ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 35,No. 5,Sep. 2020,pp. 807－823
DOI:10. 16337/j. 1004-9037. 2020. 05. 002

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

语音欺骗检测方法的研究现状及展望

张雄伟，李嘉康，孙　蒙，郑琳琳

（陆军工程大学指挥控制工程学院，南京，210007）

摘　要：语音欺骗是指通过录音、语音合成（Text-to-speech，TTS）、语音转换（Voice conversion，VC）等手段，将一段非法的、未经过自动说话人验证（Automatic speaker verification，ASV）系统认证的声音进行“修改仿冒”，以达到通过ASV系统检测的目的。随着人工智能和语音欺骗技术的发展，ASV系统在安全性方面遇到了严峻的挑战。检测输入ASV系统的语音的真实性，防止欺骗语音通过ASV的验证以提高ASV系统的安全性，是近年来语音领域研究的一个热点问题。国内外学者的最新研究从声学特征选取、识别模型选择等角度出发，探索了不同的语音欺骗方法对ASV系统的影响，并深入研究了相应的语音欺骗检测技术，在一定程度上提高了ASV系统的防欺骗性能。本文介绍了语音欺骗的基本方法，给出了语音欺骗检测的框架和典型声学特征，分两大类别总结了语音欺骗检测的主要方法和最新进展，梳理了目前语音欺骗检测中仍然存在的若干技术问题，并对语音欺骗检测技术的发展方向进行了展望。

关键词：语音欺骗检测；语音合成（TTS）；语音转换（VC）；说话人验证；ASVspoof
中图分类号：TN912　　**文献标志码**：A

Speech Anti-spoofing: The State of the Art and Prospects

ZHANG Xiongwei，LI Jiakang，SUN Meng，ZHENG Linlin

（College of Command and Control Engineering，Army Engineering University，Nanjing，210007，China）

Abstract：Speech spoofing refers to the technology of counterfeiting an illegal speech without the authentication by automatic speaker verification（ASV）system to the speech of a legally authenticated speaker by ASV through recording，text-to-speech，voice conversion and other means，so as to achieve the goal of passing the ASV system. With the development of artificial intelligence and speech anti-spoofing methods，ASV systems have encountered severe challenges in security. It is a hot topic in the field of speech research in recent years to detect the authenticity of the speech input to the ASV system and to prevent spoof speech from passing the verification of ASV to improve the security of the ASV system. The latest research of scholars at home and abroad explores the influence of different speech spoofing methods on ASV system from the perspective of acoustic feature and recognition model，and further studies the corresponding speech anti-spoofing technology，which improves the anti-spoofing ability of ASV systems to a certain extent. This paper summarizes the latest methods of speech spoofing to ASV systems and the latest anti-spoofing methods，focusing on the state-of-the-art research results around the world，and prospects the development direction of speech anti-spoofing technology.

Key words: speech anti-spoofing; text-to-speech; voice conversion; speaker verification; ASVspoof

基金项目：国家自然科学基金（61471394，62071484）资助项目；江苏省优秀青年基金（BK20180080）资助项目。
收稿日期：2020-05-20；**修订日期**：2020-07-22

引　言

近年来,基于生物识别的身份认证技术在数据安全和通过性认证中的作用越来越重要。一些常用的生物识别技术,如指纹识别、人脸识别和声纹识别等,已经在多种认证场景中得到了较为广泛的应用,给人们的生活带来了极大的便利。人们每天都要使用的手机,其解锁方式就有人脸识别、指纹识别等,微信的声纹锁也允许使用语音进行登录认证。在众多的生物识别技术中,人类的语音由于采集方便、区分度高,采集声音使用的麦克风等设备发展成熟、成本较低,因而受到了广泛的关注,自动说话人验证(Automatic speaker verification, ASV)系统也应运而生。ASV系统是一个典型的生物识别系统,该系统可以使用特定的算法对输入语音进行模式识别和匹配,判断出该待验证的说话人语音是否为合法用户的声音。随着近年来机器学习和深度学习的发展,ASV系统的识别准确率越来越高,对ASV系统的研究是当前生物识别研究的一个热点问题。

但是,任何生物识别技术都存在一定的缺陷。通过模仿、篡改特征等方法对生物特征进行修改,有可能达到非法通过生物识别系统验证的目的,这给生物识别系统的安全性带来了严峻挑战。例如,在人脸识别验证中,一个较为典型的欺骗方法就是使用已经通过验证的合法用户的照片来欺骗识别系统。因此,为了实现生物识别系统的安全性,系统必须能够准确判断输入的生物特征的真伪,对合法的用户生物特征正常接受,而对假冒的、非法的生物特征必须予以拒绝。

目前有4种典型的ASV系统语音欺骗方法:语音模仿、语音回放、语音合成(Text-to-speech, TTS)与语音转换(Voice conversion, VC)[1]。语音欺骗方法早在20世纪六七十年代就已经产生,但国际上对于语音欺骗检测的广泛关注则开始于最近十年。2013年,法国里昂举办Interspeech会议期间,召开了“ASV系统的欺骗和对策”特别会议[2],将语音欺骗检测引入了人们的关注热点。随后,2015年在德国德累斯顿的Interspeech会议期间,举行了第1次ASVspoof挑战赛,该挑战赛旨在提供一个通用的语音欺骗检测数据集和评价标准,促使人们开发出能够检测出真实语音和欺骗语音的方法。ASVspoof 2015挑战赛重点关注对TTS和VC的欺骗检测[3],该项赛事吸引了来自全世界16个国家共27支团队,掀起了语音欺骗检测的研究热潮[4]。2017年在瑞典斯德哥尔摩举行的ASVspoof 2017挑战赛则专注于语音回放的检测,这次比赛开放了语音回放检测的通用数据集,共收到来自全世界49支队伍提交的研究结果,为语音回放检测的广泛开展奠定了基础[5]。刚结束的ASVspoof 2019挑战赛则同时关注了语音回放、TTS和转换的欺骗方式,提供了2个数据库,分别针对TTS和转换欺骗,以及语音回放欺骗,这次比赛共吸引了69支队伍参加,是迄今为止针对语音欺骗检测规模最大、最全面的挑战赛[6]。

近年来,针对语音欺骗检测问题,清华大学、西北工业大学、哈尔滨工业大学、昆山杜克大学等多所国内高校以及百度、小米等多家企业都开展了相关研究,并且取得了一些优秀的研究成果。在ASVspoof 2019挑战赛上,来自“清华大学-得意音通”声纹处理联合实验室的团队取得了语音回放检测任务全球第1名的成绩。此外,中国人民银行在2018年发布的《移动金融基于声纹识别的安全应用技术规范(JR/T 0164—2018)》中也明确规范了移动金融领域中声纹识别技术需要具备的防欺骗功能,其中就包括了语音模仿、VC及合成、录音回放等。因此,语音欺骗检测是目前也是未来研究的热点。

本文介绍了常见的语音欺骗方法,重点阐述了国内外针对语音欺骗检测的最新研究进展,归纳分析了语音欺骗检测的典型方法,并展望语音欺骗检测未来的发展方向。

1　语音欺骗方法

说话人验证是一种通过说话人语音特征来验证说话人身份的技术,图1给出了一个典型的ASV系统的结构和验证流程。说话人验证系统是典型的模式识别系统,该系统可分为2个模块,分别为图中展

示的注册模块和验证模块。ASV的通用过程都是在注册模块预先存储说话人的语音特征,在测试验证模块提取待识别说话人的语音特征,与预先储存的注册特征进行对比,从而验证说话人的身份。

语音欺骗主要是针对ASV系统进行。在语音欺骗的处理阶段,非法的入侵者通过人为模仿已经通过注册的说话人的语音;或者使用录音设备偷偷录制注册说话人说出的语句;或者通过其他途径收集到的注册说话人的语音,使用TTS和转换的方法对入侵者自己的语音进行处理,使经过处理后的语音接近于注册说话人的语音。然后将处理后的语音馈送给ASV系统的麦克风,欺骗ASV系统获得准入权限,进而达到非法入侵的目的。

下文分别介绍语音欺骗的4种方法:语音模仿、语音回放、TTS和VC。

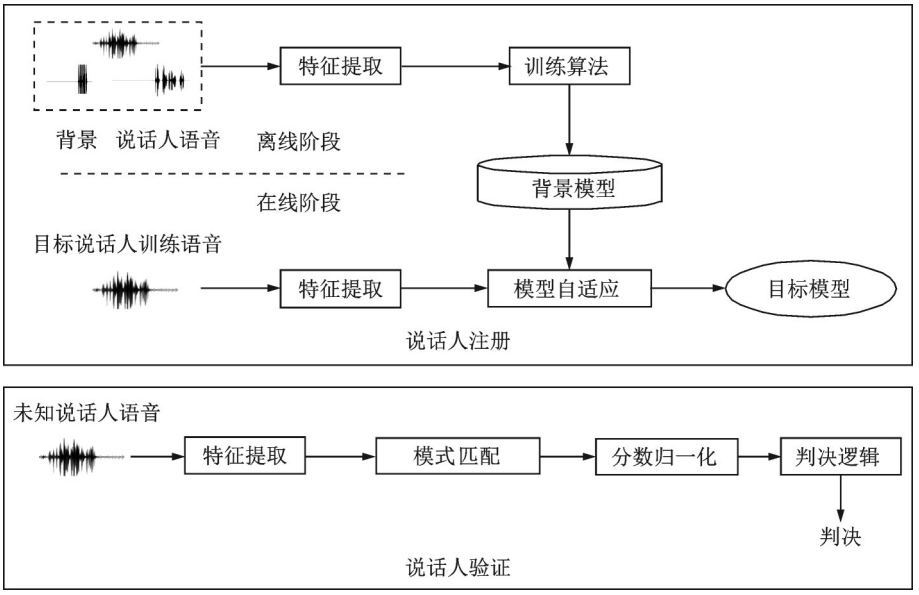


图1 典型的说话人验证系统

Fig.1 Typical automatic speaker verification system

1.1 语音模仿

语音模仿是指非法入侵者故意将其自己的声音模仿为已通过认证的目标说话人,通过模仿目标说话人说出的词汇、音色或者某些特殊的特征,使自己的声音尽可能听起来接近于目标说话人来实现对ASV系统的入侵。

语音模仿是较为简单的一种语音欺骗方法,但是该方法要求欺骗者和被模仿的注册说话人的声音较为接近,而且该方法的成功率普遍不高。

Farrus等[7]发现,专业的模仿者模仿出的语音通常在基频(F0)和共振峰频率两方面更加接近于目标说话人,他们使用若干个模仿者和一个被模仿目标说话人的韵律特征来量化模仿者的欺骗语音与目标说话人真实语音的接近程度,结果表明,与目标说话人的韵律特征更相似的模仿者的语音会增加ASV的误判率。Lau等[8]发现,如果目标说话人已知,而且模仿者与目标说话人的声音音色更加相似,那么欺骗ASV系统的成功率会大大提高。Mariéthoz等[9]的实验表明,专业的模仿者比普通的业余模仿者欺骗ASV的成功率更高。在最近的一项研究中评估了语音模仿欺骗方法对3种常见的ASV系统的影响[10],结果表明,语音模仿欺骗会导致这3种ASV系统的错误判断率提高。

因此,从总体上来看,语音模仿欺骗对ASV的安全性具有一定的威胁,会造成系统的错误识别率增

加,但是具体影响效果与模仿者的专业程度和ASV系统的识别性能有关。

1.2 语音回放

语音回放是指使用预先录制的已通过认证的目标说话人的语音,通过某些播放设备将录制好的语音播放出来馈送给ASV系统的麦克风。语音回放欺骗不需要任何专业知识或者复杂的设备,仅需要一个简单的录音和播放设备即可,因此非常易于实施。虽然语音回放欺骗的操作简单、成本低廉,但是却会给ASV系统带来严重的安全性问题。这种欺骗方法给ASV系统造成的影响要远高于语音模仿欺骗,语音回放欺骗会造成ASV系统的错误接受率(False accept rate, FAR)明显提高。

在针对语音回放欺骗的ASVspoof 2017语料库[11]发布之前,关于语音回放欺骗的研究非常有限。早期的研究主要集中在语音回放对ASV系统造成的影响。Lindberg等[12]研究了语音回放对文本相关的ASV系统的影响,结果表明,语音回放欺骗使得基于隐马尔可夫模型(Hidden Markov model, HMM)的ASV系统的男性说话人的FAR从1%增加到了89%,女性说话人的FAR从5%增加到了100%,由此可见,语音回放可以对ASV系统造成非常严重的误判。Villalba等[13]调研了远场录制的语音对文本无关的ASV系统的影响,试验结果表明,当使用回放语音对基于联合因子分析(Joint factor analysis, JFA)ASV系统进行欺骗时,ASV的等错误率(Equal error rate, EER)从1%增加到了将近70%。Wang等[14]使用语音回访欺骗对基于高斯混合模型(Gaussian mixture model, GMM)-通用背景模型(Universal background model, UBM)的ASV进行了验证,发现在语音回放欺骗下ASV的FAR为93%。Ergunay等[15]比较了不同质量的录音设备和播放设备对ASV系统的影响,其结果表明,使用高质量设备进行录音和回放时,ASV系统的FAR更高,说明了设备的质量高低也会影响ASV的准确性。此外,播放设备距离ASV的麦克风距离的远近也会影响ASV的准确性,距离越远,ASV的FAR越高。

1.3 TTS

TTS通常也称为文本到语音的转换,是一种可以将任意文本信息生成可以理解的语音的技术。TTS的应用非常广泛,包括日常生活中常用的导航系统、人机交互系统以及语言翻译系统等。TTS系统主要由2部分组成[16]:文本分析和波形生成。在文本分析中,输入的文本被转换成由单个音素组成的单元;在波形生成阶段,将各个单元合成语音的波形。在最新的端到端TTS框架中,可以直接将输入的文本信息转化为语音的波形,不需要使用其他附加模块。

随着机器学习的发展,基于参数统计的TTS成为20世纪末流行的TTS方法之一[17-18]。在这种方法中,通常使用基于时间序列的生成模型(一般为HMM)对声学参数进行建模。HMM不仅可以表示音素序列,还表示根据语音规范生成的上下文。然后使用从HMM生成的声码器生成语音波形。此外,基于HMM的TTS方法还可以使用UBM模型中的自适应技术[19],从相对较少的说话人数据中学习到特定说话人的语音模型。

近年来,深度学习的应用进一步提高了TTS的质量。首先,使用各种类型的深度神经网络提高了声学参数的预测精度[20]。常用的深度神经网络包括循环神经网络(Recurrent neural network, RNN)[21]、残差神经网络(Residual network, Resnet)[22]和生成对抗网络(Generative adversarial network, GAN)[23]等。此外,传统的基于信号处理方法的传统波形生成模块和使用自然语言处理的文本分析模块被神经网络替代,神经网络能够直接从输入的特定文本生成相对应的波形输出,可以直接对语音的波形进行建模,这种方法称为“Wavenet”。这些新型的深度学习方法可以使人工合成的语音听起来几乎和人类真实的语言一样自然[24]。

TTS的方法对ASV系统具有很强的威胁性,除了简单的语音波形拼接之外,基于HMM的语音合成方法可导致基于HMM的文本相关ASV系统的FAR从正常状态下的7%增加到70%以上[25]。De

Leon等[26]使用了基于HMM的TTS方法,在基于GMM-UBM的ASV系统和基于支持向量机(Support vector machine, SVM)的ASV系统上分别进行了测试,使其FAR分别上升到了86%和81%,结果表明TTS欺骗方法对于各种ASV系统都具有很强的威胁性。

1.4 VC

VC[27]旨在将一个说话人的声音转换为另一个说话人的声音,与TTS不同的是,VC直接在输入的语音上进行,不需要将文本转化为波形这一步操作。大多数的VC需要平行语料,即要求源说话人和目标说话人要说出相同的语音内容,并且需要源语音和目标语音的每一帧对齐。

当语VC应用于语音欺骗时,目标就是将输入的非法语音转换成新的语音信号,使得新的语音信号在某种意义上与已经通过认证的目标说话人更加相似。Perrot等[28]发现VC可以对文本无关的ASV系统造成严重的影响,当所有已注册的合法说话人语音被转换后的语音替换后,ASV系统的EER从10%增加到了60%。Kinnunen等[29]使用基于联合密度高斯混合模型(Joint density Gaussian mixture model, JDGMM)的VC方法对5种不同的ASV系统进行了测试,结果表明,即使是性能最强大的JFA系统,其在VC的欺骗方法下的FAR也从3%增加到了17%。

2 语音欺骗检测方法

语音欺骗检测是为了能够检测出输入到ASV系统的各种欺骗语音,保护ASV系统免受不法用户的侵害,提高ASV系统的安全性。本节首先给出语音欺骗检测的总体框架,并以ASVspoof 2015、2017、2019这3届挑战赛为重点,梳理总结目前语音欺骗检测的主要方法。

2.1 语音欺骗检测概述

受到语音欺骗检测数据集的限制,当前国际上对语音欺骗的几种方法还没有统一的普适性的检测手段。对于语音模仿欺骗,目前没有通用的数据集支持此项研究,同时由于语音模仿需要较为专业的模仿者,即使找到了专业的模仿人员,对于目前较为先进的ASV系统,语音模仿欺骗成功的成功率也并不高,因此,语音模仿欺骗不是当前研究的重点。对于语音回放、TTS和VC这3种语音欺骗方法,由于回放和另外两种方法所使用的技术差别较大,而TTS和VC所使用的技术具有一定的相似性,因此国际上主要将语音欺骗检测分为2大类,一类是语音回放欺骗检测,另一类则是TTS和VC欺骗检测。

2.1.1 语音欺骗检测框架

当前国际上先进的语音欺骗检测方法都是设计一个与ASV系统独立的、互不关联的欺骗检测系统。当进行欺骗检测和说话人验证时,首先对语音样本输入到欺骗检测系统中进行安全性验证,只有通过欺骗检测系统,被判定为是真实语音的样本,才能够输入到ASV系统中进行认证。一个典型的语音欺骗检测系统如图2所示。

由于当前的语音欺骗检测系统只能单独检测一种语音欺骗,例如单独检测语音回放欺骗,或者单独检测TTS与转换的欺骗。因此,如果语音的欺骗方法未知,那么就需要将各种欺骗检测系统串联起

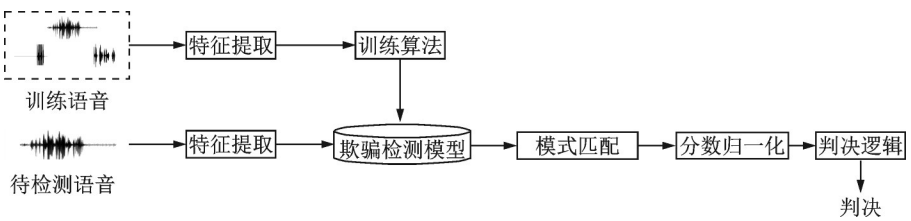


图2 典型的语音欺骗检测系统

Fig.2 A typical speech anti-spoofing system

来,分别进行检测,只有通过了所有语音欺骗检测系统的验证,才能够输入到ASV系统中进行说话人验证。整体的语音欺骗检测流程如图3所示,其中2个欺骗检测模型可以分别是语音回放欺骗检测模型和TTS与转换欺骗检测模型。

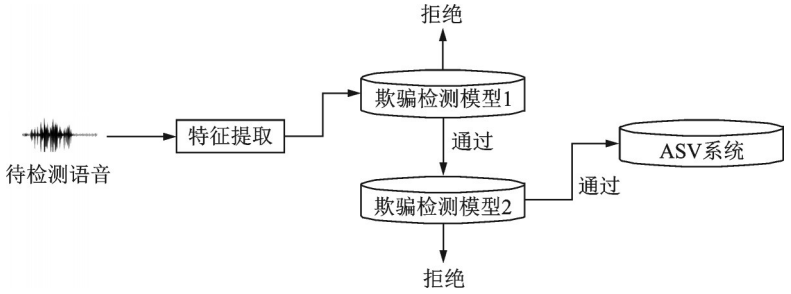


图3 语音欺骗检测流程

Fig.3 Process of speech anti-spoofing

2.1.2 评价指标

EER是评价ASV系统性能和语音欺骗检测性能的常用指标。对于说话人验证任务来说,EER是错误拒绝率(False rejection rate, FRR)和FAR相等时的数值,EER能够同时反映出系统的安全性和准确性,是衡量生物识别系统性能的重要指标。

在说话人验证系统中,ASV会判定2个语音样本是否属于相同的说话人,对比后会得到2个语音样本相似度的得分,如果得分大于某一事先设定好的阈值θ,则判定这2个语音样本来自同一个说话人,如果得分小于该阈值θ,则判定这2个语音样本来自不同的说话人。如果2个语音样本实际上属于相同的说话人,但是被ASV系统判定为不同的说话人,则称之为错误拒绝案例,FRR为错误拒绝案例在ASV系统认定为相同说话人案例中所占的比值,即

$$\mathrm{FRR}(\theta)=\frac{N_{\text{相同说话人但判定得分}\leqslant\theta}}{N_{\text{同类匹配案例}}} \tag{1}$$

式中,同类匹配案例即为应当被系统认定为相同说话人的案例。如果2个语音样本实际上属于不同的说话人,但是ASV系统判定为相同的说话人,即为错误接受案例,FAR为错误接受案例在ASV系统判定为不同说话人的案例中所占的比值,即

$$\mathrm{FAR}(\theta)=\frac{N_{\text{不同说话人但判定得分}>\theta}}{N_{\text{异类匹配案例}}} \tag{2}$$

式中,异类匹配案例即为应当被系统判定为不同说话人的案例。EER则定义为通过调整阈值为θ_{EER}时,FRR和FAR相等的数值,即

$$\mathrm{EER}=\mathrm{FRR}(\theta_{\mathrm{EER}})=\mathrm{FAR}(\theta_{\mathrm{EER}}) \tag{3}$$

这里提供计算EER使用的Bosaris工具箱。

在语音欺骗检测中,EER也和ASV系统中的EER计算方式类似,式(4)给出语音欺骗检测中的FRR、FAR和EER的计算公式

$$\begin{cases}\mathrm{FRR}(\theta)=\dfrac{N_{\text{真实语音但判定得分}\leqslant\theta}}{N_{\text{真实样本案例}}}\\ \mathrm{FAR}(\theta)=\dfrac{N_{\text{欺骗语音但判定得分}>\theta}}{N_{\text{欺骗样本案例}}}\\ \mathrm{EER}=\mathrm{FRR}(\theta_{\mathrm{EER}})=\mathrm{FAR}(\theta_{\mathrm{EER}})\end{cases} \tag{4}$$

在评价语音欺骗检测系统的性能时,如果事先指定的阈值θ过高,则会造成FRR增大,可能会造成大量真实的语音被判定为欺骗语音,给合法用户的准入造成不便;而指定的阈值θ过低,则会导致FAR提高,可能会造成大量欺骗语音被判定为真实语音,给系统的安全性造成危害。因此,EER既可以显示出欺骗检测系统的安全性,又可以显示出合法用户通过认证的可靠性,是评价语音欺骗检测系统的重要指标。

2.1.3 数据集

自2015年以来,每隔两年Interspeech就会举办一次专门针对语音欺骗检测的ASVspoof挑战赛,至今共举办了3届,每一届挑战赛都会发布专门的数据集供研究者使用。

ASVspoof 2015数据集[4]专门针对TTS和VC欺骗检测,该数据集由真实语音和欺骗语音组成。真实语音共由106名不同的说话人录制,包括45名男性和61名女性,没有对录制语音进行任何修改,并且是在干净的背景环境中进行录制,没有明显的信道或背景噪声的干扰。录制好语音后,使用了3种TTS和7种VC的算法,对原始的真实语音进行变换,生成欺骗语音。整个数据集共分为3个子集:训练集、开发集和验证集,可以使用训练集和开发集进行语音欺骗检测模型的训练和调试,用训练好的模型在验证集上进行测试,得到最终的判别结果。表1给出了ASVspoof 2015数据集的具体情况。

ASVspoof 2017挑战赛专门针对语音回放欺骗检测,该语料库来源于RedDots(https://sites.google.com/site/thereddotsproject/)。该语料库由来自全球各地的ASV研究人员使用Android智能手机进行收集和录制。ASVspoof 2017数据集中的真实语音是原始RedDots语料库中的一个子集,而回放的语音则是这些原始语音通过不同种类的设备播放后再录制的。该数据集也分为训练集、开发集和验证集3部分,表2给出了数据的具体信息。

表1 ASVspoof 2015数据集详细信息

Table 1 Detailed information of ASVspoof 2015 corpus

数据集	说话人数量		语音数量	
	男性	女性	真实语音	欺骗语音
训练集	10	15	3 750	12 625
开发集	15	20	3 497	49 875
验证集	20	26	9 204	184 000

表2 ASVspoof 2017数据集详细信息

Table 2 Detailed information of ASVspoof 2017 corpus

数据集	说话人数量	回放场景	语音数量		
			真实语音	回放语音	总计
训练集	10	6	1 508	1 508	3 016
开发集	8	10	760	950	1 710
验证集	24	163	1 298	12 922	14 220
总计	42	179	3 566	15 380	18 946

ASVspoof 2019挑战赛同时针对语音回放欺骗检测和TTS转换欺骗检测,并为此分别设立了2个赛道和相对应的数据集。这2部分数据集都是基于VCTK数据库进行开发的(http://dx.doi.org/10.7488/ds/1994),同样划分为3个子集:训练集、开发集和验证集,分别由20名(8男12女)、10名(4男6女)和48名(21男27女)不同的说话人组成。在TTS与回放欺骗检测中,使用了17种不同的TTS和VC系统生成的真实语音和欺骗语音。这17种方法中,有6种方法被指定为已知的欺骗类型,另外11种指定为未知的欺骗类型。训练集和开发集中的欺骗语音的生成方法仅包含6种已知的欺骗方法,验证集包含2种已知的欺骗方法和11种未知的欺骗方法。在已知的6种欺骗方法中,有2个VC算法和4个TTS算法,11种未知欺骗方法中,包括2个VC算法、6个TTS算法和3个TTS-VC混合算法。这些算法中包含了一些经典的和当前最先进的TTS和转换方法,包括传统的语音编码、Griffin-Lim[30]、GAN[31]、神经波形模型[32]等。

与ASVspoof 2017数据集不同，ASVspoof 2019的语音回放欺骗检测数据集设定了更加详细的声学环境，包括录音的房间大小、混响的种类和播放设备到录音设备的距离等。表3，4给出了ASVspoof 2019数据集的详细信息。

以上3个数据集可以在ASVspoof官方网站（https://www.asvspoof.org/database）下载。

表3 ASVspoof 2019的TTS数据集详细信息

Table 3 Detailed information of ASVspoof 2019 replay corpus

数据集	说话人数量	语音数量		
		真实语音	回放语音	总计
训练集	20	5 400	48 600	54 000
开发集	20	5 400	24 300	29 700
验证集	48	18 090	116 640	134 730
总计	88	28 890	189 540	218 430

表4 ASVspoof 2019的TTS与转换数据集详细信息

Table 4 Detailed information of ASVspoof 2019 TTS and VC corpus

数据集	说话人数量	语音数量		
		真实语音	回放语音	总计
训练集	20	2 580	22 800	25 380
开发集	20	2 580	22 296	24 876
验证集	48	7 355	63 882	71 237
总计	88	12 515	108 978	121 493

2.2 语音欺骗检测中的声学特征

与一般的说话人验证和语音处理所使用的声学特征不同，语音欺骗检测需要开发专门的用于语音欺骗检测的声学特征。这是由于一般的说话人验证或者其他的语音处理任务所常用的声学特征，例如，梅尔倒谱系数（Mel frequency cepstral coefficient, MFCC）在语音欺骗检测中并不能够较好地区分真实语音和欺骗语音，使得欺骗检测的性能较差。因此，专门针对语音欺骗检测开发新的声学特征就显得尤为重要。针对语音欺骗检测的声学特征需要能够较好地表征出真实语音与欺骗语音的区别，例如在语音回放检测中，来自同一个语音样本的真实语音和其回放语音，其语音内容和说话人的特征非常相似，传统的声学特征则不能显示出其区别，图4给出的是一段语音和其回放语音的功率谱图像，可以看到两者非常相似，难以区分。

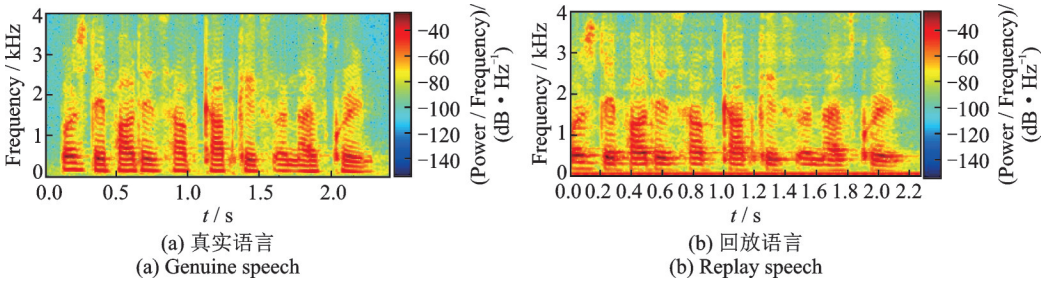


图4 真实语音与回放语音声学特征对比

Fig.4 Comparison of acoustic characteristics between genuine speech and replay speech

从2015年开始，许多国内外的研究者开始研究针对语音欺骗检测的声学特征，本节将重点介绍这些用于欺骗检测的特征。

常数Q倒谱系数（Constant Q cepstral coefficient, CQCC）[33]。该系数是基于常数Q变换（Constant Q transform, CQT）生成的一类倒谱系数。常数Q变换是一种时频分析方法，可以提供可变的时间和频率分辨率。图5阐述了CQCC的提取过程，首先对时域信号$x(n)$进行CQT变换获得CQT频谱

$$X^{CQ}(k,n)=\sum_{j=n-\lfloor N_k/2\rfloor}^{n+\lfloor N_k/2\rfloor}x(j)a_k^*(j-n+N_k/2) \tag{5}$$

式中,$k=1,2,\cdots,K$为频率索引,a_k^*为$a_k(n)$的负共轭,N_k为可变窗长,$\lfloor\cdot\rfloor$表示向下取整。

然后取对数并进行CQT几何尺度的线性化,最后通过离散余弦变换(Discrete cosine transform, DCT)获得倒谱系数,得到CQCC特征。ASVspoof 2017挑战赛中,官方给出的基线系统(https://www.asvspoof.org/data2017/baseline_CM.zip)即是使用CQCC特征和GMM进行语音回放欺骗检测的,并且取得了较好的检测结果。

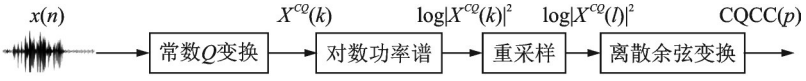


图5 CQCC提取流程

Fig.5 Block diagram of CQCC feature extraction

线性频率倒谱系数(Linear frequency cepstral coefficient, LFCC)。该系数已经被证明在语音欺骗检测中具有良好的性能表现[34]。LFCC首先对信号进行短时傅里叶变换(Short time Fourier transform, STFT)计算幅度谱,随后取对数并使用线性间隔的三角滤波器,最后使用DCT得到LFCC特征。LFCC特征在ASVspoof 2019挑战赛中在官方给出的基线系统中也有出色的性能表现。

基于瞬时频率的耳蜗倒谱系数(Cochlear filter cepstral coefficients instantaneous frequency, CFCC-IF)[35]。该系数在2015年提出,并在检测TTS和转换方面取得了较好的效果。CFCC-IF将耳蜗倒谱系数(Cochlear filter cepstral coefficients, CFCC)与瞬时频率(Instantaneous frequency, IF)相结合,CFCC基于小波变换以及人耳耳蜗的某些机制,如神经尖峰密度。为了计算具有瞬时频率的CFCC,将神经尖峰密度包络乘以瞬时频率,再进行微分和对数运算,最后进行离散余弦变换得到CFCC-IF特征。

群延迟图特征(Group delay gram, GD-gram)[36]。该特征已经被应用于欺骗检测中并且取得了不错的效果。在语音回放中,语音信号的时频表示必须要有较高的分辨率才能从特定的频谱区域中更好地提取出真实语音和欺骗语音的区别信息。与幅度谱相比,群延迟具有更高的谱分辨率,更为重要的是,GD-gram同时包含功率谱和相位谱信息,能够使真实语音与欺骗语音的区别体现得更加明显。

单频滤波倒谱系数(Single frequency filtering cepstral coefficient, SFFCC)[37]。该系数是从最近提出的单频滤波(Single frequency filtering, SFF)方法中提取出的新型语音特征,SFF的主要目的是计算信号的幅度包络随时间的变化,并且可以通过改变参数来调整频谱分辨率。该新型特征在语音欺骗检测中表现出了优秀的检测效果,其提取流程如图6所示。

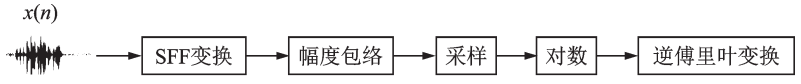


图6 SFFCC提取流程

Fig.6 Block diagram of SFFCC feature extraction

2.3 基于传统机器学习的语音欺骗检测

有了专门针对语音欺骗检测的声学特征后,还需要具有分类性能出色的后端分类模型对提取到的声学特征进行分类和判决,本节主要介绍基于传统机器学习的语音欺骗检测模型。

2.3.1 基于GMM的欺骗检测方法

GMM是一种概率统计模型,其利用期望最大估计算法(Expectation maximization algorithm, EM)

更新参数来训练GMM模型。GMM中含有多个单体高斯模型,通过将多个单体高斯进行线性加权组合,可以拟合许多十分复杂的非线性问题。在说话人验证任务中,通常利用GMM强大的数据拟合能力,来拟合说话人身份模型。而在语音欺骗检测中,则利用GMM分别来拟合真实语音和欺骗语音2个模型。

GMM的概率密度函数为

$$P(\boldsymbol{x}|\boldsymbol{\lambda})=\sum_{i=1}^{C}w_i\,p(\boldsymbol{x}|\mu_i,\boldsymbol{\Sigma}_i) \tag{6}$$

式中,$\boldsymbol{x}$为维度为F的向量,GMM模型为$\boldsymbol{\lambda}=(w_i,\mu_i,\boldsymbol{\Sigma}_i)$,高斯个数为$C$,每个高斯的权重、均值和协方差矩阵分别为$w_i$,$\mu_i$和$\boldsymbol{\Sigma}_i$,$1\leqslant i\leqslant C$。假设一个语音样本的特征矢量矩阵为$\boldsymbol{X}=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_T\}$,则该矩阵相对于GMM的对数似然得分为每个特征矢量$\boldsymbol{x}_i$相对于该模型(真实语音或欺骗语音)的对数似然得分之和。因此,通过对所有特征向量得分取平均,就可以得到最终的似然得分

$$P(\boldsymbol{x}|\boldsymbol{\lambda})=\frac{1}{T}\sum_{t=1}^{T}\log P(\boldsymbol{x}_t|\boldsymbol{\lambda}) \tag{7}$$

式中$P(\boldsymbol{x}_t|\boldsymbol{\lambda})$为特征矢量$\boldsymbol{x}_t$相对于GMM模型的似然得分。图7展示了基于GMM的语音欺骗检测流程。

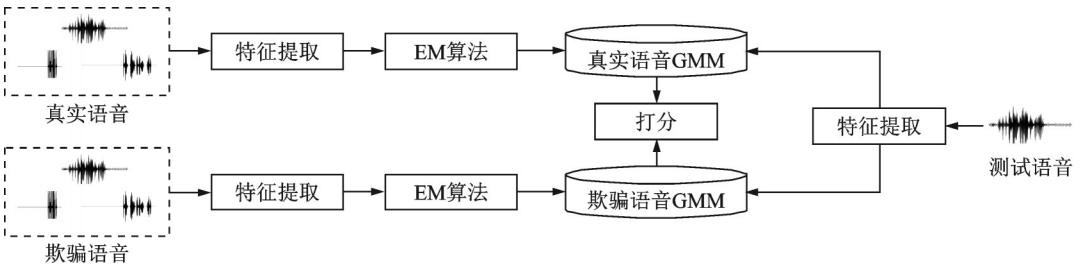


图7 基于GMM的语音欺骗检测流程图

Fig.7 Framework of anti-spoofing system based on GMM

GMM训练速度快、准确度高、使用广泛,在ASVspoof 2015中,基于GMM的欺骗检测系统取得了所有参赛队伍中排名第一的成绩[35]。其良好的表现也导致了在后来的ASVspoof 2017和ASVspoof 2019中均被官方用来作为基线系统供广大参赛者参考,在语音回放和VC欺骗检测任务中均体现了优秀的性能,也成为了众多参赛队伍普遍使用的方法。

2.3.2 基于i-vector的欺骗检测方法

i-vector是一种将不定长的语音转化为定长的可以代表说话人信息的技术(说话人超矢量),是由JFA技术扩展而来,最早由Dehak等[38]于2011年提出,该技术极大地促进了说话人验证领域的发展。总体上来说,i-vector是一种利用全变量子空间建模的技术。该技术基于以下假设:(1)说话人和信道分量具有统计独立特性;(2)这些分量符合高斯分布。i-vector通过训练一个包含说话人和信道信息的全变量子空间矩阵$\boldsymbol{T}$,从而将说话人超矢量经过全变量子空间$\boldsymbol{T}$的投影,降维成只包含说话人信息的低维矢量$\boldsymbol{w}$,即

$$\boldsymbol{M}=\boldsymbol{m}+\boldsymbol{T}\boldsymbol{w} \tag{8}$$

式中,$\boldsymbol{m}$为均值超矢量,和说话人以及信道都独立;$\boldsymbol{T}$为全变量空间矩阵,用来表示跨越大量训练数据的主要方向变换,$\boldsymbol{w}$为全变量因子,也就是i-vector。下面简要介绍i-vector的提取流程。

假设语音的声学特征(如MFCC)的维度为F,GMM的高斯混合数为C,那么i-vector提取过程可以

按照式(9)计算

$$\boldsymbol{w}=(\boldsymbol{I}+\boldsymbol{T}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{N}\boldsymbol{T})^{-1}\boldsymbol{T}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{A} \tag{9}$$

式中,$\boldsymbol{I}$为一个$F\times F$的身份矩阵,$\boldsymbol{N}$为一个$CF\times CF$的对角矩阵,它的对角元素为$N_C\boldsymbol{I}(c=1,2,\cdots,C)$,超矢量$\boldsymbol{A}$是由归一化的一阶Baum-Welch统计量串联起来形成的。$\boldsymbol{\Sigma}$为没有包含在$\boldsymbol{T}$矩阵中的残余变量的协方差矩阵。i-vector中计算全变量空间的过程和JFA特征因子空间的训练过程类似,但是有一个地方不同:在JFA特征因子空间的训练中,通常认为给定说话人的所有语音的归属者为相同的说话人;而在全变量子空间的训练过程中,为了捕捉信道变化,通常认为这些语音属于不同的说话人。i-vector的维度要远远低于说话人超矢量,因此,许多在处理维度较高的超矢量时失效的技术,都可以用来处理i-vector。i-vector在ASVspoof 2015挑战赛中对TTS和转换的欺骗检测取得了第2名的成绩[39],充分证明了i-vector不仅可以用于说话人验证,同样可以用于语音欺骗检测,且能够取得良好的表现效果。

2.3.3 基于SVM的欺骗检测方法

SVM是基于统计学习理论的一种机器学习算法,具有完备的理论、强大的实用性和优秀的泛化能力,是一种优秀的二分类算法,非常适用于语音欺骗检测任务。在欺骗检测任务中,SVM需要区分的两类分别为真实语音的特征和欺骗语音的特征。

对于样本$(x_1,y_1),\cdots,(x_l,y_l),x\in\mathbf{R},y\in\{+1,-1\}$,$l$为样本数,训练SVM就是尽最大能力寻找一个可以将测试数据进行正确分类的函数,可以称其为决策函数,也就是寻找一个可以将2类样本完全隔开的超平面。如果此超平面可以将训练样本准确的隔开,并且可以使每类数据距离超平面的距离最大,则称其为最优超平面,其中,2类样本中距离超平面最近的2个(每类1个,若多于1个则选择1个)到超平面的距离的和称为分类间隔(Margin)。

SVM的训练速度快,分类效果好,在ASVspoof挑战赛中,SVM作为各支参赛队伍广泛使用的后端分类器已经在欺骗检测中表现出了良好的检测效果,在ASVspoof 2015中,使用基于SVM的语音欺骗检测系统取得了第2名的好成绩[39],体现出了优秀的判别效果,将是未来一段时间内仍然受到广泛关注和使用的分类方法。

2.3.4 基于PLDA的欺骗检测方法

概率线性判别分析(Probabilistic linear discriminant analysis, PLDA)打分通常用于i-vector等嵌入式(Embedding)特征后端常用的打分策略。在i-vector中存在一个假设,即说话人信息和信道分量是相互独立且均符合高斯分布。在PLDA中有同样的假设,假设$\boldsymbol{X}_s$和$\boldsymbol{X}_t$为2个语音样本的i-vector,则它们之间的PLDA打分定义为

$$S_{LR}(\boldsymbol{X}_s,\boldsymbol{X}_t)=\frac{P(\boldsymbol{X}_s,\boldsymbol{X}_t|\text{相同类型})}{P(\boldsymbol{X}_s,\boldsymbol{X}_t|\text{不同类型})}=\text{const}+\boldsymbol{X}_s^{\mathrm{T}}\boldsymbol{Q}\boldsymbol{X}_s+\boldsymbol{X}_t^{\mathrm{T}}\boldsymbol{Q}\boldsymbol{X}_t+2\boldsymbol{X}_s^{\mathrm{T}}\boldsymbol{P}\boldsymbol{X}_t \tag{10}$$

式中

$$\boldsymbol{P}=\boldsymbol{\Lambda}^{-1}\boldsymbol{\Gamma}(\boldsymbol{\Lambda}-\boldsymbol{\Gamma}\boldsymbol{\Lambda}^{-1}\boldsymbol{\Gamma})^{-1};\boldsymbol{\Lambda}=\boldsymbol{V}\boldsymbol{V}^{\mathrm{T}}+\boldsymbol{\Sigma} \tag{11}$$

$$\boldsymbol{Q}=\boldsymbol{\Lambda}^{-1}-(\boldsymbol{\Lambda}-\boldsymbol{\Gamma}\boldsymbol{\Lambda}^{-1}\boldsymbol{\Gamma})^{-1};\boldsymbol{\Gamma}=\boldsymbol{V}\boldsymbol{V}^{\mathrm{T}} \tag{12}$$

式中,$\boldsymbol{V}$为因子载荷矩阵,$\boldsymbol{\Sigma}$为PLDA模型的协方差矩阵。

该方法通常和i-vector方法共同使用,作为常用的后端打分系统,具有良好的分类性能,能够较为明显地区别出不同类型的样本,目前已广泛应用于说话人验证,语音欺骗检测等任务中。

2.4 基于深度学习的语音欺骗检测

近年来,随着深度学习(Deep learning, DL)的快速发展,能够区分复杂非线性特征的深度神经网络

层出不穷，极大地提高了对复杂样本的分类准确性，如卷积神经网络(Convolutional neural networks, CNN)[40]、RNN[41]、生成式对抗网络(Generative adversarial network, GAN)[42]和它们的改进方法。现主要介绍用于语音欺骗检测的深度学习方法。

2.4.1 基于DNN的欺骗检测方法

深度神经网络(Deep neural network, DNN)是应用最为广泛的深度学习算法之一，其按照内部结构可以分为输入层、隐含层和输出层，每层之间都是全连接的，具有非常强的非线性问题的拟合性能。在语音处理领域，DNN一般的层数在4层左右。在语音欺骗检测中，首先提取前面所提到的声学特征，再将这些声学特征送入DNN中进行学习和训练，在测试阶段，使用训练好的DNN对待测样本进行分类和判别，具有良好的区分性。在ASVspoof 2015挑战赛中，Yu等[43]提出了一种具有5个隐含层的DNN来进行欺骗检测，并且采用了一种新型的评分方法——人类对数似然值(Human log-likelihoods, HLLs)对检测结果进行评价。网络使用CQCC作为输入，网络中的每个隐含层具有2 048个节点，激活函数采用sigmoid函数，使用softmax层作为网络的输出层，在比赛中取得优异的成绩，证明了DNN在语音欺骗检测中的优秀效果。

2.4.2 基于CNN的欺骗检测方法

CNN是目前深度学习技术领域中非常具有代表性的神经网络之一，在图像分析和处理领域取得了众多突破性的进展，在学术界常用的标准图像标注集ImageNet上，基于CNN取得了很多成就，包括图像特征提取分类、场景识别等。CNN通常被用来从统一大小的样本数据(如图像)中提取鲁棒性的特征，因此需要对数据进行预处理，对时频数据使用固定窗长的窗口化处理从而使数据具有相同的格式。

在ASVspoof 2017挑战赛中，取得语音回放检测第1名的团队使用的就是CNN的变种方法Light CNN (LCNN)[44]，该方法基于最大特征激活(Max-feature-map activation, MFM)，基于MFM的神经网络能够选择对任务求解至关重要的特征，因此可以成功实现音频分类任务，尤其是语音欺骗检测。MFM定义为

$$y_{ij}^{k}=\max\left(x_{ij}^{k},x_{ij}^{k+\frac{N}{2}}\right) \tag{13}$$

$$\forall i=\overline{1,H};j=\overline{1,W};k=\overline{1,N/2} \tag{14}$$

式中，$\boldsymbol{x}$表示$H\times W\times N$的输入矢量，$\boldsymbol{y}$表示$H\times W\times 2$的输出矢量，i和j表示频域和时域，k表示信道索引。图8展示了卷积层的MFM，MFM的使用能够减少CNN架构，因此这也是称为Light CNN的原因。

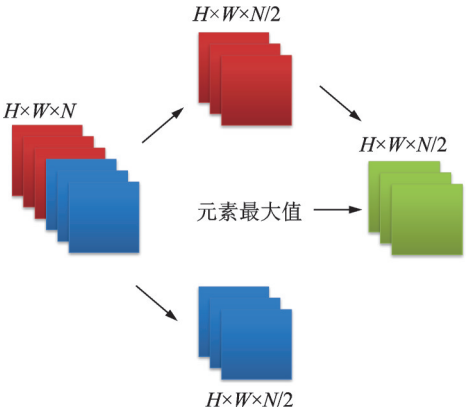


图8 卷积层的MFM

Fig.8 MFM for convolutional layer

Lavrentyeva等[44]对语音提取CQCC特征，后端分类器使用了5个卷积层、4个网络内网络(Network-in-network, NIN)、10个MFM、4个最大池化层和2个全连接层的网络结构进行语音回放检测，此方法证明了CNN在语音欺骗检测中的超强能力，得到了广泛的认可。

2.4.3 基于RNN的欺骗检测方法

RNN的研究始于20世纪八九十年代，并在21世纪初发展为深度学习算法之一，其中双向RNN (Bidirectional RNN, Bi-RNN)和长短期记忆网络(Long short-term memory networks, LSTM)是常见的RNN。RNN通过循环单元和门限结构而具有记忆性，因此在对时间序列问题的处理中具有一定的优势。目前，RNN已经广泛应用于自然语言处理、语音识别、机器翻译等领域。LSTM是最早被提出的

RNN门控算法,其对应的循环单元包括输入门、遗忘门和输出门,相对于RNN对系统状态建立的递归计算。3个门控对LSTM单元的内部状态建立了自循环,即输入门决定当前时间步的输入和前一个时间步的系统状态对内部状态的更新,遗忘门决定前一个时间步内部状态对当前时间步内部状态的更新,输出门决定内部状态对系统状态的更新。LSTM的更新方式如下

$$h^{(t)}=g_{\mathrm{o}}^{(t)}f_h(s^{(t)}) \tag{15}$$

$$s^{(t)}=g_{\mathrm{f}}^{(t)}s^{(t-1)}+g_{\mathrm{i}}^{(t)}f_s(wh^{(t-1)}+uX^{(t)}+b) \tag{16}$$

$$g_{\mathrm{i}}^{(t)}=\mathrm{sigmoid}(w_{\mathrm{i}}h^{(t-1)}+u_{\mathrm{i}}X^{(t)}+b_{\mathrm{i}}) \tag{17}$$

$$g_{\mathrm{f}}^{(t)}=\mathrm{sigmoid}(w_{\mathrm{f}}h^{(t-1)}+u_{\mathrm{f}}X^{(t)}+b_{\mathrm{f}}) \tag{18}$$

$$g_{\mathrm{o}}^{(t)}=\mathrm{sigmoid}(w_{\mathrm{o}}h^{(t-1)}+u_{\mathrm{o}}X^{(t)}+b_{\mathrm{o}}) \tag{19}$$

式中,$s^{(t)}$表示输出状态单元,$h^{(t)}$表示隐藏状态单元,g_{i}表示输入门,g_{f}表示遗忘门,g_{o}表示输出门,f代表激活函数,t表示当前时间节点,b表示偏置,u代表输入层到隐含层的权重,w表示隐藏层节点到下一隐藏层节点的权重。

基于RNN的语音欺骗检测方法在ASVspoof 2017挑战赛中取得了第1名的成绩[5]。此外,Li等[45]使用了基于注意力机制的LSTM结构,对提取出的CQCC进行判别,在ASVspoof 2017语音回放检测数据及上也取得了良好的结果,证明了基于RNN方法在欺骗检测中的适用性。

2.4.4 基于深度特征的欺骗检测方法

i-vector在说话人验证领域取得了非常好的效果。然而,和任何基于统计理论的机器学习模型一样,i-vector系统由若干个独立的无监督子系统组成,这些子系统的训练目标均不相同。在有大量数据作为训练数据的前提下,i-vector系统的性能提升相对有限,为了能够使系统的各个部分联合优化,且能够在大数据训练量的情况下获得更优异的效果,研究者们基于深度神经网络提出了x-vector框架。x-vector是一个基于深度学习的有监督的识别系统,该系统将聚类和提取统计量的步骤合而为一,通过训练时延神经网络(Time delay neural network, TDNN)[46]来区分不同的类别,如图9所示。在统计池化层后的输出层就可以用来当做该语音样本的嵌入式矢量,即x-vector。同时可以设计和使用不同种类的损失函数来满足不同的目标任务,例如使用了AMSoftmax[47]的TDNN通过最大化类间距离和最小化类内距离,进一步提高了x-vector的性能。从TDNN网络中提取到嵌入式矢量后(即图8中的l_6),使用线性判别分析(Linear discriminative analysis, LDA)对嵌入矢量进行降维,得到x-vector。通过将提取到的x-vector表示为函数$g(\cdot)$,真实语音和伪装语音的x-vector可表示为$g(x)$和$g(y)$。因此,优化语音伪装检测任务就可以转化为计算距离$d(g(x),g(y))$,其中d是距离度量,例如余弦距离。Li等[48]使用了x-vector用于语音欺骗检测,在ASVspoof 2019语音回放检测中取得了良好的结果。此外,基于其他深度特征的语音欺骗检

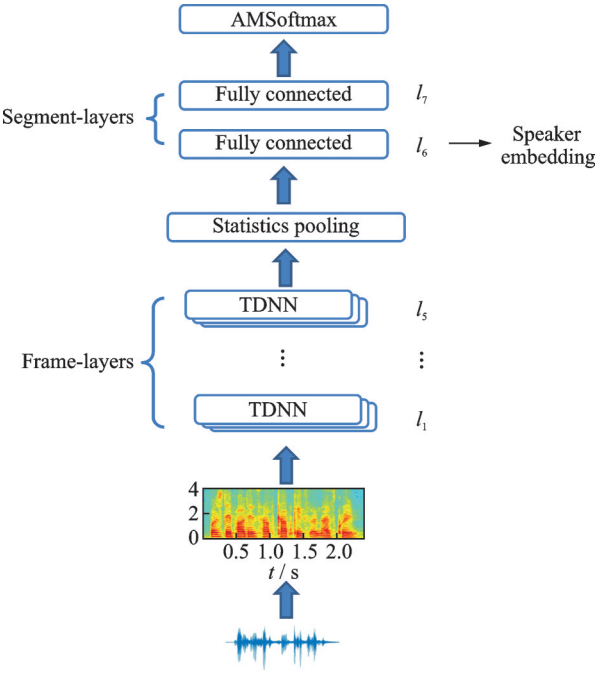


图9 x-vector网络结构图

Fig.9 Framework of x-vector

测方法也在语音欺骗检测中取得了良好的性能[49]。这种基于深度特征的欺骗检测方法虽然准确率高，但是其网络结构较为复杂，所需要的计算时间较长，需要有良好的硬件条件作为支撑。

2.5 语音欺骗检测方法比较

以上总结了语音欺骗检测的总体流程、欺骗检测所使用的声学特征以及后端分类算法，表5将ASVspoof 2015、2017两次语音欺骗检测挑战赛中前5名所使用的方法进行总结，以提供更直观的认识和对比，也为将来的研究提供相应的参考。

表5 ASVspoof挑战赛方法比较

Table 5 Summary of ASVspoof challenge

挑战赛	特征	分类器	EER/%
ASVspoof 2015	MFCC, CFCCIF	GMM	1.211
	MFCC, MFPC, CosPhasePCs	SVM+i-vector	1.965
	DNN-based feature	s-vector	2.528
	LMS, MGD, IF, PSP	Multilayer perceptron	2.617
	PS-MFCC, MGD, WLP-GDCCs, MFCC-CNPCCs	GMM	2.694
ASVspoof 2017	LPCC	CNN, GMM, TV, RNN	6.73
	CQCC, MFCC, PLP	GMM-UBM, TV-PLDA, SVM, GBDT, RF	12.34
	MFCC, IMFCC, RFCC, LFCC, CQCC, SSFC	GMM, FF-ANN	14.06
	RFCC, MFCC, IMFCC, LFCC, SSFC, SCMC	GMM	14.66
	Linear filterbank feature	GMM, CT-DNN	15.97

3 总结与展望

本文介绍了不同的语音欺骗方法以及相应的检测策略，梳理总结了近年来国内外的专家学者在欺骗检测方面所取得进展。过去的几年里，随着各种针对语音欺骗检测数据库的发布，语音欺骗检测方法研究取得了很大的进步。当前最先进的语音欺骗检测技术已经可以取得很高的准确性，具有较高的实用价值，但仍需要以下几个方面进行进一步的研究。

(1) 欺骗检测方法的鲁棒性

最近的研究表明，尽管目前欺骗检测方法在干净环境下的检测效果比较理想，但是在噪声、混响和信道效应的作用下，各种欺骗检测方法基本上就失去了作用。这是由于环境的变化，导致欺骗语音与真实语音的差异变得更加不明显。因此需要进一步研究在复杂的声学环境条件下语音欺骗检测方法的有效性，找到在噪声环境下的检测方法，更加贴近真实使用场景。

(2) 欺骗检测方法的普适性

目前的欺骗检测方法都是针对某种特定类型的欺骗方法，如针对VC的欺骗检测方法在TTS上就表现出较差的性能。此外，针对未知类型的欺骗方法，现有的欺骗检测方法也不能较好区分真实语音和欺骗语音。因此，应该进一步研究更加具有通用性和普适性的欺骗检测方法，使其能够同时应对和检测出多种的欺骗类型，这将是未来语音欺骗检测的重点发展方向。

(3) 欺骗检测和说话人验证联合检测

开发欺骗检测方法的最终目的是保护ASV系统免于受到欺骗,免遭具有欺骗语音的非法者的影响。到目前为止,绝大多数的欺骗检测方法都是独立于ASV的系统。但是将欺骗检测和ASV结合起来并不是一个简单的问题。首先,欺骗检测的判别得分和说话人验证的得分是两种完全不同的计算方法;其次,没有达到很高判别准确率的欺骗检测系统可能会拒绝真实的说话人而使ASV的FAR大大提高;最后,从本质上来看,欺骗检测的改进是否能够改善整个ASV系统目前还并没有一个准确的结论,如果欺骗检测和ASV没有经过适当的匹配,可能无法在实际情况下保护ASV系统。最新的研究工作初步探索了用于联合评估欺骗检测和ASV系统的损失函数以及新型的融和方法,具有一定的借鉴参考价值,为今后的联合检测系统提供了思路。

语音的欺骗检测研究是当前的研究热点,在语音处理和生物识别领域得到了广泛的关注。随着录音设备质量的提高和TTS、VC等语音处理技术的发展,真实的人类语音与人工加工后的语音将越来越难以区分,给语音欺骗检测和ASV系统的安全性带来越来越严峻的挑战。随着越来越多国内外研究者的高度重视和积极参与,有理由相信语音欺骗检测技术将会得到越来越快的发展和进步。

参考文献:

[1] WU Z, EVANS N, KINNUNEN T, et al. Spoofing and countermeasures for speaker verification: A survey[J]. Speech Communication, 2015, 66: 130-153.

[2] EVANS N, KINNUNEN T, YAMAGISHI J. Spoofing and countermeasures for automatic speaker verification[C]// Proceedings of Conference of the International Speech Communication Association (INETSPEECH). Lyon, France: [s.n.], 2013: 25-29.

[3] WU Z, KINNUNEN T, EVANS N, et al. ASVspoof 2015: Automatic speaker verification spoofing and countermeasures challenge evaluation plan[EB/OL]. (2015-01-15) [2020-03-15]. http://www.spoofingchallenge.org/asvSpoof.pdf.

[4] WU Z, KINNUNEN T, EVANS N, et al. ASVspoof 2015: The first automatic speaker verification spoofing and countermeasures challenge[J]. IEEE Journal on Selected Topics in Signal Process, 2017, 11(4): 588-604.

[5] KINNUNEN T, SAHIDULLAH M, DELGADO H, et al. The ASVspoof 2017 challenge: Assessing the limits of replay spoofing attack detection[C]//Proceedings of Conference of the International Speech Communication Association (INETSPEECH). Stockholm, Sweden: [s.n.], 2017: 20-24.

[6] TODISCO M, WANG X, VESTMAN V, et al. ASVspoof 2019: Future horizons in spoofed and fake audio detection[EB/OL]. (2019-04-09) [2020-03-20]. https://arxiv.org/abs/1904.05441.

[7] FARRUS C M, WAGNER M, ERRO D, et al. Automatic speaker recognition as a measurement of voice imitation and conversion[J]. The International Journal of Speech Language and the Law, 2010, 1(17): 119-142.

[8] LAU Y W, TRAN D, WAGNER M. Testing voice mimicry with the YOHO speaker verification corpus[C]//Proceedings of International Conference on Knowledge-Based and Intelligent Information and Engineering Systems. Melbourne, Australia: [s.n.], 2005: 14-16.

[9] MARIÉTHOZ J. BENGIO S. Can a professional imitator fool a GMM-based speaker verification system?[EB/OL]. (2005-03-02) [2020-03-20]. https://core.ac.uk/display/23274186.

[10] HAUTAMAKI R G, KINNUNEN T, HAUTAMAKI V, et al. Automatic versus human speaker verification: The case of voice mimicry[J] Speech Communication, 2015, 72: 13-31.

[11] DELGADO H, TODISCO M, SAHIDULLAH M, et al. ASVspoof 2017 Version 2.0: Metadata analysis and baseline enhancements[C]//Proceedings of Odyssey: the Speaker and Language Recognition Workshop. Les Sables D′Olonne, France: [s.n.], 2018.

[12] LINDBERG J, BLOMBERG M. Vulnerability in speaker verification—A study of technical impostor techniques[C]// Proceedings of The European Conference on Speech Communication and Technology. [S.l.]: [s.n.], 1999.

[13] VILLALBA J, LLEIDA E. Speaker verification performance degradation against spoofing and tampering attacks [C]// Proceedings of FALA 10 workshop. [S.l]: [s.n.], 2010.

[14] WANG Z F, WEI G, HE Q H. Channel pattern noise based playback attack detection algorithm for speaker recognition[C]// Proceedings of 2011 International Conference on Machine Learning and Cybernetics. [S.l.]: [s.n.], 2011: 1708-1713,.

[15] ERGUNAY S K, KHOURY E, LAZARIDISS A, et al. On the vulnerability of speaker verification to realistic voice spoofing [C]//Proceedings of IEEE International Conference on Biometrics: Theory, Applications and Systems. [S.l.]: IEEE, 2015: 1-8.

[16] TAYLOR P. Text-to-speech synthesis[M]. 1st ed. Cambridge: Cambridge University Press, 2009.

[17] YOSHIMURA T, TOKUDA K, MASUKO T T, et al. Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis[C]// Proceedings of Eurospeech. [S.l.]: [s.n.], 1999: 2347-2350.

[18] ZEN H, TODA T, NAKAMURA M, et al. Details of the Nitech HMM-based speech synthesis system for the blizzard challenge 2005[J]. IEICE Transactions on Information Systems, 2007, E90-D(1): 325-333.

[19] WOODLAND P C. Speaker adaptation for continuous density HMMs: A review [C]//Proceedings of Proc ISCA Workshop on Adaptation Methods for Speech Recognition. [S.l.]: [s.n.], 2001: 119.

[20] LING Z H, DENG L, YU D. Modeling spectral envelopes using restricted Boltzmann machines and deep belief networks for statistical parametric speech synthesis[J]. IEEE Transactions on Audio, Speech, and Language Processing, 2013, 21(10): 2129-2139.

[21] WU Z, KING S. Investigating gated recurrent networks for speech synthesis[C]//Proceedings of ICASSP. [S.l.]: IEEE, 2016: 5140-5144.

[22] WANG X, TAKAKI S, YAMAGISHI J. Investigating very deep highway networks for parametric speech synthesis[J]. Speech Communication, 2018, 96: 1-9.

[23] SAITO Y, TAKAMICHI S, SARUWATARI H. Statistical parametric speech synthesis incorporating generative adversarial networks[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2018, 26(1): 84-96.

[24] SHEN J, SCHUSTER M, JAITLY N, et al. Natural TTS synthesis by conditioning wavenet on mel spectrogram predictions [C]//Proceedings of ICASSP. [S.l.]: IEEE, 2018.

[25] MASUKO T, TOKUDA K, KOBAYASHI T, et al. Voice characteristics conversion for HMM-based speech synthesis system[C]//Proceedings of ICASSP. [S.l.]: IEEE, 1997.

[26] DE LEON P L, PUCHER M, YAMAGISHI J, I. et al. Evaluation of speaker verification security and detection of HMM-based synthetic speech[J]. IEEE Transactions on Audio, Speech, and Language Processing, 2012, 8: 2280-2290.

[27] 张雄伟, 苗晓孔, 曾歆, 等. 语音转换技术研究现状及展望[J]. 数据采集与处理, 2019, 34(5): 753-769.
ZHANG Xiongwei, MIAO Xiaokong, ZENG Xin, et al. Voice conversion: The state of the art and prospects[J]. Journal of Data Acquisition and Processing, 2019, 34(5): 753-769.

[28] PERROT P, AVERSANO G, BLOUET R, et al. Voice forgery using ALISP: Indexation in a client memory[C]// Proceedings of ICASSP. [S.l.]: IEEE, 2005: 17-20.

[29] KINNUNEN T, WU Z, LEE K A, et al. Vulnerability of speaker verification systems against voice conversion spoofing attacks: The case of telephone speech[C]//Proceedings of ICASSP. [S.l.]: IEEE, 2012: 4401-4404.

[30] GRIFFIN D, LIM J. Signal estimation from modified short-time Fourier transform[J]. IEEE Transactions on Acoustics, Speech, and Signal Processing, 1984, 32(2): 236-243.

[31] TANAKA K, KANEKO T, HOJO N, et al. Synthetic-to-natural speech waveform conversion using cycle-consistent adversarial networks[C]//Proceedings of SLT. [S.l.]: IEEE, 2018: 632-639.

[32] OORD A, DIELEMAN S, ZEN H, et al. Wavenet: A generative model for raw audio[EB/OL]. (2016-09-12) [2020-03-25]. https://arxiv.org/abs/1609.03499.

[33] TODISCO M, DELGADO H, EVANS N. Constant *Q* cepstral coefficients: a spoofing countermeasures for automatic speaker verification[J]. Computer, Speech and Language, 2017, 45: 516-535.

[34] SAHIDULLAH M, KINNUNEN T, HANILCI C. A comparison of features for synthetic speech detection[C]// Proceedings of Conference of the International Speech Communication Association (INETSPEECH). Dresden, Germany: [s.n.], 2015.

[35] PATEL T B, PATIL H A. Combining evidences from mel cepstral, cochlear filter cepstral and instantaneous frequency features for detection of natural vs. spoofed speech[C]// Proceedings of Conference of the International Speech Communication Association (INETSPEECH). Dresden, Germany: [s.n.], 2015.

[36] TOM F, JAIN M, DEY P. End-to-end audio replay attack detection using deep convolutional networks with attention[C]// Conference of the International Speech Communication Association (INETSPEECH). Hyderabad, India: [s.n.], 2018.

[37] ALLURI K, ACHANTA S, KADIRI S, et al. Detection of replay attacks using single frequency filtering cepstral coefficients [C]// Proceedings of Conference of the International Speech Communication Association (INETSPEECH). Stockholm, Sweden: [s.n.], 2017.

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

[39] NOVOSELOV S, KOZLOV A, LAVRENTYEVA G, et al. STC antispoofing systems for the ASVspoof 2015 challenge [C]//Proceedings of ICASSP. [S.l.]: IEEE, 2016: 5475-5479.

[40] LECUN Y, BENGIO Y. Convolutional networks for images, speech, and time series[M]// The Handbook of Brain Theory and Neural Networks. Boston, USA: MIT Press, 1995, 10: 3361.

[41] GOODFELLOW I J, BENGIO Y, COURVILLE A. Deep learning[M]. Cambridge: MIT Press, 2016: 367-415.

[42] GOODFELLOW I J, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial networks[J]. Advances in Neural Information Processing Systems, 2014, 3: 2672-2680.

[43] YU H, TAN Z H, MA Z, et al. Spoofing detection in automatic speaker verification systems using DNN classifiers and dynamic acoustic features[J]. IEEE Transactions on Neural Networks and Learning Systems, 2018, 99: 1-12.

[44] LAVRENTYEVA G, NOVOSELOV S, MALYKH E, et al. Audio replay attack detection with deep learning frameworks [C]//Proceedings of Conference of the International Speech Communication Association (INETSPEECH). Stockholm, Sweden: [s.n.], 2017.

[45] LI J, ZHANG X, SUM M, et al. Attention-based LSTM algorithm for audio replay detection in noisy environments[J]. Applied Sciences, 2019, 9(8): 1539.

[46] PEDDINTI V, POVEY D, KHUDANPUR S. A time delay neural network architecture for efficient modeling of long temporal contexts[C]//Proceedings of Conference of the International Speech Communication Association (INETSPEECH). Dresden, Germany: [s.n.], 2015.

[47] WANG F, LIE W, LIE H, et al. Additive margin softmax for face verification[J]. IEEE Signal Processing Letters, 2018, 25 (7): 926-930.

[48] LI J, SUM M, ZHANG X, et al. Joint decision of antispoofing and automatic speaker verification by multi-task learning with contrastive loss[J]. IEEE Access, 2020, 8: 7907-7915.

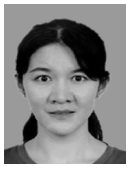
[49] CHEN N, QIAN Y, DINKEL H, et al. Robust deep feature for spoofing detection—The SJTU system for ASVspoof 2015 challenge[C]//Proceedings of Conference of the International Speech Communication Association (INETSPEECH). Dresden, Germany: [s.n.], 2015.

作者简介:

张雄伟(1965-),男,教授,研究方向:语音与图像处理、智能信息处理,E-mail: xwzhang9898@163.com。

李嘉康(1993-),**通信作者**,男,博士研究生,研究方向:语音处理与网络安全,E-mail:jkangli@163.com。

孙蒙(1984-),男,副教授,研究方向:智能语音处理、机器学习。



郑琳琳(1992-),女,硕士研究生,研究方向:语音处理与网络安全。

(编辑:张彤)