, 2010: 1678-1681.

[17] 张利鹏,曹翠,徐明星,等. 防止假冒者闯入说话人识别系统[J]. 清华大学学报:自然科学版, 2008, 48(S1): 699-703.
Zhang Lipeng, Cao Jiang, Xu Mingxing, et al. Prevention of impostors entering speaker recognition systems[J]. Journal of Tsinghua University: Science and Technology, 2008, 48(S1): 699-703.

[18] Villalba J, Lleida E. Preventing replay attacks on speaker verification systems[C]// ICCST. San Franci: IEEE, 2011: 1-8.

[19] Villalba J, Lleida E. Speaker verification performance degradation against spoofing and tampering attacks[C]// FALA workshop. [S.l.]: IEEE, 2010: 131-134.

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

[21] Farrs M, Wagner M, Anguita J, et al. How vulnerable are prosodic features to professional imitators? [C]// Odyssey: 2008, The Speaker and Language Recongnition Workshop. Stellenbosch: ISCA,2008:5-8.

[22] Jin Qin, Toth A R, Black A W, et al. Is voice transformation a threat to speaker identification? [C]// ICASSP. Las Vegas: IEEE, 2008: 4845-4848.

[23] De Leon P L, Apsingekar V R, Pucher M, et al. Revisiting the security of speaker verification systems against imposture using synthetic speech[C]// ICASSP. Dallas: IEEE, 2010: 1798-1801.

[24] Kinnunen T, Wu Z Z, Lee K A, et al. Vulnerability of speaker verification systems against voice conversion spoofing attacks: The case of telephone speech[C]// ICASSP. Kyoto: IEEE, 2012: 4401-4404.

[25] Wang Zhifeng, Wei Gang, He Qianhua. Channel pattern noise based playback attack detection algorithm for speaker recognition[C]// Machine Learning and Cybernetics, Proceedings of the 2011 International Conference on. Guilin: IEEE, 2011: 1708-1713.

[26] 丁琦,平西建. 针对语音变换的语音篡改检测[J]. 数据采集与处理,2012, 27(1): 57-62.
Ding Qi, Ping Xijian. Speech tampering detection for voice transformation[J]. Journal of Data Acquisition and Processing, 2012, 27(1): 57-62.

[27] 钟巍,孔祥维,尤新刚,等. 基于分数倒谱变换的取证语音拼接特征提取与分析[J]. 数据采集与处理,2014, 29(2): 248-253.
Zhong Wei, Kong Xiangwei, You Xingang, et al. Splicing feature extraction and analysis based on fractional cepstrum transform in voice forensics[J]. Journal of Data Acquisition and Processing, 2014, 29(2): 248-253.

[28] Larcher A, Lee K A, Ma B, et al. Imposture classification for text-dependent speaker verification[C]// ICASSP. Florence: IEEE, 2014: 739-743.

[29] Min H P. Challenge-response based ACK message authentication[J]. Electronics Letters, 2012, 48(16): 1021-1023.

[30] Delacretaz D P, Hennebert J. Text-prompted speaker verification experiments with phoneme specific MLPs[C]// ICASSP. Seattle: IEEE, 1998:777-780.

[31] Johnson R C, Boult T E, Scheirer W J. Voice authentication using short phrases: Examining accuracy, security and privacy issues[C]// Biometrics: Theory, Applications and Systems (BTAS), IEEE Sixth International Conference on. Washington DC : IEEE, 2013: 1-8.

[32] Baloul M, Cherrier E, Rosenberger C. Challenge-based speaker recognition for mobile authentication[C]// Biometrics Proceedings of the International Conference of the Special Interest Group (BIOSIG). Darmstadt: IEEE, 2012: 1-7.

[33] Ross A, Jain A K. Multimodal biometrics: An overview[C]// EUSIPCO. Vienna: EURASIP, 2004: 1221-1224.

[34] Deriche M. Trends and challenges in mono and multi biometrics[C]// Image Processing Theory, Tools & Applications, 2008 IEEE First-Workshop on. Sousse: IEEE, 2008:1-9.

[35] Bredin H, Miguel A, Witten I H, et al. Detecting replay attacks in audiovisual identity verification[C]// ICASSP. Toulouse: IEEE, 2006: 621-624.

[36] Cetingul H E, Yemez Y, Erzin E, et al. Discriminative lip-motion features for biometric speaker identification[C]// ICIP. Singapore: IEEE, 2004: 2023-2026.

[37] Li Y , Narayanan S, Kuo C J. Audiovisual-based adaptive speaker identification[C]// ICASSP. Hong Kong: IEEE, 2003: 812-15.

[38] Ichino M, Sakano H, Komatsu N. Multimodal biometrics of lip movements and voice using kernel fisher discriminant analysis[C]// Control, Automation, Robotics and Vision, 9th International Conference on. Singapore: IEEE, 2006: 1-6.

[39] Faraj M, Bigun J. Synergy of lip-motion and acoustic features in biometric speech and speaker recognition[J]. Computers, IEEE Transactions on, 2007, 56(9): 1169-1175.

[40] Fox N A, Gross R, Cohn J F, et al. Robust biometric person identification using automatic classifier fusion of speech, mouth, and face experts[J]. Multimedia, IEEE Transactions on, 2007, 9(4): 701-714.

[41] Slaney M, Covell M. Facesync: A linear operator for measuring synchronization of video facial images and audio track[C]// Neural Information Processing Systems 2000(NIPS). Denver: NIPS, 2000: 814-820.

[42] Chetty G, Wagner M. Multi-level liveness verification for face-voice biometric authentication[C]// Biometric Consortium Conference. Baltimore: BSYM, 2006: 1-6.

[43] Kumagai S, Doman K, Takahashi T, et al. Detection of inconsistency between subject and speaker based on the co-occurrence of lip motion and voice towards speech scene extraction from news videos[C]// Multimedia, IEEE International symposium on. California: IEEE, 2011: 311-318.

[44] 朱铮宇,贺前华,奉小慧,等. 基于时空相关度融合的语音唇动一致性监测算法[J]. 电子学报, 2014, 42(4): 779-785.
Zhu Zhengyu, He Qianhua, Feng Xiaohui, et al. Lip motion and voice consistency algorithm based on fusing spatiotemporal correlation degree[J]. Chinese Journal of Electronics, 2014, 42(4): 779-785.

[45] Eveno N, Besacier L. A speaker independent "liveness" test for audio-visual biometrics[C]// Nineth European Conference on Speech Communication and Technology. Lisbon: ISCA, 2005: 3081-3084.

[46] Bredin H, Chollet G. Making talking-face authentication robust to deliberate imposture[C]// ICASSP. Las Vegas: IEEE, 2008: 1693-1696.

[47] Chetty G. Biometric liveness detection based on cross modal fusion[C]// Information Fusion, 12th International Conference on. Seattle: IEEE, 2009: 2255-2262.

[48] Tsuge S, Koizumi D, Fukumi M, et al. Speaker verification method using bone-conduction and air-conduction speech[C]//

Intelligent Signal Processing and Communication Systems, 2009 International Symposium on. Kanazawa: IEEE, 2009: 449-452.

[49] 王志锋, 贺前华, 张雪源, 等. 基于信道模式噪声的录音回放攻击检测[J]. 华南理工大学学报:自然科学版, 2011, 39(10): 7-12.
Wang Zhifeng, He Qianhua, Zhang Xueyuan, et al. Playback attack detection based on channel pattern noise[J]. Journal of South China University of Technology: Natural Science Edition, 2011, 39(10): 7-12.

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

[51] 王志锋. 基于信道信息的数字音频取证关键问题研究[D]. 广州: 华南理工大学电子与信息学院, 2013.
Wang Zhifeng. Research on key issues of digtial audio forensics with channel information[D]. Guangzhou: School of Electronic and Information Engineering, South China University of Technology, 2013.

[52] Chatlani N, Soraghan J J. EMD-based filtering (EMDF) of low-frequency noise for speech enhancement[J]. Audio, Speech, and Language Processing, IEEE Transactions on, 2012, 20(4): 1158-1166.

[53] 胡永健, 刘琲贝, 贺前华. 数字多媒体取证技术综述[J]. 计算机应用, 2010, 4(3): 657-662.
Hu Yongjian, Liu Beibei, He Qianhua. Survey on techniques of digital multimedia forensics[J]. Journal of Computer Applications, 2010, 4(3): 657-662.

[54] Alegre F, Vipperla R, Evans N, et al. On the vulnerability of automatic speaker recognition to spoofing attacks with artificial signals[C]// EUSIPCO. Bucharest: EURASIP, 2012: 36-40.

作者简介:**贺前华**(1965-),男,教授,研究方向:语音信号处理,E-mail:eeqhhe@scut.edu.cn;**潘伟锵**(1972-),男,高级工程师,研究方向:语音通信;**胡永健**(1962-),男,教授,研究方向:数字图像取证;**朱铮宇**(1984-),男,博士研究生,研究方向:音视频一致性分析;**李艳雄**(1980-),男,讲师,研究方向:音频事件检测;**奉小慧**(1981-),女,讲师,研究方向:多模态语音信号处理。