ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 36,No. 5,Sep. 2021,pp. 831－849
DOI:10. 16337/j. 1004-9037. 2021. 05. 001

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

# 说话人验证系统攻击方法的研究现状及展望

张雄伟，张星昱，孙　蒙，邹　霞

（陆军工程大学指挥控制工程学院，南京 210007）

**摘　要：**自动说话人验证（Automatic speaker verification，ASV）技术的发展正在深刻地影响和改变着当前的人机交互系统，ASV作为一些智能设备的语音核心功能，可以接受目标说话人的语音并准确识别出该说话人的身份。近年来，人工智能技术的快速进展推动了ASV系统实现跨越式发展。然而，随着人工神经网络和深度学习技术的发展，越来越多的研究者开始研究如何攻击ASV系统。如何通过对原始语音进行一系列处理实现对ASV系统的攻击，是近年来语音领域研究的一个热点问题。目前，对ASV系统的攻击方法大致可分为欺骗攻击（Spoofing attack）和对抗攻击（Adversarial attack）两大类。本文对两大类的典型方法和基本原理进行综述，梳理了目前一些攻击手段中存在的若干问题，揭示了ASV系统存在的安全隐患，对今后ASV系统安全性的发展做了简要的展望，并为未来进一步提高ASV系统的安全性和可靠性提供了参考。

**关键词：**说话人识别；自动说话人验证；欺骗攻击；对抗攻击；深度学习

**中图分类号：**TN912　　**文献标志码：**A

## Attack Methods in Speaker Verification System：The State of the Art and Prospects

ZHANG Xiongwei，ZHANG Xingyu，SUN Meng，ZOU Xia

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

**Abstract：**The development of automatic speaker verification （ASV） technology is profoundly affecting and changing the current human-computer interaction system. As the core speech function of some smart devices，ASV can accept the voice of the target speakers and accurately identify the speakers' identities. In recent years，the rapid development of artificial intelligence technology has promoted the leapfrog development of ASV systems. However，with the development of artificial neural network and deep learning technology，more and more researchers begin to study the way to attack ASV systems. How to attack ASV systems through a series of processing of raw speech has been a hot topic in speech research in recent years. At present，the attack methods of ASV systems can be roughly divided into spoofing attacks and adversarial attacks. In this paper，the typical methods and basic principles of the two kinds of attacks are summarized，some problems existing in current attack methods are sorted out，the safety problems existing in the system of ASV are revealed，a brief outlook on the future development of ASV system security is given，and the development directions of improving the security and reliability of ASV systems are provided.
**Key words:** speaker recognition; automatic speaker verification (ASV); spoofing attack; adversarial attack; deep learning

---

**基金项目：**国家自然科学基金（62071484）资助项目；江苏省优秀青年基金（BK20180080）资助项目。
**收稿日期：**2021-04-01；**修订日期：**2021-07-07

## 引　言

自动说话人验证(Automatic speaker verification, ASV)技术是一种重要的生物特征识别技术,主要用于门禁控制、司法取证和军事侦察等领域[1-2]。然而,攻击者们利用包括欺骗攻击和对抗攻击在内的各种手段,有可能攻破未受保护的ASV系统。因此,如何检测并防御针对ASV系统的攻击行为,成为学者们研究的热点[3]。这些针对攻击行为的检测与防御已经成为提高ASV系统安全性的重要课题之一。

ASVspoof系列挑战赛[4]是社区驱动的标准化项目,旨在解决语音欺骗攻击及其防御中存在的种种问题。其中,语音欺骗攻击包括语音转换(Voice conversion,VC)、语音合成(Text-to-speech synthesis,TTS)、语音模仿(Impersonation)和语音重放(Replay)[5]。在语音欺骗攻击中,攻击者通常会利用各种算法生成和目标说话人尽可能相似的语音,而不会直接使用被攻击的ASV系统的内部信息。例如,VC和TTS任务中,通常以最大化生成音频的语音质量和感知相似度为目标函数,而不是以攻破ASV系统为直接目标。

除了研究鲁棒的欺骗对抗对策之外,研究ASV系统的脆弱性,从而保护其免受多种类型的攻击同样十分重要。为了找出ASV系统的缺陷,需要进一步审视语音欺骗攻击的局限性。作为一名希望攻破ASV系统的黑客,最理想的攻击方式是能够通过ASV系统内部的功能模块进行攻击[4]。但是这类要求通常来说都难以实现,因为上述攻击需要访问ASV系统内部的各种模块,而ASV系统内部模块一般均拒绝非开发者进行访问。另一种能够攻破ASV系统的方法是利用对抗样本(Adversarial examples)[6]进行语音对抗攻击。对抗样本是利用被攻击系统的先验知识,在原始语音中通过故意添加细微的扰动所形成的语音样本。将这些语音样本输入ASV系统,将会导致ASV系统输出错误的识别结果。在诸如图像处理、自然语言处理等分类任务中,对抗攻击受到了很多的关注[7],但是在语音领域,尤其是ASV领域,关于对抗样本的研究还相对较少。关于对抗样本攻击和防御的研究一方面可以提升ASV系统的安全性,另一方面还可以用生成的对抗样本对训练集进行扩充,从而提升ASV系统的鲁棒性。

图1为欺骗攻击和对抗攻击的示意图。由图1可以看出,欺骗攻击不需要与被攻击的ASV系统进行交互,而对抗攻击需要与被攻击的ASV系统进行交互。语音欺骗攻击主要是通过生成或者获取与目标说话人特征接近的语音,从而使ASV系统在判别时产生错误。当攻击者实施欺骗攻击时,需要通过语音欺骗系统得到欺骗音频样本,然后用欺骗音频样本对ASV系统进行欺骗。由于语音欺骗攻击并未利用ASV系统的先验知识,因此攻击成功率相对并没有那么高。

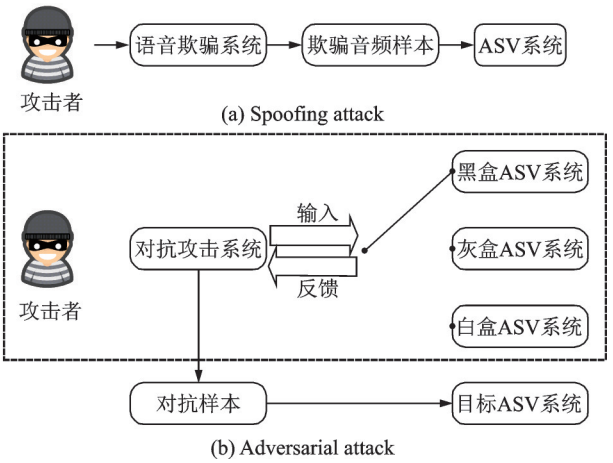


图1　欺骗攻击和对抗攻击

Fig.1　Spoofing attack and adversarial attack

语音对抗攻击是指利用语音对抗样本实现的攻击。在对抗攻击中,攻击者可以利用被攻击的ASV系统(或者其他类似的ASV系统)的先验知识,来生成对抗样本。对抗样本攻击大致可以分为黑盒、灰盒和白盒攻击[8]。在黑盒攻击中,攻击者只能够获取到ASV系统的输出结果(说话人相似性得分、接收/拒绝的结果等),以此作为先验知识来指导对抗样本的生成[9]。灰盒攻击则需要攻击者掌握更多的信

息,例如说话人的特征或者这些特征的实现过程,但是不需要掌握ASV系统的具体模型结构[10]。白盒攻击是最具威胁性的攻击,因为在白盒攻击中,攻击者完全掌握了ASV系统的模型结构,因此具有丰富的先验信息。最近针对对抗攻击的研究表明,对抗样本具有迷惑机器学习系统行为的威胁性[6,11-12]。近些年来,涌现出一些针对ASV系统进行对抗攻击的研究工作[13-17],这些研究工作揭示了ASV系统中存在着新型的潜在安全性威胁。

本文总结和探讨了针对ASV系统的欺骗攻击和对抗攻击方法,并展望了这两类攻击及其应对措施的未来发展方向。

## 1 说话人验证

说话人验证系统的任务是通过测试语音样本和已注册的说话人模型进行比较,来决定接受或者拒绝该说话人[1]。其中,ASV系统又分为文本相关和文本无关两类。文本相关的ASV系统采用固定或者带提示的短语,这些短语通常在说话人测试和验证时保持不变。文本无关的ASV系统则允许说话人用任意语句进行注册和测试。文本相关的ASV系统通常更适用于身份验证场景,因为使用固定的、较短的语句能够实现更高的识别率。文本无关的ASV系统同样具有实用价值,例如电话银行中的说话人验证。典型的ASV系统如图2所示。

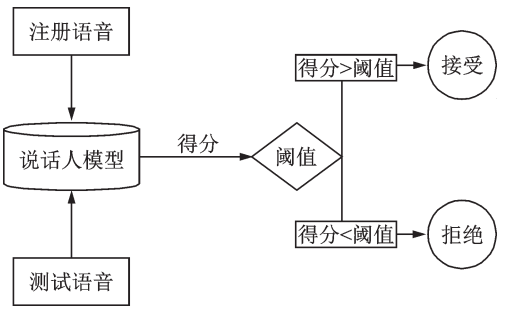


图2 典型的ASV系统

Fig.2 Typical ASV

通常来讲,ASV算法可以分为分级(Stage-wise)ASV算法和端到端(End-to-end)ASV算法。分级ASV算法的前端用于提取说话人特征,后端用于计算特征相似性。前端将时域或时频域表征的语音转化为高维特征矢量。后端首先计算注册说话人特征和测试说话人特征之间的相似性得分,然后将分数与阈值比较

$$f(\boldsymbol{x}^{\mathrm{e}},\boldsymbol{x}^{\mathrm{t}};\boldsymbol{w})\begin{cases}\geqslant\xi & \mathrm{Decision}=H_0\\ <\xi & \mathrm{Decision}=H_1\end{cases} \tag{1}$$

式中:$f(\cdot)$表示计算相似度的函数,$\boldsymbol{w}$表示后端的参数,$\boldsymbol{x}^{\mathrm{e}}$和$\boldsymbol{x}^{\mathrm{t}}$分别表示注册和测试说话人的特征,$\xi$表示阈值,$H_0$表示$\boldsymbol{x}^{\mathrm{e}}$和$\boldsymbol{x}^{\mathrm{t}}$属于同一个说话人,$H_1$表示$\boldsymbol{x}^{\mathrm{e}}$和$\boldsymbol{x}^{\mathrm{t}}$属于不同说话人。后端的主要作用之一是消除信道变异性,降低干扰,例如:减弱语言间的不匹配带来的干扰[18]。相比分级ASV算法,端到端ASV算法直接以一组语音作为输入,直接输出它们之间的相似度。基于Speaker embedding的说话人验证方法是当前的研究重点,接下来对常见的Embedding方法进行概述。

### 1.1 GMM/i-vector和DNN/i-vector

20世纪90年代以后,高斯混合模型(Gaussian mixture model, GMM)以其简单、灵活、有效以及较好的鲁棒性,迅速成为当时文本无关说话人识别领域中的主流技术,将说话人识别研究带入了崭新的阶段。2000年,Reynolds在说话人确认任务中提出了GMM通用背景模型(GMM universal background model, GMM-UBM)结构,为说话人识别从实验室走向实用做出了重要贡献[1]。但是基于GMM-UBM的ASV算法在很大程度上受到说话人本身和信道变化的影响。为解决这一问题,Dehak等[19]提出利用联合因子分析(Joint factor analysis, JFA)方法将GMM-UBM中的超矢量降为低维矢量,并命名为i-vector。GMM/i-vector系统能够有效地消除说话人内部和信道带来的可变性,从而显著地改进ASV

系统的性能。这类能够表征说话人身份的特征矢量又被称作Speaker embedding。GMM/i-vector系统如图3所示。其中常用梅尔倒谱系数(Mel-frequency cepstral coefficient, MFCC)作为表征说话人的声学特征。

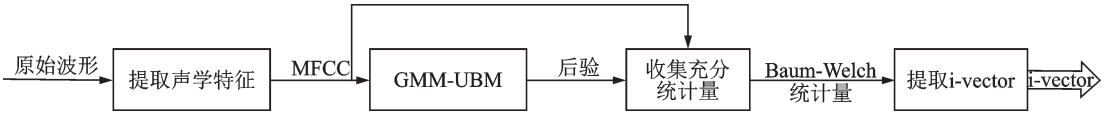


图3 GMM/i-vector系统

Fig.3 GMM/i-vector framework

由于深度学习在语音识别领域的成功应用,有许多研究者做了很多努力,从而将GMM/i-vector中的GMM-UBM模型用深度神经网络(Deep neural network, DNN)来替代,这一类算法有两个分支,分别是DNN-UBM/i-vector和基于DNN的瓶颈特征(Bottleneck feature, BNF)DNN-BNF/i-vector。接下来对这两类算法进行简要介绍。

### 1.2 DNN-UBM/i-vector

由于在收集充分统计量时,只需要用到语音帧的后验概率即可生成i-vector,因此,理论上可以使用除GMM-UBM以外的任意概率模型来计算后验。基于这一观点,Lei等[20]提出了DNN-UBM/i-vector框架,如图4所示。该框架利用了经过自动语音识别(Automatic speech recognition, ASR)系统训练的DNN声学模型,记为DNN-UBM,以此来替代GMM-UBM生成后验概率[20]。其中的后验概率由DNN声学模型产生,充分统计量由ASV系统计算得出。

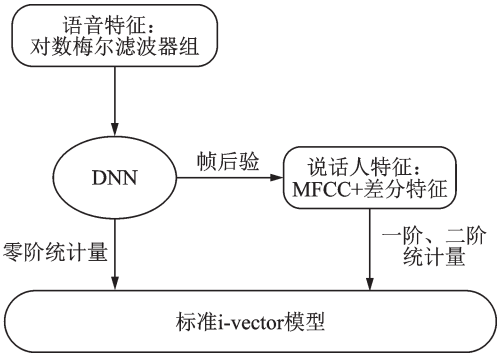


图4 DNN-UBM/i-vector系统

Fig.4 DNN-UBM/i-vector framework

具体来说,DNN-UBM使用一组多元音素(例如三音态)来模拟GMM-UBM中的混合分量。首先训练一个基于DNN的ASR声学模型,使每个训练帧都与每个多元音素对齐,然后从DNN声学模型的softmax输出层生成多元音素上每个帧的后验概率。由于DNN相比GMM有更加强大的表示能力,因此基于DNN-UBM/i-vector架构的ASV系统相比基于GMM/i-vector架构的ASV系统有30%左右的性能提升。

DNN声学模型对内容相关的音素状态的清晰建模能力强,不仅能够生成高度紧凑的数据表示,还能够提供精准的帧间对齐。这种优势在文本相关的ASV任务中尤其明显。然而,相比传统的GMM-UBM/i-vector架构,DNN带来了急剧增加的计算复杂度。此外,基于DNN的声学模型需要大量已标记训练数据进行训练。为了克服这一缺点,Snyder等[21]基于DNN声学模型,提出了有监督的GMM—UBM。尽管有监督的GMM降低了训练中的计算复杂度,但是训练DNN声学模型依然需要大量已标记的训练数据。

### 1.3 DNN-BNF/i-vector

DNN-BNF/i-vector框架的基本思想是,从DNN的瓶颈层提取一个紧凑的特征,将其输入JFA模块。瓶颈层是DNN中的一种特殊隐藏层,它的隐藏单元比其他隐藏层要少得多。

在实际使用中,DNN-BNF/i-vector有多种变体,如图5所示。JFA的输入可以是瓶颈层产生的

BNF,也可以是BNF和其他声学特征的串接[22-23],也可以是经过主成分分析(Principal component analysis,PCA)或线性判别分析(Linear discriminant analysis,LDA)处理过后的特征[22-23]。无论是单独使用BNF[24],还是将其与其他声学特征串接[25],DNN-BNF/i-vector的性能都要明显优于传统的GMM/i-vector。

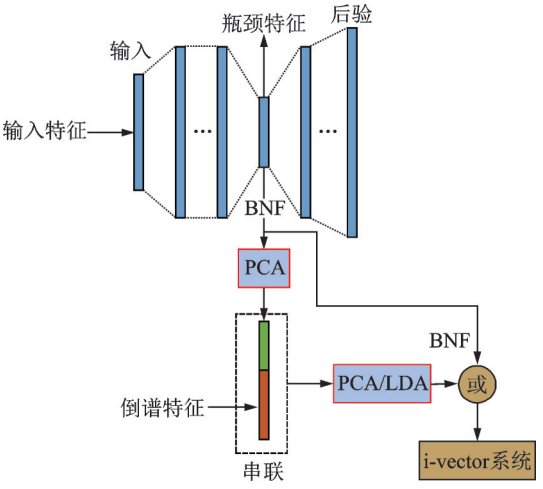


图5 DNN-BNF/i-vector框架

Fig.5 DNN-BNF/i-vector framework

### 1.4 帧级speaker embedding(d-vector)

d-vector是最早的基于DNN的Speaker embedding之一[26]。d-vector的核心思想是,在训练阶段,将一条训练语音对应的真实说话人身份分配给每一帧作为其标签,这种做法可以将模型训练转化为分类问题。如图6所示,d-vector使用上下文信息对每条训练帧进行扩展,并使用带有maxout激活函数的DNN将训练语言中的各帧分类到该语音对应的说话人身份上去。其中DNN使用softmax作为输出层,从而最小化各帧真实标签与网络输出之间的交叉熵损失。

在测试阶段,提取DNN最后隐藏层的输出激活函数作为每一帧的深度身份特征,然后将一条语音中所有帧对应的身份特征进行平均,便得到了该语音紧凑的身份特征表示矢量,命名为d-vector。

### 1.5 段级speaker embedding(x-vector)

x-vector是d-vector的一种重要的变体[27-28],x-vector通过聚合过程将ASV任务从逐帧分析发展到逐句分析。x-vector的网络结构如图7所示。首先,通过时延层提取帧级(Frame-level)的特征矢量。然后通过统计池化层将帧级特征矢量的均值和标准差连接起来生成段级特征(Segment-level)。最后,通过标准的前馈网络将段级特征进行分类。在训练时,时延层、统计池化层和前馈网络进行联合训练。一般取倒数第2个隐层作为表征说话人身份的矢量,称为x-vector。

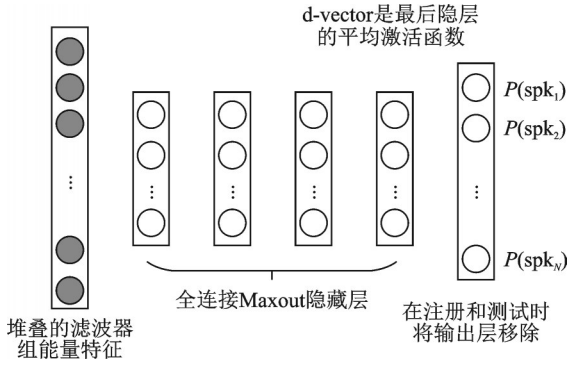


图6 d-vector框架

Fig.6 d-vector framework

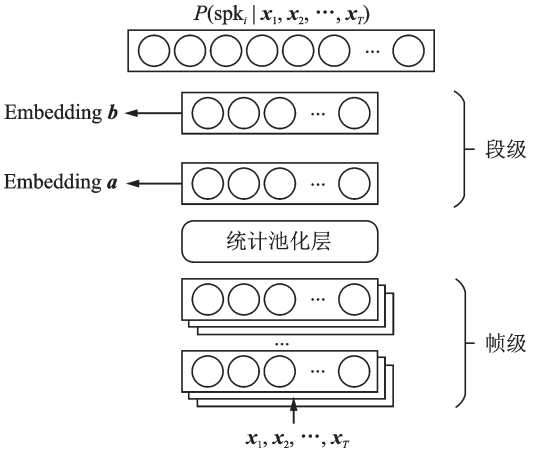


图7 x-vector框架

Fig.7 x-vector framework

## 2 语音欺骗攻击方法

攻击者将语音伪装成目标说话人,以此来获取ASV系统的准入权限,这一过程被称为语音欺骗攻击。每个人的声音信息都很容易被他人获取,因此语音欺骗攻击的发生难以避免。在语音欺骗过程中,攻击者需要利用各种手段获得和目标说话人接近的语音,再将该语音送入ASV系统的麦克风,从而实现欺骗。语音欺骗攻击主要包含4种主要的攻击方式,分别是:语音模仿、语音重放、VC和TTS。下文分别介绍这4种欺骗攻击方式。

### 2.1 语音模仿

语音模仿被定义为产生与目标说话人声音相似的声音模式或者语音行为的过程[29-31]。语音模仿是利用人类改变的声音进行攻击,因此也称为“人类模仿”。语音模仿攻击可以由专业的模仿者执行(利用行为特征),或者由双胞胎执行(利用生理特征)[32]。模仿者通常不需要任何的技术背景或者机器辅助,即可模仿目标说话人。

Lau等[33]的研究表明,如果仿冒者知道目标说话人的声音,并且用类似的声音模式说话,就可以攻破声纹认证系统。通常为了更好地模仿目标说话人,专业的模仿者都会尝试模仿韵律、口音、发音、常用词等高级特征[34]。这种模仿可能会导致人类产生误判,但是对攻击ASV系统的作用不大,因为大部分ASV系统都使用基于频谱的声学特征来生成判决结果。

Hautamäki等[35]分析了GMM-UBM和i-vector系统面对语音模仿攻击时的性能。在这项研究中,5名芬兰的公共人物被选定为目标说话人,之后用专业的模仿者去模仿他们的声音,以此来攻破ASV系统。和Mariéthoz等[36]的研究相似,模仿者依然无法成功攻破ASV系统。Hautamäki等[37]还研究了语音模仿攻击对常见3种ASV系统的影响,实验结果同样证明语音模仿攻击会导致ASV系统产生一定的误判。

在ASV任务中,需要从语音数据中提取每个人独特的说话人特征,然而,双胞胎之间的说话人特征十分相似,区分性不强[38]。通常ASV系统中使用频谱特征来区分说话人身份。但是Kersta等[39]的研究表明,同样的技术手段无法有效区分双胞胎。Patil等[40]的研究也表明,双胞胎之间的语音信号模式,基音周期($F_0$)轮廓,共振峰轮廓和频谱尽管不完全相同,但也非常相似。由于缺少语音特征独特性,因此在面对双胞胎时,ASV系统的错误接受率(False accept rate, FAR)会显著提升。

总之,语音模仿攻击对ASV系统的安全性有一定的影响,但由于这类攻击的效果和具体模仿者的模仿水平有强相关关系,因此一般不被认为是ASV系统的主要威胁之一。但是关于双胞胎模仿攻击的有效性和原理值得进一步进行研究。

### 2.2 语音重放

语音重放是最常见的欺骗攻击手段之一。攻击者希望通过重放预先录制的目标说话人的语音,来达到获取ASV系统准入权限的目的[41-43]。在高质量的音频录音重放设备的加持下,重放语音将会和原始语音高度相似,由于设备的脉冲响应,只有频谱内容会产生微弱的变化。因此,语音重放对ASV系统具有比较严重的威胁。ASVspoof 2017挑战赛将注意力瞄准了针对文本相关ASV系统的语音重放及其检测[44]。ASVspoof2019挑战赛中同样也涉及了对重放攻击的研究,包括模拟场景下的重放攻击和真实场景下的重放攻击[5]。在此之前,针对语音重放的工作比较有限。语音重放主要包括两种,一种是通过录音和重放设备进行攻击,另一种是通过拷贝的语音副本进行攻击,如图8所示。

在语音重放中,真实语音信号记为$s[n]$,该信号可以视为声门气流$p[n]$和声带冲激响应$h[n]$的卷积

$$s[n]=p[n]*h[n] \tag{2}$$

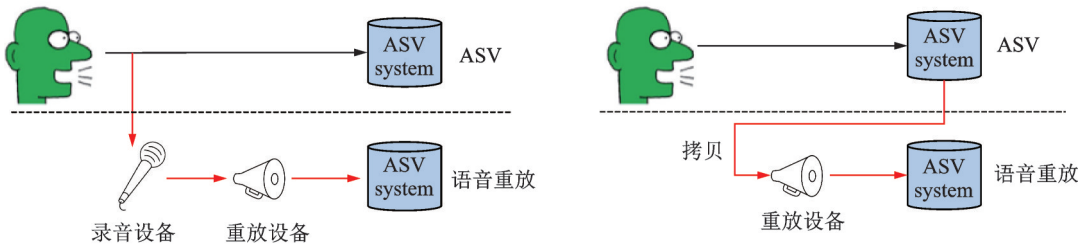


图8 语音重放场景

Fig.8 Replay attack scenario

因此,重放语音信号$r[n]$可以建模为真实语音信号$s[n]$和中间设备(录音和重放设备)的冲激响应的卷积$\eta[n]$

$$r[n]=s[n]*\eta[n] \tag{3}$$

式中,$\eta[n]$中是多种因素的卷积,包括录音设备的冲激响应$h_{\mathrm{mic}}[n]$,录音环境$a[n]$,重放设备(多媒体麦克风)$h_{\mathrm{spk}}[n]$,以及重放环境$b[n]$

$$\eta[n]=h_{\mathrm{mic}}[n]*a[n]*h_{\mathrm{spk}}[n]*b[n] \tag{4}$$

下面介绍一些关于语音重放的研究。Lindberg等[41]首先在隐马尔科夫模型(Hidden Markov model, HMM)模型上研究了针对文本相关系统的语音重放。男性说话人的FAR从1.1%增长到89.5%,女性说话人的FAR从5.6%增长到100%。Shang等[45]对说话人进行了多种录音,并评估了这些录音用来进行语音重放的效果。Wang和Villalba等的团队还研究了基于信道噪声的语音重放。Wang等[46]采用支持向量机(Support vector model, SVM)对信道噪声进行训练,以此来评估输入语音是重放语音还是原始真实语音,实验发现经语音重放后的系统错误率约为40%。Villalba等[47]采用JFA方法的研究基于两种因子:第1种是麦克风中是否有扩音器的重放录音;第2种是语音样本是复制粘贴的可能性。实验结果表明,经过欺骗后的系统等错误率(Equal error rate, EER)为20%,FAR为40%。

Wu等[48]在RSR2015数据集上验证了语音重放对GMM-UBM和HMM-UBM系统的有效性。预先录制好的语音样本对HMM和GMM模型进行重放。EER从2.92%提高到25.56%,FAR则高达78.36%。Galka等[49]研究了电话信道中的语音重放,并提出了一种针对电话信道语音重放的检测手段,并且能够以极高成功率检测出这类语音重放。Delgado等[50]的一项分析表明,在干净环境下通过高质量的录音和重放设备生成的重放语音很难被检测出来。Yoon等[51]提出了一种新型的重放攻击及防御方式,这种攻击只包括嵌入在ASV系统中的一个录音设备的属性,真正的语音只通过录音设备一次,重放语音则要通过同一录音设备两次。针对现实场景中的语音控制系统,Gong等[52]开发了新一代的重放攻击数据集(Realistic replay attack microphone array speech corpus, ReMASC),数据集中包含真实的语音控制指令和重放设备播放的指令,该数据集对语音控制系统中的重放攻击研究提供了公开的研究素材。

总之,ASV系统很容易受到预先录制好的说话人语音样本带来的语音重放攻击。由于语音重放的语音样本中包含了大量目标说话人本身的特征,因此语音重放对任何不受保护的ASV系统都具有严重的威胁,尤其是对文本无关ASV和没有错误口令保护的文本相关ASV。对于拥有错误口令保护的ASV系统来说,由于需要预先录制相同内容的语音样本,因此无法实现较为灵活的语音重放攻击。

### 2.3 语音转换

语音转换(VC)的目标是将源说话人的语音波形进行变换处理,使其听起来像目标说话人的语音[53]。VC是对声调、语音时长、响度和音色等不同语音特征的频谱映射。VC的一般流程分为3步,分别是语音特征提取、语音特征转换和重新合成语音信号。在提取特征阶段,常用的算法包括谐波噪声模型(Harmonic noise model, HNM)[54]、自适应加权谱内插(Speech transformation and representation using adaptive interpolation of weighted spectrum, STRAIGHT)方法[55]等。通常来说,最常使用也是最重要的语音特征是频谱包络特征(或者MFCC),它表示具体发音。除此之外,基频(表示音高)、语速、韵律等特征有时也会用于语音转换。在重新合成语音的过程中,一般使用声码器。传统的声码器合成出的语音质量通常很差,因此,在2018年开展的语音转换挑战赛报告中提出[56],使用WaveNet[57]来替代传统的声码器,可以获得语音质量的提升。在ASVspoof2019中[58],利用了两种语音转换算法来生成欺骗语音,分别是基于神经网络的方法和基于转换函数的方法。

#### 2.3.1 传统语音转换方法

传统的语音转换方法中,需要固定源说话人和目标说话人的身份,同时需要帧间对齐的训练数据。因为有数据对齐的要求,因此传统的语音转换方法一般难以进行跨语种的语音转换,即源说话人和目标说话人的训练数据不能是不同语种。

Abe等[59]提出了基于统计频谱映射的矢量量化(Vector quantization, VQ)方法。Pellom等[60]针对GMM-UBM系统,在包含138个说话人的YOHO数据集上进行了VC攻击实验,实验结果表明,FAR从1%急剧上升到了81%,这说明VC对未经防御的ASV系统存在严重的威胁。Patrick等[61]和Matrouf等[62]利用VC攻击GMM-UBM系统,错误率分别从16%和8%上升到26%和63%。Bonastre等[63]同样对GMM-UBM系统进行语音转换攻击,使系统错误率从6.61%提升到28.07%。Kinnunen等[64]对JFA系统进行攻击后,使错误率从3.24%提升到7.61%。Wu等[65-66]和Alegre等[67-69]则使用VC对不同ASV系统实现了攻击。

**高斯混合模型**(Gaussian mixture model, GMM)[70-72] GMM是最主流的传统方法,这种方法的基本思路为,用一个GMM对输入特征和转换后的特征的联合分布进行拟合,然后在转换时,根据输入特征和得到的GMM即可推断出转换后的特征。语音转换中,GMM的每一个分量表示如下

$$\begin{bmatrix} \boldsymbol{X} \\ \boldsymbol{Y} \end{bmatrix} \sim N\left( \begin{bmatrix} \boldsymbol{\mu}_X \\ \boldsymbol{\mu}_Y \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_{XX} & \boldsymbol{\Sigma}_{XY} \\ \boldsymbol{\Sigma}_{XY} & \boldsymbol{\Sigma}_{YY} \end{bmatrix} \right) \tag{5}$$

式中:$\boldsymbol{X}$和$\boldsymbol{Y}$分别代表输入的特征和转换后的特征;$\boldsymbol{\Sigma}_{XX}$,$\boldsymbol{\Sigma}_{XY}$和$\boldsymbol{\Sigma}_{YY}$均为对角矩阵,即只有$\boldsymbol{X}$和$\boldsymbol{Y}$对应维度之间是相关的,$\boldsymbol{X}$和$\boldsymbol{Y}$内部各维度、以及$\boldsymbol{X}$和$\boldsymbol{Y}$的不同维度之间相互独立。因此在选取特征时,最好选取各维度之间本来就相互独立的特征(如MFCC)。

**频率弯折法**(Frequency warping)[73] 频率弯折法主要包含3步。首先,对训练数据中的输入、输出语音分别提取共振峰信息;然后从匹配的输入、输出的共振峰数据中,拟合出一个分段线性弯折函数[74];在转换时,利用拟合出的弯折函数对语音的频谱包络进行伸缩变换。分段线性的弯折函数可以调整频谱包络中各个共振峰的位置和宽度,从而使输入、输出频谱包络尽可能相似。但是频率弯折法由于对频谱包络改动过少,因此局限较大。不过也正因为改动较少,导致转换后的音质较好。

**基于模板的方法**(Example-based method)[75] 这种方法的一般思路是,将语音的语谱图分解成许多基本元素(即模板)的叠加。如图9

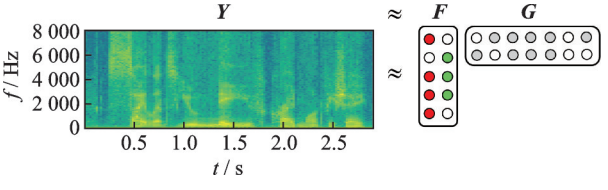


图9 语谱图NMF

Fig.9 Non-negative matrix factorization of spectrogram

所示。$Y$表示语谱图;$F$表示词典,每一列表示一个元素;$G$表示增益矩阵,其元素表示每个样本的强度,一般来说$G$是稀疏的。词典和增益矩阵都是非负矩阵,因此这个过程称为非负矩阵分解(Non-negative matrix factorization, NMF)。基于模板的语音转换步骤主要分为两步:

(1) 在训练时,把对齐后的源说话人和目标说话人的语谱图进行NMF。用$\boldsymbol{X}_i$和$\boldsymbol{Y}_i$表示源说话人和目标说话人的第$i$条语音的语谱图,则有

$$\boldsymbol{X}_i \approx F_1 \cdot \boldsymbol{G}_i \tag{6}$$

$$\boldsymbol{Y}_i \approx F_2 \cdot \boldsymbol{G}_i \tag{7}$$

式中,词典矩阵$F$只和说话人相关,同一说话人全部语音的词典都相同;而增益矩阵$\boldsymbol{G}$与说话人无关。

(2) 在转换时,先将输入语音的语谱图$\boldsymbol{X}$用原说话人的词典$F_1$分解,得到增益矩阵$\boldsymbol{G}$,再根据该增益矩阵和目标说话人词典$F_2$合成转换后的语谱图$\boldsymbol{Y}$

$$\boldsymbol{X} \approx F_1 \cdot \boldsymbol{G} \tag{8}$$

$$\boldsymbol{Y} = F_2 \cdot \boldsymbol{G} \tag{9}$$

2.3.2 *新型语音转换方法*

相比传统的语音转换方法,新型的语音转换方法除了在音质和说话人相似性两方面对传统方法实现了超越,而且新型语音转换不需要训练数据帧间对齐,其中有些方法还突破了源说话人和目标说话人身份固定的限制[76]。

**生成对抗式网络**(Generative adversarial network, GAN) GAN是一种生成式网络,GAN的模型中除了有一个生成器$G$以外,还有一个判别器$D$。GAN应用于图像、语音等数据时,生成的结果往往能以假乱真,但缺点是训练比较困难。Kaneko等[77]利用CycleGAN实现了语音转换。CycleGAN中含有两个生成器$G$和$F$,如图10所示。$G$负责把源说话人语音$x$转换成目标说话人语音$y$,$F$的作用则是将$y$变成$x$。使用CycleGAN进行语音转换,突破了需要平行语音训练数据的限制,但是仍然要提前指定源说话人和目标说话人的身份。

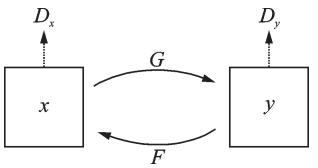


图10 CycleGAN示意图

Fig.10 Schematic diagram of CycleGAN

**i-vector + PLDA** (Probabilistic linear discriminant analysis) 之前介绍的语音转换方法都需要指定源说话人和目标说话人的身份,Kinnunen等[78]借鉴了ASV中的i-vector和PLDA,只需要训练一个系统,就可以处理多个源说话人和目标说话人。首先提取输入语音的i-vector $\boldsymbol{\omega}_1$,假设源和目标说话人的i-vector分别为$\boldsymbol{y}_1$和$\boldsymbol{y}_2$,为了保持语音内容不变,只改变说话人身份,则按照式(10)转换得到转换后的i-vector $\boldsymbol{\omega}_2$。

$$\boldsymbol{\omega}_2 = \boldsymbol{\omega}_1 + \boldsymbol{S}(\boldsymbol{y}_2 - \boldsymbol{y}_1) \tag{10}$$

然后通过i-vector的改变量,逆推出GMM中各分量偏移量,并求出每帧语音特征属于GMM各分量的概率,以这些概率为权重对各分量偏移量进行加权平均,得到每帧语音应该改变的数值。最后用修改后的MFCC重新合成语音。

**自编码器**(Autoencoder) 通过自编码器进行语音转换,也不需要指定源说话人和目标说话人。自编码器中含有一个编码器和一个解码器,编码器负责把数据的表层特征转换成隐特征,解码器负责从隐特征中恢复出表层特征。在语音转换任务中,数据的表层特征可以是波形、语谱图、MFCC序列等,隐特征则蕴含了语音的内容和说话人的身份信息。如果能通过某种手段,将隐特征中内容和身份两部分信息分开,那就可以实现任意替换隐特征中的身份信息,实现语音转换。Hsu等[79]、Chou等[80]、Qian等[81]利用自编码器的这种特性,提出了多种语音转换方法。

使用语音转换对ASV系统进行攻击，通常是通过最小化生成语言和目标语音之间的频谱距离来实现的。虽然频谱距离测度和ASV系统中的说话人相似性度量之间的联系性很弱，但是多项研究表明，ASV系统仍然对这些攻击很敏感[63-64,82-83]。

### 2.4 语音合成

语音合成(TTS)是一种可以将任意文本信息转换为自然语音的技术，如图11所示。TTS系统包含两个主要过程：第一步是文本分析，在这一步文本被转化成语音或者其他形式；第二步是利用上一步生成的信息合成语音信号，其中第一步通常被称为前端，第二步被称为后端。之前有许多关于ASV系统面对TTS时表现出脆弱性的研究[41-42,84]。早在2000年，Masuko等[85]就发现了ASV系统面对基于HMM的TTS技术时具有脆弱性。后来，De Leon等[86]分别对GMM-UBM和SVM两类ASV系统进行了TTS攻击，攻击使FAR分别从0.28%和0%上升到86%和81%。Galou等[87]针对商用的ASV系统进行攻击，也取得了类似的效果。在ASVspoof2019挑战赛中[58]，使用了多种TTS算法，包括波形串接、使用源滤波编码器的参数TTS和使用Wavenet的TTS。这些合成语音均由公开的TTS工具Merlin[88]、CURRENT[89]和MaryTTS[90]生成。

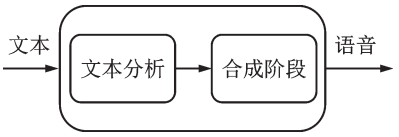

图11 TTS流程图
Fig.11 Flow chart of TTS

#### 2.4.1 结合深度学习的传统TTS

TTS中传统方法主要包括拼接法、参数法。由于深度学习的快速发展，将神经网络引入传统的TTS系统中，用于替代各个模块，是一种有效的方法。代表方法有Deep Voice-1[91]和Deep Voice-2[92]。Deep Voice-1的合成速度较快，合成质量也很高。Deep Voice-2则进一步将i-vector引入了模型训练过程。但是这两种系统是模块化系统，因此在训练时难以进行联合优化。到Deep Voice-3系统已经实现了端到端TTS。

#### 2.4.2 端到端TTS

Wang等[93]于2016年首次提出了端到端TTS模型。后来，在Interspeech2017上发布了Tacotron-1 TTS端到端系统[94]。2018年开发者又对其进行了改进，在Tacotron-1的基础上改用Wavenet作为声码器[95]。总体上讲，相比非端到端TTS系统，Tacotron系列系统架构相对较为简单，同时也能得到高质量的合成语音。百度于2018年在Deep Voice-2的基础上也开发了自己的端到端TTS系统——Deep Voice-3[96]。Deep Voice-3是一个基于全卷积注意力机制的TTS系统，其中的声学模型可以生成多种中间表征形式。Deep Voice-3相比之前的TTS系统，大幅度提升了训练速度和TTS速度。

现代的VC和TTS系统都不是为特定的说话人量身打造的。通过将多说话人语音数据训练出的模型自适应调整到预期目标[83]，或者使用全局说话人变量来调节模型[97]，可以生成高质量的目标说话人语音。这些说话人调节变量和ASV系统中的说话人身份矢量类似。这些技术的发展使得ASV系统和TTS/VC系统更加接近，也因此会对ASV系统产生更大的威胁。

## 3 语音对抗攻击

语音对抗攻击是指利用语音对抗样本对ASV系统进行攻击。对抗样本是指在数据集中通过故意添加细微的扰动所形成的输入样本，该样本会导致机器学习模型以高置信度给出一个错误的输出。包括深度学习模型在内的机器学习模型对于对抗攻击十分敏感，攻击者可以只对原始语音样本进行微弱的改动，即可导致ASV系统无法正常进行识别、分类任务。随着DNN在诸如ASR、说话人识别、情感识别和行为识别等语音信号处理任务中的应用，研究ASV系统针对对抗样本的脆弱性并研究如何防御对抗样本变得越来越重要。下面简要介绍国内外学者针对ASV系统的对抗样本攻击研究。

### 3.1 语音对抗攻击概述

给定音频样本$\boldsymbol{x}$,则语音对抗样本的生成过程表示为

$$\widetilde{\boldsymbol{x}}=\boldsymbol{x}+\boldsymbol{\eta}\quad \text{s.t.}\ \|\delta\|_{\text{p}}\leqslant\varepsilon \tag{11}$$

式(11)的目标是使分类器无法正确完成对$\tilde{x}$的分类。如果原始音频$\boldsymbol{x}$的标签为$y_{\text{ori}}$,则攻击者的目标就是使对抗样本$\widetilde{\boldsymbol{x}}$的分类结果$\tilde{y}\neq y_{\text{ori}}$。在语音领域,式(11)中的范数p通常为无穷范数或者2范数。

目前针对ASV系统进行的对抗攻击,按照攻击者是否掌握ASV内部信息(包括模型结构,参数,损失函数和梯度信息等),可以分为白盒攻击、灰盒攻击和黑盒攻击。通常来说,白盒攻击和灰盒的攻击成功率更高,但是黑盒攻击更加符合现实攻击场景。对抗攻击还可以分为有目标和无目标攻击。在无目标攻击中,只需要使ASV系统产生错误的输出结果即可;在有目标攻击中,需要指定ASV系统输出特定的识别结果。

关于说话人识别系统的对抗样本研究在近几年刚刚开始起步。2018年,Kreuk[98]和Gong等[99]首次提出了针对端到端长短时记忆网络(Long short term memory, LSTM)说话人识别系统的对抗样本生成方法。之后,又出现了针对x-vector和i-vector系统的对抗样本攻击。2020年,Li等[100]使用快速梯度符号法(Fast gradient sign method, FGSM)实现了对GMM/i-vector和x-vector系统的攻击。Xie等[101]提出了一种实时的通用黑盒攻击方法,该方法生成的对抗样本不仅能够实现实时攻击,而且可以适应不同说话人发出的不同时长的语音。在实验中,该通用对抗攻击方法成功攻击了基于Kaldi[102]时延神经网络(Time-delay neural network, TDNN)的x-vector系统。Li等[103]引入了通用对抗扰动(Universal adversarial perturbations, UAPs),文中提出了一种生成模型,该模型能够学习从低维正态分布到UAPs子空间的映射,因此使用任何输入语音均可生成UAPs。实验表明生成的UAPs可以以高成功率欺骗已训练的ASV模型。Wang等[104]利用心理声学概念,对x-vector系统实现了白盒攻击,并且显著减弱了扰动的可感知性。Villalba等[105]研究了利用对抗样本攻击基于x-vector的ASV系统,并且成功利用小型白盒ASV系统中生成的对抗样本攻击了规模更大的黑盒ASV系统。Zhang等[106]提出了动量迭代FGSM(Momentum iterative FGSM, MI-FGSM),并用该方法成功攻击了ASV欺骗对抗系统。Chen等[107]提出了FakeBob对抗攻击系统。FakeBob通过向原始语音添加细微扰动从而实现了黑盒攻击,并且能够在多种现实场景中实现攻击。研究中包含了多种说话人识别系统的架构(包含商用系统),攻击的可转移性,不可感知性分析和在现实场景中进行播放实现攻击。Li等[108]和Xie等[101]的研究通过在训练中添加混响,探究了真实场景下对抗攻击的实时性和可行性。

Marras等[109]尝试使用字典攻击(Dictionary attack)来攻击ASV系统。这类攻击允许有大量的目标说话人,而且不需要了解关于目标说话人的声音特点或者语音模型。字典攻击通过向主声(Master voice)中添加对抗扰动,来最大化主声和大部分说话人的语谱图相似度。当语谱图相似度超过阈值,主声和人群中大量说话人相接近时,就通过语谱图反向生成时域波形。

Nakamura等[110]利用白盒ASV系统实现了一种验证-合成攻击(Verification-to-synthesis, VTS)。在这种对抗攻击中,使用没有目标说话人训练数据的白盒ASV模型对VC系统进行训练。由于训练后的网络可能会对输入语音的语音特性进行扭曲,因此在优化过程中,文中添加了一个ASR模型,以弥补语音信息的损失。这样输出的语音不仅能够欺骗ASV系统,而且保持了感知质量。

Liu等[111]针对ASV系统的抗欺骗系统(Anti-spoof)进行了黑盒和白盒攻击。作者使用FGSM和投影梯度下降法生成对抗样本,以此来攻击基于轻量卷积神经网络(Light convolutional neural network, LCNN)的抗欺骗系统。通过实验表明,使用黑盒和白盒生成对抗样本,均可以有效欺骗性能良好的ASV抗欺骗系统。

### 3.2 对抗样本生成算法

接下来,对已经应用于ASV系统的几种对抗样本生成算法进行简要介绍。

3.2.1 FGSM

FGSM是一种计算效率高的单步攻击方法[11],计算中仅仅使用梯度函数的符号,并沿着梯度的方向来最大化误分类,从而生成对抗样本

$$\widetilde{\boldsymbol{x}} = \boldsymbol{x} + \varepsilon \mathrm{sign}(\nabla_{\boldsymbol{x}} L(\boldsymbol{x}, y, \boldsymbol{\theta})) \tag{12}$$

式中:$\boldsymbol{x}$为给定的源说话人的语音样本,$\widetilde{\boldsymbol{x}}$为最终生成的对抗样本,$y$为待攻击的目标说话人标签,$\boldsymbol{\theta}$为网络结构参数,$L$为交叉熵函数,$\varepsilon$为限定的最大扰动值。通常来说,最大扰动$\varepsilon$越大,ASV系统产生误分类的可能性越高,但是扰动的不可感知性也越差。FGSM是一种能够快速生成对抗样本的算法,但是攻击成功率并不是很高。

3.2.2 迭代FGSM或基本迭代法

相比FGSM在梯度方向上取单步下降,迭代FGSM(Iterative FGSM, IFGSM)或基本迭代法(Basic iterative method, BIM)[112]在梯度方向上进行多次步长较小的迭代,迭代步长为$\alpha$,即

$$\widetilde{\boldsymbol{x}}_{i+1} = \boldsymbol{x} + \mathrm{clip}_{\varepsilon}(\widetilde{\boldsymbol{x}}_i + \alpha \mathrm{sign}(\nabla_{\widetilde{\boldsymbol{x}}_i} L(\widetilde{\boldsymbol{x}}_i, y, \boldsymbol{\theta})) - \boldsymbol{x}) \tag{13}$$

式中:$\widetilde{\boldsymbol{x}}_0 = \boldsymbol{x}$,$i$为优化迭代的轮数。clip函数则保证每次迭代后扰动的最大值均小于$\varepsilon$。IFGSM或BIM的攻击成功率要高于FGSM,但同时也需要消耗更多的时间。

3.2.3 Carlini-Wagner(CW)攻击

CW攻击[113]尝试找到能够欺骗分类器并且保持不可感知性的最小扰动

$$\min \left\| \eta \right\|_{\mathrm{p}} + c \cdot g(\widetilde{\boldsymbol{x}}) \quad \text{s.t. } \widetilde{\boldsymbol{x}} \in [0, 1] \tag{14}$$

式中$g(\cdot)$定义的目标函数表示为

$$g(\widetilde{\boldsymbol{x}}) = \left[ Z(\widetilde{\boldsymbol{x}})_t - \max_{j \neq t} (Z(\widetilde{\boldsymbol{x}})_j) + \delta \right]_+ \tag{15}$$

式中:$Z(\cdot)$为包含所有类别后验概率的输出矢量,$t$表示真实标签对应的输出节点,$\delta$为置信边界参数,$[\cdot]_+$表示$\max(\cdot, 0)$。直观上看,CW攻击尝试找到错误类别中后验概率最大的类别,并使其后验概率超过真实标签的后验概率。范数p可以取2或者$\infty$。

3.2.4 投影梯度下降攻击(Projected gradient descent, PGD)

Madry等[114]提出了迭代梯度$l_\infty$攻击的广义版本

$$\widetilde{\boldsymbol{x}}_{i+1} = \prod_{\boldsymbol{x}+S} [\widetilde{\boldsymbol{x}}_i + \alpha \mathrm{sign}(\nabla_{\boldsymbol{x}} L(\boldsymbol{x}, y, \boldsymbol{\theta}))] \tag{16}$$

式中:$\alpha$为梯度下降更新的步长,$\boldsymbol{x}+S$表示如果扰动超出了一定范围,就要映射回规定的范围$S$内。PGD算法会规定最大的迭代次数$T$。因此经过$T$轮迭代的PGD一般记为PGD-T。

## 4 总结与展望

本文从语音欺骗攻击和对抗样本攻击两个角度,介绍了针对ASV系统的攻击方法,梳理总结了近些年来国内外专家学者对ASV系统安全性研究方面所取得的进展。总体上说,目前关于语音欺骗攻击和检测的研究远多于对抗样本的攻击与防御研究。但是由于对抗样本攻击的攻击成功率更高,不可感知性更强,因此比语音欺骗攻击对ASV系统的威胁更大。当前最先进的语音欺骗攻击和对抗样本攻击都已经取得了很高的攻击成功率,但是仍需要以下几个方面进行进一步的研究。

语音欺骗攻击方面:

(1) 多欺骗攻击手段联合

目前主流的研究都关注于单一欺骗手段的研究,未来攻击者可能结合多种欺骗攻击手段,从而实现攻击。例如语音转换合语音模仿相结合,可以在频谱域和时间域对ASV系统进行欺骗。因此,对于这类多欺骗攻击手段的联合攻击,需要研究者们继续关注。

(2) 欺骗检测方法的普适性

目前的欺骗检测和防御手段大多都只能降低某种特定欺骗攻击的威胁。同时,针对未知类型的欺骗攻击手段,目前的检测方法还不能做到较好地区分欺骗语音和真实语音。未来应该着重研究具有通用性和普适性的欺骗检测方法,从而能够在没有任何先验知识的情况下,检测出未知的欺骗攻击手段。

(3) 欺骗攻击和检测方法的鲁棒性

现实场景中存在大量的噪声和混响,会带来注册和测试之间的不匹配问题,从而影响欺骗攻击或者检测的效果。因此需要进一步降低噪声和混响带来的不利影响,提高复杂的声学环境下欺骗攻击和检测的成功率,使欺骗攻击和检测更加贴近真实应用场景[115]。

语音对抗攻击方面:

(1) 攻击与防御的评判标准

目前,针对ASV系统的对抗样本攻击与防御的研究较少,现有的对抗攻击研究还没有在评估数据集和评估指标上实现统一,大多数工作都利用现有的说话人识别数据集进行实验。因此,考虑到对抗样本攻击与防御的实用性,需要有一个共同的协议、评估指标和评估数据集来统一评判标准。此外,无论语音欺骗攻击和对抗样本攻击,都是对ASV系统严重的威胁,如何将二者统一到同一评判标准仍值得进一步研究。

(2) ASV系统防御手段

针对ASV系统的对抗样本攻击,深刻地揭示了ASV系统的脆弱性。因此,应该进一步研究针对此类攻击的防御方法。如何能够同时应对和防御多种攻击类型,将是防御领域的重点之一。此外,对抗样本攻击可以利用有关于任何系统的先验信息,因此可以将对抗样本攻击施加在带有欺骗对抗策略的ASV系统上。目前尚未有类似工作发表。由于现实场景中的许多ASV系统都结合了欺骗对抗系统,因此,针对此类组合系统的攻击具有很强的现实意义,同时可以进一步提高此类系统的安全性。

(3) 提高不可感知性和攻击成功率

通过生成具有高度不可感知性的对抗样本,从而使人类无法察觉对抗样本的存在,进而以高成功率实现攻击,是目前许多研究者正在努力的方向。同时,如何对这类对抗样本进行检测和防御,也是一个值得研究的课题。

语音的欺骗攻击和对抗样本攻击是当前的研究热点,在语音信号处理领域和信息安全领域均受到了广泛关注。随着录音设备质量的提高,对抗样本攻击技术的发展,以及TTS、VC等语音生成技术的进步,ASV系统面对的安全性威胁将会越来越严重。当前,越来越多的国内外研究者参与到了语音系统的安全性研究之中,相信在众多研究者的努力下,语音系统的安全性将会得到显著提升。

**参考文献:**

[1] KINNUNEN T, LI H. An overview of text-independent speaker recognition: From features to supervectors[J]. Speech Communication, 2010, 52(1): 12-40.

[2] PODDAR A, SAHIDULLAH M, SAHA G. Speaker verification with short utterances: A review of challenges, trends and opportunities[J]. IET Biometrics, 2017, 7(2): 91-101.

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

[4] WU Z, YAMAGISHI J, KINNUNEN T, et al. ASVspoof: The automatic speaker verification spoofing and countermeasures challenge[J]. IEEE Journal of Selected Topics in Signal Processing, 2017, 11(4): 588-604.

[5] TODISCO M, WANG X, VESTMAN V, et al. ASVspoof 2019: Future horizons in spoofed and fake audio detection[C]// Proceedings of Conference of the International Speech Communication Association (INTERSPEECH). Graz, Austria: [s.n.], 2019: 1008-1012.

[6] SZEGEDY C, ZAREMBA W, SUTSKEVER I, et al. Intriguing properties of neural networks[C]//Proceedings of International Conference on Learning Representations (ICLR). Banff, Canada: [s.n.], 2014.

[7] BIGGIO B, ROLI F. Wild patterns: Ten years after the rise of adversarial machine learning[J]. Pattern Recognition, 2018(84): 317-331.

[8] YUAN X, HE P, ZHU Q, et al. Adversarial examples: Attacks and defenses for deep learning[J]. IEEE Transactions on Neural Networks and Learning Systems, 2019, 30(9): 2805-2824.

[9] PAPERNOT N, MCDANIEL P, GOODFELLOW I, et al. Practical black-box attacks against machine learning[C]// Proceedings of Asia Conference on Computer and Communications Security (AsiaCCS). Abu Dhabi, United Arab Emirates: [s.n.], 2017: 506-519.

[10] VIVEK B S, MOPURI K R, BABU R V. Gray-box adversarial training[C]//Proceedings of the European Conference on Computer Vision (ECCV). Munich, Germany: Springer, 2018: 203-218.

[11] GOODFELLOW I J, SHLENS J, SZEGEDY C. Explaining and harnessing adversarial examples[C]//Proceedings of International Conference on Learning Representations (ICLR). San Diego, CA, USA: [s.n.], 2015: 1-11.

[12] EYKHOLT K, EVTIMOV I, FERNANDES E, et al. Robust physical-world attacks on deep learning visual classification [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Salt Lake City, UT, USA: IEEE, 2018: 1625-1634.

[13] KREUK F, ADI Y, CISSE M, et al. Fooling end-to-end speaker verification with adversarial examples[C]//Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Calgary, AB, Canada: IEEE, 2018: 1962-1966.

[14] VESTMAN V, KINNUNEN T, HAUTAMÄKI R G, et al. Voice mimicry attacks assisted by automatic speaker verification [J]. Computer Speech & Language, 2020(59): 36-54.

[15] NAKAMURA T, SAITO Y, TAKAMICHI S, et al. V2S attack: Building DNN-based voice conversion from automatic speaker verification[C]//Proceedings of ISCA Speech Synthesis Workshop (SSW). Vienna, Austria: [s.n.], 2019: 161-165.

[16] LIU S, WU H, LEE H, et al. Adversarial attacks on spoofing countermeasures of automatic speaker verification[C]// Proceedings of IEEE Automatic Speech Recognition and Understanding Workshop (ASRU). Singapore: IEEE, 2019: 312-319.

[17] TIAN X, DAS R K, LI H. Black-box attacks on automatic speaker verification using feedback-controlled voice conversion[C]// Proceedings of Odyssey 2020: The Speaker and Language Recognition Workshop. Tokyo, Japan: [s.n.], 2020: 159-164.

[18] WANG Q, OKABE K, LEE K A, et al. A generalized framework for domain adaptation of PLDA in speaker recognition[C]// Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Barcelona, Spain: IEEE, 2020: 6619-6623.

[19] DEHAK N, KENNY P J, DEHAK R, et al. Front-end factor analysis for speaker verification[J]. IEEE Transactions on Audio, Speech, and Language Processing, 2010, 19(4): 788-798.

[20] LEI Y, SCHEFFER N, FERRER L, et al. A novel scheme for speaker recognition using a phonetically-aware deep neural network[C]//Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Florence, Italy: IEEE, 2014: 1695-1699.

[21] SNYDER D, GARCIA-ROMERO D, POVEY D. Time delay deep neural network-based universal background models for speaker recognition[C]//Proceedings of IEEE Automatic Speech Recognition and Understanding Workshop (ASRU). Scottsdale, AZ, USA: IEEE, 2015: 92-97.

[22] DO C T, BARRAS C, LE V B, et al. Augmenting short-term cepstral features with long-term discriminative features for speaker verification of telephone data[C]//Proceedings of Conference of the International Speech Communication Association (INTERSPEECH). Lyon, France: [s.n.], 2013: 2484-2488.

[23] SARKAR A K, DO C T, LE V B, et al. Combination of cepstral and phonetically discriminative features for speaker verification[J]. IEEE Signal Processing Letters, 2014, 21(9): 1040-1044.

[24] RICHARDSON F, REYNOLDS D A, DEHAK N. A unified deep neural network for speaker and language recognition[C]//

Proceedings of Annual Conference of the International Speech Communication Association (INTERSPEECH). Dresden, Germany: [s.n.], 2015: 1146-1150.

[25] MCLAREN M, LEI Y, FERRER L. Advances in deep neural network approaches to speaker recognition[C]//Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). South Brisbane, Queensland, Australia: IEEE, 2015: 4814-4818.

[26] VARIANI E, LEI X, MCDERMOTT E, et al. Deep neural networks for small footprint text-dependent speaker verification [C]//Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Florence, Italy: IEEE, 2014: 4052-4056.

[27] SNYDER D, GARCIA-ROMERO D, POVEY D, et al. Deep neural network embeddings for text-independent speaker verification[C]//Proceedings of Conference of the International Speech Communication Association (INTERSPEECH). Stockholm, Sweden: [s.n.], 2017: 999-1003.

[28] SNYDER D, GARCIA-ROMERO D, SELL G, et al. X-vectors: Robust DNN embeddings for speaker recognition[C]// Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Calgary, AB, Canada: IEEE, 2018: 5329-5333.

[29] MARKHAM D. Phonetic imitation, accent, and the learner[M]. Sweden: Lund University, 1997.

[30] LAU Y W, WAGNER M, TRAN D. Vulnerability of speaker verification to voice mimicking[C]//Proceedings of International Symposium on Intelligent Multimedia, Video and Speech Processing. Hong Kong, China: IEEE, 2004: 145-148.

[31] HAUTAMÄKI R G, KINNUNEN T, HAUTAMÄKI V, et al. I-vectors meet imitators: On vulnerability of speaker verification systems against voice mimicry[C]//Proceedings of Conference of the International Speech Communication Association (INTERSPEECH). Lyon, France: [s.n.], 2013: 930-934.

[32] ROSENBERG A E. Automatic speaker verification: A review[J]. Proceedings of the IEEE, 1976, 64(4): 475-487.

[33] 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. Berlin, Heidelberg: Springer, 2005: 15-21.

[34] FARRÚS M, WAGNER M, ANGUITA J, et al. How vulnerable are prosodic features to professional imitators? [C]// Proceedings of Odyssey 2008: The Speaker and Language Recognition Workshop. Stellenbosch, South Africa: [s.n.], 2008: 1-4.

[35] HAUTAMÄKI R G, KINNUNEN T, HAUTAMÄKI V, et al. I-vectors meet imitators: on vulnerability of speaker verification systems against voice mimicry[C]// Proceedings of Conference of the International Speech Communication Association (INTERSPEECH). Lyon, France: [s.n.], 2013: 930-934.

[36] MARIÉTHOZ J, BENGIO S. Can a professional imitator fool a GMM-based speaker verification system?[EB/OL]. (2006-1-11) [2021-4-25]. https://publications.idiap.ch/downloads/reports/2005/mariethoz-idiap-rr-05-61.pdf.

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

[38] JAIN A K, PRABHAKAR S, PANKANTI S. On the similarity of identical twin fingerprints[J]. Pattern Recognition, 2002, 35(11): 2653-2663.

[39] KERSTA L G, COLANGELO J A. Spectrographic speech patterns of identical twins[J]. The Journal of the Acoustical Society of America, 1970, 47(1A): 58-59.

[40] PATIL H A, PARHI K K. Variable length teager energy based mel cepstral features for identification of twins[C]// Proceedings of International Conference on Pattern Recognition and Machine Intelligence (PReMI). Berlin, Heidelberg: Springer, 2009: 525-530.

[41] LINDBERG J, BLOMBERG M. Vulnerability in speaker verification—A study of technical impostor techniques[C]// Proceedings of European Conference on Speech Communication and Technology (EuroSpeech). Budapest, Hungary: [s.n.], 1999: 1-4.

[42] VILLALBA J, LLEIDA E. Speaker verification performance degradation against spoofing and tampering attacks[C]// Proceedings of FALA workshop. Vigo, Spain: [s.n.], 2010: 131-134.

[43] WU Z, KINNUNEN T, EVANS N, et al. ASVspoof 2015: The first automatic speaker verification spoofing and countermeasures challenge[C]//Proceedings of Conference of the International Speech Communication Association (INTERSPEECH). Dresden, Germany: [s.n.], 2015: 2037-2041.

[44] KINNUNEN T, EVANS N, YAMAGISHI J, et al. ASVspoof 2017: Automatic speaker verification spoofing and

countermeasures challenge evaluation plan[EB/OL]. (2018-09-19) [2021-04-25]. https:// www. asvspoof. org/data2017/asvspoof-2017_evalplan_v1.2.pdf.

[45] SHANG W, STEVENSON M. Score normalization in playback attack detection[C]//Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Dallas, Texas, USA: IEEE, 2010: 1678-1681.

[46] WANG Z F, WEI G, HE Q H. Channel pattern noise-based playback attack detection algorithm for speaker recognition[C]//Proceedings of International Conference on Machine Learning and Cybernetics (ICMLC). Guilin, China: IEEE, 2011: 1708-1713.

[47] VILLALBA J, LLEIDA E. Detecting replay attacks from far-field recordings on speaker verification systems[C]//Proceedings of European Workshop on Biometrics and Identity Management (BioID). Berlin, Heidelberg: Springer, 2011: 274-285.

[48] WU Z, GAO S, CLING E S, et al. A study on replay attack and anti-spoofing for text-dependent speaker verification[C]//Proceedings of Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA). Chiang Mai, Thailand: IEEE, 2014: 1-5.

[49] GAŁKA J, GRZYWACZ M, SAMBORSKI R. Playback attack detection for text-dependent speaker verification over telephone channels[J]. Speech Communication, 2015, 67: 143-153.

[50] DELGADO H, TODISCO M, SAHIDULLAH M, et al. ASVspoof 2017 Version 2.0: Meta-data analysis and baseline enhancements[C]//Proceedings of Odyssey 2018: The Speaker and Language Recognition Workshop. Les Sables d'Olonne, France: [s.n.], 2018: 296-303.

[51] YOON S H, KOH M S, PARK J H, et al. A new replay attack against automatic speaker verification systems[J]. IEEE Access, 2020, 8: 36080-36088.

[52] GONG Y, YANG J, HUBER J, et al. ReMASC: Realistic replay attack corpus for voice controlled systems[C]//Proceedings of Conference of the International Speech Communication Association (INTERSPEECH). Graz, Austria: [s.n.], 2019: 2355-2359.

[53] STYLIANOU Y, CAPPÉ O, MOULINES E. Continuous probabilistic transform for voice conversion[J]. IEEE Transactions on Speech and Audio Processing, 1998, 6(2): 131-142.

[54] STYLIANOU Y, LAROCHE J, MOULINES E. High-quality speech modification based on a harmonic+ noise model[C]//Proceedings of European Conference on Speech Communication and Technology (EUROSPEECH). Madrid, Spain: [s.n.], 1995: 451-454.

[55] KAWAHARA H, MASUDA-KATSUSE I, DE CHEVEIGNE A. Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds[J]. Speech Communication, 1999, 27(3/4): 187-207.

[56] LORENZO-TRUEBA J, YAMAGISHI J, TODA T, et al. The voice conversion challenge 2018: Promoting development of parallel and nonparallel methods[C]//Proceedings of Odyssey 2018: The Speaker and Language Recognition Workshop. Les Sables d'Olonne, France: [s.n.], 2018: 195-202.

[57] OORD A, DIELEMAN S, ZEN H, et al. WaveNet: A generative model for raw audio[C]//Proceedings of ISCA Speech Synthesis Workshop (SSW). Sunnyvale, CA, USA: [s.n.], 2016: 125-129.

[58] ASVSPOOF Consortium. ASVspoof2019: Automatic speaker verification spoofing and countermeasures challenge evaluation plan[EB/OL]. (2019-01-15) [2021-04-25]. https://www.asvspoof.org/ asvspoof2019/asvspoof2019_ evaluation_plan.pdf.

[59] ABE M, NAKAMURA S, SHIKANO K, et al. Voice conversion through vector quantization[C]//Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). New York, USA: IEEE, 1988: 655-658.

[60] PELLOM B L, HANSEN J H L. An experimental study of speaker verification sensitivity to computer voice-altered imposters [C]//Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Phoenix, Arizona, USA: IEEE, 1999: 837-840.

[61] PATRICK P Z, AVERSANO G, BLOUET R, et al. Voice forgery using ALISP: Indexation in a client memory[C]//Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Philadelphia, Pennsylvania, USA: IEEE, 2005: 17-20.

[62] MATROUF D, BONASTRE J F, FREDOUILLE C. Effect of speech transformation on impostor acceptance[C]//Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Toulouse, France: IEEE, 2006: 933-936.

[63] BONASTRE J F, MATROUF D, FREDOUILLE C. Artificial impostor voice transformation effects on false acceptance rates

[C]//Proceedings of Conference of the International Speech Communication Association (INTERSPEECH). Antwerp, Belgium: [s.n.], 2007: 2053-2056.

[64] 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]//Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Kyoto, Japan: IEEE, 2012: 4401-4404.

[65] WU Z, KINNUNEN T, CHNG E S, et al. A study on spoofing attack in state-of-the-art speaker verification: The telephone speech case[C]//Proceedings of Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA). Hollywood, CA, USA: IEEE, 2012: 1-5.

[66] WU Z, LI H. Voice conversion and spoofing attack on speaker verification systems[C]//Proceedings of Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA). Kaohsiung, Taiwan, China: IEEE, 2013: 1-9.

[67] ALEGRE F, VIPPERLA R, EVANS N. Spoofing countermeasures for the protection of automatic speaker recognition systems against attacks with artificial signals[C]//Proceedings of Conference of the International Speech Communication Association (INTERSPEECH). Portland, Oregon, USA: [s.n.], 2012: 1688-1691.

[68] ALEGRE F, AMEHRAYE A, EVANS N. A one-class classification approach to generalized speaker verification spoofing countermeasures using local binary patterns[C]//Proceedings of IEEE International Conference on Biometrics: Theory, Applications and Systems (BTAS). Arlington, VA, USA: IEEE, 2013: 1-8.

[69] ALEGRE F, AMEHRAYE A, EVANS N. Spoofing countermeasures to protect automatic speaker verification from voice conversion[C]//Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Vancouver, BC, Canada: IEEE, 2013: 3068-3072.

[70] STYLIANOU Y, CAPPÉ O, MOULINES E. Continuous probabilistic transform for voice conversion[J]. IEEE Transactions on Speech and Audio Processing, 1998, 6(2): 131-142.

[71] TODA T, BLACK A W, TOKUDA K. Spectral conversion based on maximum likelihood estimation considering global variance of converted parameter[C]//Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Philadelphia, Pennsylvania, USA: IEEE, 2005: 9-12.

[72] DESAI S, RAGHAVENDRA E V, YEGNANARAYANA B, et al. Voice conversion using artificial neural networks[C]//Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Taipei, China: IEEE, 2009: 3893-3896.

[73] SHUANG Z W, BAKIS R, SHECHTMAN S, et al. Frequency warping based on mapping formant parameters[C]//Proceedings of Conference of the International Speech Communication Association (INTERSPEECH). Pittsburgh, PA, USA: [s.n.], 2006: 2290-2293.

[74] ERRO D, MORENO A. Weighted frequency warping for voice conversion[C]//Proceedings of Conference of the International Speech Communication Association (INTERSPEECH). Antwerp, Belgium: [s.n.], 2007: 1965-1968.

[75] WU Z, VIRTANEN T, KINNUNEN T, et al. Exemplar-based voice conversion using non-negative spectrogram deconvolution[C]//Proceedings of ISCA Speech Synthesis Workshop (SSW). Barcelona, Spain: [s.n.], 2013: 201-206.

[76] 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): 753-770.
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): 753-770.

[77] KANEKO T, KAMEOKA H. Parallel-data-free voice conversion using cycle-consistent adversarial networks[EB/OL]. (2017-12-20) [2021-04-25]. https://arxiv.org/abs/1711.11293.

[78] KINNUNEN T, JUVELA L, ALKU P, et al. Non-parallel voice conversion using i-vector PLDA: Towards unifying speaker verification and transformation[C]//Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). New Orleans, LA, USA: IEEE, 2017: 5535-5539.

[79] HSU C C, HWANG H T, WU Y C, et al. Voice conversion from non-parallel corpora using variational auto-encoder[C]//Proceedings of Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA). Jeju, South Korea: IEEE, 2016: 1-6.

[80] CHOU J, YEH C, LEE H, et al. Multi-target voice conversion without parallel data by adversarial learning disentangled audio representations[C]//Proceedings of Conference of the International Speech Communication Association (INTERSPEECH). Hyderabad, India: [s.n.], 2018: 501-505.

[81] QIAN K, ZHANG Y, CHANG S, et al. Autovc: Zero-shot voice style transfer with only autoencoder loss[C]//Proceedings of

International Conference on Machine Learning (ICML). Long Beach, California, USA: PMLR, 2019: 5210-5219.

[82] KAMBLE M R, SAILOR H B, PATIL H A, et al. Advances in anti-spoofing: From the perspective of ASVspoof challenges [J]. APSIPA Transactions on Signal and Information Processing, 2020. DOI: 10.1017/ATSIP.2019.21.

[83] TIAN X. Average modeling approach to voice conversion with non-parallel data[C]//Proceedings of Odyssey 2018: The Speaker and Language Recognition Workshop. Les Sables d'Olonne, France: [s.n.], 2018: 227-232.

[84] FOOMANY F H, HIRSCHFIELD A, INGLEBY M. Toward a dynamic framework for security evaluation of voice verification systems[C]//Proceedings of IEEE Toronto International Conference Science and Technology for Humanity (TIC-STH). Toronto, Canada: IEEE, 2009: 22-27.

[85] MASUKO T, TOKUDA K, KOBAYASHI T. Imposture using synthetic speech against speaker verification based on spectrum and pitch[C]//Proceedings of Conference of the International Speech Communication Association (INTERSPEECH). Beijing, China: [s.n.], 2000: 302-305.

[86] 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.

[87] GALOU G, CHOLLET G. Synthetic voice forgery in the forensic context: a short tutorial[C]//Proceedings of Forensic Speech and Audio Analysis Working Group (ENFSI-FSAAWG). Rome, Italy: [s.n.], 2011: 1-3.

[88] WU Z, WATTS O, KING S. Merlin: An open-source neural network speech synthesis system[C]//Proceedings of ISCA Speech Synthesis Workshop (SSW). Sunnyvale, CA, USA: [s.n.], 2016: 202-207.

[89] WENINGER F, BERGMANN J, SCHULLER B. Introducing CURRENNT: The Munich open-source CUDA recurrent neural network toolkit[J]. Journal of Machine Learning Research, 2015(16): 547-551.

[90] SCHRÖDER M, TROUVAIN J. The German text-to-speech synthesis system MARY: A tool for research, development and teaching[J]. International Journal of Speech Technology, 2003, 6(4): 365-377.

[91] ARIK S Ö, CHRZANOWSKI M, COATES A, et al. Deep voice: Real-time neural text-to-speech[C]//Proceedings of International Conference on Machine Learning (ICML). Sydney, NSW, Australia: PMLR, 2017: 195-204.

[92] ARIK S Ö, DIAMOS G, GIBIANSKY A, et al. Deep voice 2: Multi-speaker neural text-to-speech[C]//Proceedings of International Conference on Neural Information Processing Systems (NeurIPS). Long Beach, CA, USA: [s.n.], 2017: 2966-2974.

[93] WANG W, XU S, XU B. First step towards end-to-end parametric TTS synthesis: Generating spectral parameters with neural attention[C]//Proceedings of Conference of the International Speech Communication Association (INTERSPEECH). San Francesco, CA, USA: [s.n.], 2016: 2243-2247.

[94] WANG Y, SKERRY-RYAN R J, STANTON D, et al. Tacotron: Towards end-to-end speech synthesis[C]//Proceedings of Conference of the International Speech Communication Association (INTERSPEECH). Stockholm, Sweden: [s.n.], 2017: 4006-4010.

[95] SHEN J, PANG R, WEISS R J, et al. Natural TTS synthesis by conditioning wavenet on mel spectrogram predictions[C]//Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Calgary, AB, Canada: IEEE, 2018: 4779-4783.

[96] PING W, PENG K, GIBIANSKY A, et al. Deep voice 3: 2000-speaker neural text-to-speech[C]//Proceedings of International Conference on Learning Representations (ICLR). Vancouver, BC, Canada: [s.n.], 2018: 214-217.

[97] HSU W N, ZHANG Y, WEISS R J, et al. Hierarchical generative modeling for controllable speech synthesis[C]//Proceedings of International Conference on Learning Representations (ICLR). New Orleans, LA, USA: [s.n.], 2019: 1-27.

[98] KREUK F, ADI Y, CISSE M, et al. Fooling end-to-end speaker verification with adversarial examples[C]//Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Calgary, AB, Canada: IEEE, 2018: 1962-1966.

[99] GONG Y, POELLABAUER C. Crafting adversarial examples for speech paralinguistics applications[C]//Proceedings of Dynamic and Novel Advances in Machine Learning and Intelligent Cyber Security (DYNAMICS) Workshop. San Juan, Puerto Rico, USA: [s.n.], 2018: 1-8.

[100] LI X, ZHONG J, WU X, et al. Adversarial attacks on GMM i-vector based speaker verification systems[C]//Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Barcelona, Spain: IEEE, 2020: 6579-6583.

[101] XIE Y, SHI C, LI Z, et al. Real-time, universal, and robust adversarial attacks against speaker recognition systems[C]//

Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Barcelona, Spain: IEEE, 2020: 1738-1742.

[102] POVEY D, GHOSHAL A, BOULIANNE G, et al. The Kaldi speech recognition toolkit[C]//Proceedings of IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU). Hawaii, USA: IEEE, 2011: 1-5.

[103] LI J, ZHANG X, JIA C, et al. Universal adversarial perturbations generative network for speaker recognition[C]//Proceedings of IEEE International Conference on Multimedia and Expo (ICME). London, UK: IEEE, 2020: 1-6.

[104] WANG Q, GUO P, XIE L. Inaudible adversarial perturbations for targeted attack in speaker recognition[C]//Proceedings of Conference of the International Speech Communication Association (INTERSPEECH). Shanghai, China: [s.n.], 2020: 4228-4232.

[105] VILLALBA J, ZHANG Y, DEHAK N. X-vectors meet adversarial attacks: benchmarking adversarial robustness in speaker verification[C]//Proceedings of Conference of the International Speech Communication Association (INTERSPEECH). Shanghai, China: [s.n.], 2020: 4233-4237.

[106] ZHANG Y, JIANG Z, VILLALBA J, et al. Black-box attacks on spoofing countermeasures using transferability of adversarial examples[C]//Proceedings of Conference of the International Speech Communication Association (INTERSPEECH). Shanghai, China: [s.n.], 2020: 4238-4242.

[107] CHEN G, CHEN S, FAN L, et al. Who is real Bob? Adversarial attacks on speaker recognition systems[C]//Proceedings of IEEE Symposium on Security and Privacy Workshops (SPW). San Francisco, CA, USA: [s.n.], 2021: 1-18.

[108] LI Z, SHI C, XIE Y, et al. Practical adversarial attacks against speaker recognition systems[C]//Proceedings of Annual International Workshop on Mobile Computing Systems and Applications (HotMobile). Austin, Texas, USA:ACM, 2020: 9-14.

[109] MARRAS M, KORUS P, MEMON N D, et al. Adversarial optimization for dictionary attacks on speaker verification[C]//Proceedings of Conference of the International Speech Communication Association (INTERSPEECH). Graz, Austria: [s.n.], 2019: 2913-2917.

[110] NAKAMURA T, SAITO Y, TAKAMICHI S, et al. V2S attack: Building DNN-based voice conversion from automatic speaker verification[C]//Proceedings of ISCA Speech Synthesis Workshop (SSW). Vienna, Austria: [s.n.], 2019: 161-165.

[111] LIU S, WU H, LEE H, et al. Adversarial attacks on spoofing countermeasures of automatic speaker verification[C]//Proceedings of IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU). Singapore: IEEE, 2019: 312-319.

[112] KURAKIN A, GOODFELLOW I J, BENGIO A S. Adversarial examples in the physical world[C]//Proceedings of International Conference on Learning Representations (ICLR). Toulon, France: [s.n.], 2017: 1-14.

[113] CARLINI N, WAGNER D. Towards evaluating the robustness of neural networks[C]//Proceedings of IEEE Symposium on Security and Privacy (S&P). San Jose, CA, USA: IEEE, 2017: 39-57.

[114] MADRY A, MAKELOV A. Towards deep learning models resistant to adversarial attacks[C]//Proceedings of International Conference on Learning Representations (ICLR). Vancouver, BC, Canada: [s.n.], 2018: 1-28.

[115] 张雄伟, 李嘉康, 孙蒙, 等. 语音欺骗检测方法的研究现状及展望[J]. 数据采集与处理, 2020, 35(5): 807-823.
ZHANG Xiongwei, LI Jiakang, SUN Meng, et al. Speech anti-spoofing: The state of the art and prospects[J]. Journal of Data Acquisition and Processing, 2020, 35(5): 807-823.

**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。

**张星昱**(1994-),**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:zxynbnb@126.com。

**孙蒙**(1984-),男,副教授,研究方向:智能语音处理、机器学习。

**邹霞**(1979-),男,副教授,研究方向:语音信号处理、语音编码。

(编辑:张彤)