计算机研究与发展
Journal of Computer Research and Development
DOI:10.7544/issn1000-1239.2021.20210632
58(11):2350-2363,2021

基于单“音频像素”扰动的说话人识别隐蔽攻击

沈轶杰[1] 李良澄[1] 刘子威[1] 刘天天[1] 罗 浩[1] 沈 汀[3] 林 峰[1,2] 任 奎[1]
[1](浙江大学网络空间安全研究中心 杭州 310027)
[2](浙江省区块链与网络空间治理重点实验室(浙江大学) 杭州 310027)
[3](浙江东安检测技术有限公司 杭州 310063)
(shenyijie@zju.edu.cn)

Stealthy Attack Towards Speaker Recognition Based on One-“Audio Pixel” Perturbation

Shen Yijie[1], Li Liangcheng[1], Liu Ziwei[1], Liu Tiantian[1], Luo Hao[1], Shen Ting[3], Lin Feng[1,2], and Ren Kui[1]
[1](*Institute of Cyberspace Research, Zhejiang University, Hangzhou* 310027)
[2](*Key Laboratory of Blockchain and Cyberspace Governance of Zhejiang Province (Zhejiang University), Hangzhou* 310027)
[3](*Zhejiang Dong'an Testing Technology Co., Ltd., Hangzhou* 310063)

Abstract Attacks towards the speaker recognition system need to inject a long-time perturbation, so it is easy to be detected by machines or administrators. We propose a novel attack towards the speaker recognition based on one-“audio pixel”. Such attack uses the black-box characteristics and search mode of the differential evolution algorithm that does not rely on the model and the gradient information. It overcomes the problem in previous works that the disturbance duration cannot be constrained. Thus, our attack effectively spoofs the speaker recognition via one-“audio pixel” perturbation. In particular, we design a candidate point construction model based on the audio-point-disturbance tuple targeting time series of audio data. It solves the problem that candidate points of differential evolution algorithm are difficult to be described against our attack. The success rate of our attack achieves 100% targeting 60 people in LibriSpeech dataset. In addition, we also conduct abundant experiments to explore the impact of different conditions (e.g., gender, dataset and speaker recognition method) on the performance of our stealthy attack. The result of above experiments provides guidance for effective attacks. At the same time, we put forward ideas based on denoising, reconstruction algorithm and speech compression to defend against our stealthy attack, respectively.

Key words one-“audio pixel” perturbation; black-box attack; speaker recognition; differential evolution algorithm; perturbation attack

收稿日期:2021-06-11;**修回日期**:2021-08-12

基金项目:国家重点研发计划项目(2020AAA0107700);国家自然科学基金项目(62032021,61772236,61972348);浙江省重点研发计划项目(2019C03133);浙江省引进培育领军型创新创业团队项目(2018R01005);阿里巴巴-浙江大学前沿技术联合研究中心项目;网络空间国际治理研究基地项目

This work was supported by the National Key Research and Development Program of China (2020AAA0107700), the National Natural Science Foundation of China (62032021, 61772236, 61972348), Zhejiang Key Research and Development Plan (2019C03133), the Leading Innovative and Entrepreneur Team Introduction Program of Zhejiang (2018R01005), the Fund of Alibaba-Zhejiang University Joint Institute of Frontier Technologies, and the Fund of Research Institute of Cyberspace Governance in Zhejiang University.

通信作者:林峰(flin@zju.edu.cn)

摘 要 目前针对说话人识别的攻击需要对音频注入长时间的扰动,因此容易被机器或者管理人员发现.提出了一种新颖的基于单“音频像素”扰动的针对说话人识别的隐蔽攻击.该攻击利用了差分进化算法不依赖于模型的黑盒特性和不依赖梯度信息的搜索模式,克服了已有攻击中扰动时长无法被约束的问题,实现了使用单“音频像素”扰动的有效攻击.特别地,设计了一种基于音频段-音频点-扰动值多元组的候选点构造模式,针对音频数据的时序特性,解决了在攻击方案中差分进化算法的候选点难以被描述的问题.攻击在 LibriSpeech 数据集上针对 60 个人的实验表明这一攻击能达到 100% 的成功率.还开展了大量的实验探究不同条件(如性别、数据集、说话人识别方法等)对于隐蔽攻击性能的影响.上述实验的结果为进行有效地攻击提供了指导.同时,提出了分别基于去噪器、重建算法和语音压缩的防御思路.

关键词 单“音频像素”扰动;黑盒攻击;说话人识别;差分进化算法;扰动攻击

中图法分类号 TP309

说话人识别技术通过对说话人声纹的分析识别说话人身份,是目前应用广泛的生物认证技术之一.该技术已经被广泛应用于个人安全和社会安全领域(如个人设备管理[1]、电子取证[2]以及电子监控[3]等).然而,目前主流的说话识别系统存在着重大安全隐患,即攻击者可以通过来源于第三方的音频(即非来源于受害者的音频)获取目标系统(如安卓操作系统、支付宝、微信等)中受害者的权限,执行查看隐私、交易支付、登入社交账号等操作.这些操作会威胁受害者的隐私信息、经济利益甚至人身安全.

前人的工作提出了一系列基于机器学习的攻击方案.这些攻击运用对抗学习的技术,生成特殊的扰动使第三方音频伪装成受害者的身份,实现入侵系统的目的.根据攻击者对模型信息的获取程度可以分为白盒攻击[4-5]和黑盒攻击[6-8].其中,白盒攻击假设攻击者需要获取模型的完整参数,而黑盒攻击假设攻击者不需要获得模型的任何参数.上述方案取得了一定的效果,但是也存在 2 点不足:1)白盒攻击依赖于被攻击模型的完整参数,这一约束降低了攻击的实用性.2)人工复查是目前检测语音识别是否遭受攻击的主要手段之一.因此,提升攻击对于人耳的隐蔽性是提高攻击实用性的重要环节.由于增加扰动会引入宽频噪声,这样的噪声根据人耳的“掩蔽效应”[9]容易被人耳所察觉.因此注入扰动的时长越长,攻击被察觉的可能性越大.然而目前的攻击方案需要向第三方音频注入亚秒级甚至秒级的扰动,导致攻击易被察觉.存在这一不足是因为现有的白盒和黑盒攻击方案都依赖于梯度信息,所以容易陷入局部最优解,形成对于攻击能力的限制,即攻击者无法通过修改对攻击成功增益最大的采样点实现将第三方音频伪装成受害者的目的.

随着深度学习的发展,基于深度学习的说话人识别技术(如 x-vector[10]和 d-vector[11-12])由于其高精确以及高鲁棒性成为了目前该领域的主流技术.现有的工作[13-14]指出了深度学习技术在实现特征提取的过程中位于决策边界附近的数据点对于特定方向的扰动的敏感性.特别地,本文利用这一特性试图实现一种针对说话人识别的高隐蔽性扰动攻击.为了实现针对说话人识别系统的攻击并克服现有工作不足,本文攻击方案需要满足 3 个特性:

1) 黑盒攻击.攻击者不需要获取任何说话人识别系统中模型的参数信息,这一特性增强攻击的实用性.

2) 有目标的攻击.攻击能够将第三方音频伪装成目标受害者,这一特性使攻击具有有效性和针对性.

3) 单“音频像素”扰动.单“音频像素”指音频中的单个独立采样点,类比于图像中的一个像素点,是音频采集过程中最小的记录单位.在扰动生成过程中能够搜索对攻击增益最高的“音频像素”并进行注入.这一特性强化攻击的高效性和隐蔽性.

为了实现以上 3 个特性,图 1 展示了我们针对说话人识别基于单“音频像素”扰动的攻击流程,下文简称这种攻击为基于单“音频像素”扰动的攻击.攻击者先在第三方音频上搜索对攻击增益最高的“音频像素”,并通过向搜索到的“音频像素”注入扰动产生能够伪装成受害者身份的攻击音频,最后攻击者使用该音频实现攻击.为了实现这套方案需要解决 2 个技术挑战:1)如何在黑盒条件下实现针对说话人识别系统的攻击?2)如何使得攻击能够搜索并修改对攻击增益最高的“音频像素”?

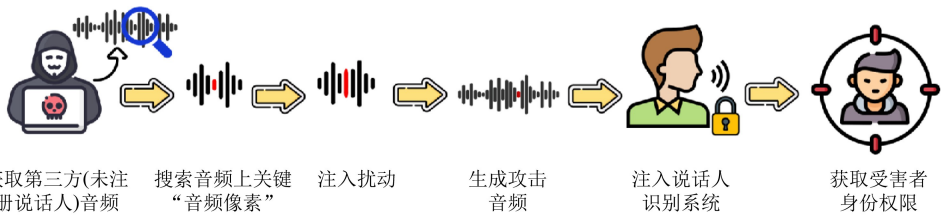


Fig. 1 The workflow of the attack aiming the speaker verification system via one-“audio pixel” perturbation

图 1 单“音频像素”扰动实现针对说话人识别系统的攻击流程

本文提出了一种基于差分进化算法[15]的针对说话人识别技术的攻击方案,以下简称基于单“音频像素”扰动的攻击.首先,利用差分进化算法的启发式特性,实现不依赖于系统参数的攻击.其次,提出基于音频段-音频点-扰动值多元组的候选点构造模式,结合差分进化算法不依赖梯度的特点,对候选点进行迭代和优化,实现搜索并修改对攻击增益最高的“音频像素”,生成能够伪装为受害者的攻击音频.我们在 LibriSpeech 数据集[16]上测试了该方案,攻击成功率达到了 100%.另外,我们对影响基于单“音频像素”扰动的攻击因素进行了评估,为获得高性能攻击提供了指导.除此之外,利用实验探究了不同平台和跨数据集下攻击的性能,说明了攻击不依赖于平台和数据集.

本文的主要贡献有 3 个方面:

1) 提出和探索了利用单“音频像素”扰动实现对于说话人识别系统的隐蔽攻击,攻击注入扰动的时长仅为几十微妙.

2) 设计了基于音频段-音频点-扰动值多元组的候选点构造模式,结合差分进化算法,实现了在音频上搜索对攻击增益最高的“音频像素”并注入有效的扰动.

3) 开展了广泛的实验,探究了不同条件对于攻击的影响.同时,通过实验说明基于单“音频像素”扰动的攻击隐蔽性优于目前最先进的攻击工作.

1 背景及相关工作

1.1 说话人识别系统

相比于其他生物认证技术,说话人识别技术具有低成本、高精确度和非接触的特点.经过数十年的发展,说话人识别系统已经具有大量的开源平台(如 Kaldi[17], MSR Identity Toolkit[18], ALIZE[19]和 SIDEKIT[20])以及商业项目(Google home[21], Talentedsoft[22]).根据对于识别时语料内容的约束,说话人识别系统可以分为文本相关的[23-24]和文本无关的[25-26]两类,前者要求用户使用文本内容一致的语料进行注册和识别,而后者没有这一限制.显然,文本无关的说话人识别系统更加便利和实用,因此本文设计的攻击针对文本无关的说话人识别系统展开,即下文中的说话人识别系统默认为文本无关的说话人识别系统.说话人识别系统由 3 个子模块构成,分别为音频预处理模块、特征提取模块以及决策模块.

1) 音频预处理模块.使用人声提取算法如 G.279[27]提取原音频中的人声部分从而消除环境噪声对于识别结果的影响,接着通过梅尔频率倒谱系数(Mel-frequency cepstral coefficients, MFCC)[28]等时频分析方法,获取音频信号的时频信息为进一步的特征提取提供数据.

2) 特征提取模块.特征提取模块通过特定的提取器,提取出与身份相关的特征为身份决策提供可靠的依据.特征提取模块是说话人识别系统的核心模块,也是目前说话人识别技术的主要研究内容.目前的特征提取器大致可以分为基于概率分析的特征提取器和基于深度学习的特征提取器 2 类.

① 基于概率分析的特征提取器.基于概率分析的特征提取器使用概率模型提取语音中的特征.基于高斯混合-通用背景模型(Gaussian mixture model-universal background model, GMM-UBM)框架[29]的特征提取是其中最成熟的特征提取方式,它通过大量无关语料预先构建通用的背景模型,实现不依赖用户的说话人特征提取.而基于概率分析的特征提取器中最先进的是 Dehak 等人[30]提出的 i-vector,它优化 GMM-UBM 模型,基于 i-vector 因子分析技术将信道和说话人作为一个整体分析,实现高性能的说话人特征提取.

② 基于深度学习的特征提取器.基于深度学习的特征提取器使用深度神经网络提取描述对象身份的特征.按照优化提取器性能的方法不同,主要可以

分为2类:一类使用独特的深度神经网络模型来实现高效的特征提取.比如,Google公司[31]提出的d-vector;Shi等人[32]提取的j-vector;Snyder等人[33]提出的x-vector.另一类使用独特的损失函数来实现高效的特征提取.比如,rahman等人[34]和Wan等人[11]先后提出基于元组的端到端损失(tuple-based end-to-end, TE2E)和广义端到端损失(generalized end-to-end, GE2E).

目前基于深度学习的特征提取器性能表现远高于基于统计分析的特征提取器,因此本文的工作针对基于深度学习的特征提取器展开.

3) 决策模块.决策模块可以使用余弦相似度[11]、K临近算法[35]、支持向量机[36]和概率线性判别分析[37]等算法根据特征提取模块提取的特征识别说话人身份.在这一模块根据目的的不同,说话人识别可以细分为说话人确认和说话人辨认2类.由于在日常使用的支付、解锁、登入社交帐号等操作中都以说话人确认为目的,因此本文围绕以说话人确认为目的的说话人识别展开.在发起一次识别后,将输入语音 x 与受害者的记录进行比较.在比较过程中,识别系统通过函数 $f(\cdot)$ 计算 x 与受害人的相似度作为得分.当得分高于分数阈值 θ 时,输出受害者的身份信息;若得分低于 θ 时,则输出失败的标识,具体公式为

$$\text{识别结果}=\begin{cases}\text{受害者的身份信息}, & f(x)\geqslant\theta,\\ \text{失败的标识}, & f(x)<\theta.\end{cases} \tag{1}$$

1.2 对抗音频攻击

2013年Szegedy等人[38]在图像领域提出对抗样本的概念.随着这一技术的应用和发展,Cisse等人[39]将对抗概念应用于语音识别领域,实现了对抗音频攻击.随后,不同研究人员针对语音识别提出了成功率更高和隐蔽性更强的攻击方案[40-42].紧接着,研究人员发现对抗音频攻击同样可以用于攻击说话人识别.因此,随后的工作[4-8]针对说话人识别的对抗音频攻击进行了探索.其中有基于白盒攻击[4-5]的工作,通过对于说话人识别模型参数的学习,生成能够伪装成受害者的攻击音频.然而,白盒攻击需要依赖于被攻击模型的参数,这些参数在实际场景中往往被第三方所保护,攻击者难以获取,因此这类攻击的实际应用价值很低.Li等人[7]和Chen等人[6]的工作分别提出了针对说话人识别系统的黑盒攻击方案,解决了以往工作中对于被攻击模型参数的依赖问题.然而,这些工作中对于注入时长没有约束,攻击时需要对于整段音频进行注入,引入长时间的宽带噪声易于被察觉.Li等人[8]为了降低长时间噪声的影响提出了一种亚秒级的注入攻击,但是注入扰动的时间仍然超过0.2 s,无法实现"音频像素"级的注入攻击.因此,已有工作在实现高隐蔽性对抗音频攻击的方面仍然存在不足.

相较于已有工作,基于单"音频像素"扰动的攻击方案不依赖于梯度信息,利用差分进化算法和独特的候选点构造模式,通过修改单个"音频像素"实现有效的黑盒攻击,大幅度地减小了注入时间,从而获得更高的隐蔽性.

1.3 差分进化算法

差分进化算法(differential evolution, DE)[15]是一种基于种群的优化算法,适合用来实现基于单"音频像素"扰动的攻击,因为该算法具有3个特性:

1) 全局最优性.差分进化算法在迭代过程中利用随机性选择的候选点进行择优进化.这使得差分进化算法能够在全局搜索解,避免陷入局部最优解.

2) 可迁移性.差分进化算法是一种不依赖于目标模型的黑盒算法,因此对于攻击的目标系统有良好的可迁移性,即当更换攻击目标系统时不需要对算法实现进行重写.

3) 可优化性.本文中所使用的是标准的差分进化算法.差分进化算法发展至今已经出现了多种变体以满足不同的优化需求,如使用模糊逻辑控制器加快收敛速度[43]和使用自适应参数选择优化算法性能[44].

除了这3个算法理论的特性以外,差分进化算法已经在攻击图像识别领域上展现了优异的性能.Su等人[13]利用差分进化算法实现了一种基于单像素的攻击方案,通过改变图像中关键像素的像素值,实现干扰系统识别结果的目的.然而,由于图像和音频在表现形式和特征蕴含方式上都有巨大的差异.因此我们设计了一套独特的差分进化算法构造模式,辅助我们将差分进化算法应用于攻击说话人识别系统,实现基于单"音频像素"扰动的攻击方案,具体的方案会在第3节中详细讨论.

2 攻击模型

在基于单"音频像素"扰动的攻击中,攻击者通过在第三方音频上注入单"音频像素"扰动,即改变第三方音频中的一个采样点,将第三方音频伪装成受害者身份.为了使攻击者具有足够的能力和约束,我们对攻击者有4方面假设:

1) 攻击者有足够大的语料库用于攻击.目前,开源社区中包含大量语音开源数据库(如VCTK[45],TIMIT[46],CMU_ARCTIC[47]等)以及网络上存在大量音频信息,因此攻击者可以通过网络获取语料库从而满足这一假设.

2) 攻击者进行黑盒攻击,即攻击者只能获取说话人识别系统反馈的评分数值和最终给出的身份.这一假设符合目前主流说话人识别系统给出识别结果的形式.

3) 说话人识别系统提供用户接口进行查询,例如 Talentedsoft[22].攻击者可以通过这些接口注入攻击.同时,它也符合已有工作对于注入扰动攻击的假设[6].

4) 受害者的语料数据不能被获得,攻击者无法使用偷录等方式获得被害者语料.受害者对于自身的保护可以满足这一假设.

相较于现有的攻击方案,基于单"音频像素"扰动的攻击在保证高性能的前提下,利用差分进化算法,突破以往攻击注入时间长而易被发现的问题,实现基于单"音频像素"扰动的隐蔽攻击.

3 基于单"音频像素"扰动的攻击设计

3.1 攻击目标

图 2 为基于单"音频像素"扰动的攻击目标.受害者的身份标识为 v,目标说话人识别的相似度比较为函数 $f(\cdot)$,函数的输入值为一段语音数据,输出值为输入值和系统中受害者记录的相似度,当输出值大于阈值 θ 时,系统认为输入值来源于受害者.

基于单"音频像素"扰动的攻击需要满足 3 个目标:1)攻击者能伪装成受害者进入系统;2)攻击者能够修改的采样点数量为 1;3)生成的扰动注入第三方音频后,幅值绝对值不能超出音频编码允许的最大值.

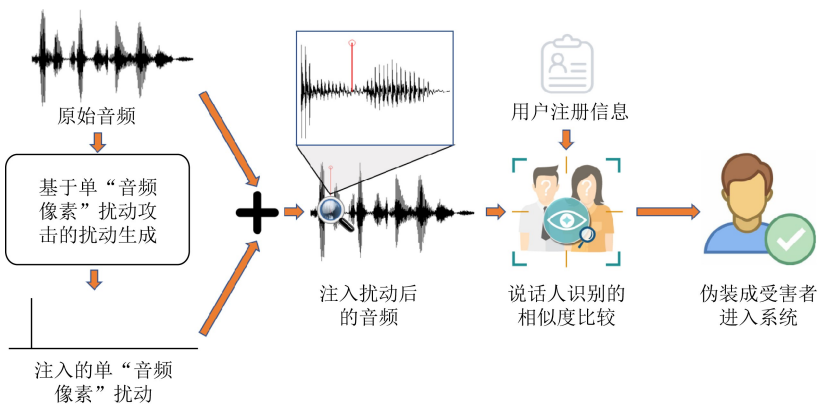


Fig. 2 The goal of the attack based on one-'audio pixel' perturbation

图 2 基于单"音频像素"扰动的攻击目标

在这 3 个目标下,为了更好地建立攻击的数学模型,我们对这 3 个目标进行了公式化的描述.假设一个 n 维向量 $\boldsymbol{x}=(x_1,x_2,\cdots,x_n)$ 表示第三方音频,攻击的目标是生成一个 n 维扰动 $\boldsymbol{p}=(p_1,p_2,\cdots,p_n)$,使得系统识别 $\boldsymbol{x}+\boldsymbol{p}$ 的结果是 v,定义为

$$f(\boldsymbol{x}+\boldsymbol{p})\geqslant\theta. \tag{2}$$

接着,为了实现通过单"音频像素"扰动的目标,我们进一步地约束扰动为

$$p_i=\begin{cases}0, & i\notin k,\\ D_i, & i\in k,\end{cases} \tag{3}$$

其中,k 为算法搜索到的进行攻击的"音频像素"位置的集合,当基于单"音频像素"扰动实现攻击时,k 中的元素数量为 1,D_i 为对应位置 i 处需要引入扰动的大小.在此基础上,为了保证扰动 $\boldsymbol{p}$ 注入第三方音频 $\boldsymbol{x}$ 后的采样点幅值的绝对值小于音频幅值的绝对值上限 l,我们需要扰动满足:

$$|x_i+p_i|\leqslant l,\ i\in\{1,2,\cdots,n\}. \tag{4}$$

攻击可以被描述为在式(3)和式(4)约束下对式(2)的优化问题.

3.2 基于单"音频像素"扰动的攻击概述

基于单"音频像素"扰动的攻击的工作流程图如图 3 所示,可以分为 2 个子模块:1)候选语料选择;2)扰动生成.其中扰动生成包含候选点构造、候选点的迭代优化和最优点测试攻击 3 个步骤.当攻击者发起攻击,候选语料选择子模块首先从语料库中选择最有可能实现攻击的前 50 条语料送入扰动生成子模块,扰动生成模块对于当前语料生成候选点集并开始迭代,当其中任意一条语料被注入扰动后

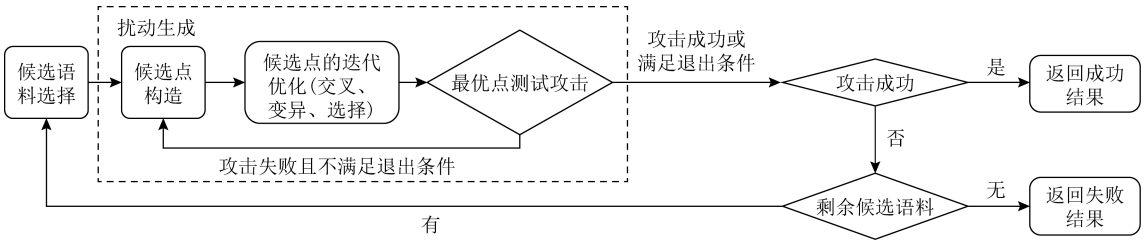


Fig. 3 The workflow of the attack base on the one-‘audio pixel’ perturbation

图 3 基于单“音频像素”扰动的攻击的工作流程图

可以伪装成受害者身份进入目标系统时,则认为攻击成功;否则,若 50 条语料全部进行尝试后,则认为攻击失败.

3.3 候选语料选择

攻击者利用一个数量巨大的语料库对目标受害者进行攻击.然而,针对特定的受害者,并不是所有语料都能实现有效的攻击,因此随机选择候选语料尝试攻击的方法将消耗大量无意义的时间.为了优化攻击的时间成本,我们利用深度神经网络中,位于决策边界附近的数据对于扰动的敏感度高的特性[14],对语料库进行筛选,通过说话人识别系统对于语料库中语料与受害者之间的相似度,把语料库中的语料按相似度从高到低排序.最后,我们选择相似度最高的前 50 条语料作为攻击的候选语料,以相似度从高到低的顺序送入扰动生成子模块.

3.4 扰动生成

扰动生成子模块采用差分进化算法,利用我们提出的音频段-音频点-扰动值多元组的候选点构造模式,解决针对基于单“音频像素”扰动的攻击差分进化算法的候选点难以被描述的问题,实现能够有效攻击目标说话人识别系统的扰动的生成.

对于每一条语料,基于单“音频像素”扰动的生成的实现可以分为 3 个步骤:

1) 以特殊的构造模式构建候选点,每个候选点是一个包含多个元素的元组.在初始化过程中,系统会根据攻击者给出的参数随机生成一个包含多个候选点的集合,称为候选集.

2) 当前候选集中的候选点称为父代候选点,对于每个父代候选点,攻击者先通过对于整体父代候选集中的最优点 b_{best},即注入语料后生成的音频与受害者相似度最高的候选点,进行变异得到一个子代候选点,再将子代和父代候选点进行交叉,获得新的候选点.最后,保留父代候选点和新的候选点之间对于攻击表现更好的候选点,完成一次变异—交叉—选择的过程.对所有候选点完成一次变异—交叉—选择后,则完成一次迭代.

3) 重复步骤 2,每次迭代结束后,当新一代的父代候选点集中的 b_{best} 注入语料后能够以受害者的身份被系统所识别时则退出迭代,且返回候选点 b_{best}.当迭代次数大于设定阈值后结束迭代,且返回攻击失败.

3.4.1 候选点构造模式

候选点是为基于单“音频像素”扰动的攻击提供优化所需要的启发信息的重要描述方法.我们结合音频数据的时间维度特性提出了基于音频段-音频点-扰动大小多元组的候选点构造模式,实现候选点的构造.为了使候选点构造方法更具有普适性,我们介绍当需注入扰动点数量为 n 的情况下,扰动点的构造模式.构造候选点 b 为

$$b=\{x_b,x_0,D_0,x_1,D_1,\cdots,x_{n-1},D_{n-1}\},\quad(5)$$

其中,x_b 为候选点插入扰动点时的基准值,x_i 为第 i 个扰动点的相对偏移,即 x_b+x_i 为实际扰动在音频中插入的绝对位置,D_i 为第 i 个扰动点位置加入扰动的大小.

同时,假设采样率为 s,“音频像素”分布时间的宽度为 Δt,初始化时约束候选点 b 的不等式为

$$\text{Max}(b)\leqslant s\Delta t,\quad(6)$$

其中,Max(b)是计算候选点中最大 x_i 的函数.初始化阶段攻击者总共生成 140 个候选点.

3.4.2 候选点的迭代优化

在初始化候选点后,我们需要对候选点进行迭代优化,从而获得能够实现攻击的候选点.迭代优化过程可以分为变异—交叉—选择 3 个阶段,是基于单“音频像素”扰动的攻击的核心过程.变异阶段使用已有候选点产生新的候选点,实现候选集的扩大;交叉阶段使用对父代和子代的交叉重组,实现通过离散杂交增加子代随机解的多样性;选择阶段淘汰对于实现目标能力较弱的候选点,使得候选集整体朝着设定目标不断进化.

1) 变异.我们采用的是 best1bin 策略进行变异.变异阶段通过父代候选点中的最优点 b_{best} 生成新的

子代候选点.值得注意的是,变异—交叉—选择的过程是针对一个父代候选点进行的,但是在 best1bin 策略的变异中并不会使用该父代候选点的元素值.子代候选点生成:

$$b_i'=b_{\text{best}}+\varepsilon(b_{r_1}-b_{r_2}),\tag{7}$$

其中,b_i' 是变异后的子代候选点,b_{best}是父代候选点中的最优点,b_{r_1} 和 b_{r_2} 是父代候选集中的任意 2 个候选点,并且 $r_1\neq r_2$,ε 是变异算子为 0.5～1 之间的随机值.$\varepsilon(b_j-b_k)$的整体称为差异向量,决定了子代变异的方向.

2) 交叉.交叉阶段首先需要定义交叉概率 cr 因子:

$$cr=\begin{cases}cr_1+(cr_{\text{u}}-cr_1)\dfrac{f_i-f_{\min}}{f_{\max}-f_{\min}}, & f_i>\bar{f},\\ cr_1, & f_i\leqslant\bar{f},\end{cases}\tag{8}$$

其中,$cr_1=0.1$ 和 $cr_{\text{u}}=0.6$ 是设定的固定值,f_i 是当前父代候选点注入语料后和受害者的相似度,$f_{\min}$,$f_{\max}$和 $\bar{f}$ 是父代候选集中所有候选点注入语料后和受害者相似度的最小值、最大值和平均值.

当确定 cr 因子后,需要对父代候选点和子代候选点的每一个维度进行交叉重组:

$$b''(j)=\begin{cases}b_i'(j), & rand(0,1)<cr,\\ b_i(j), & \text{其他},\end{cases}\tag{9}$$

其中,$b''(j)$为新生成候选点的第 j 维数值,$b_i'(j)$和 $b_i(j)$分别是子代和父代中第 j 维数值,$rand(0,1)$是随机生成的 0～1 之间的随机数.b''为完成交叉后的新候选点.

3) 选择.在完成交叉以后,算法需要对父代候选点 b 和新生成的侯选点 b''进行选择.选择依据为将 2 个候选点分别注入语料后,比较与受害者的相似度,将相似度更高的候选点保留作为下一代的父代候选点,并淘汰另一个候选点.

3.4.3 迭代设置

在实验过程中,攻击设定最大迭代次数为 1 000 次.同时,为了提高攻击效率,攻击设定了一个附加约束:每 100 次迭代计算最优候选点与受害者相似度的提升,当提升小于等于 0 时,则也提前终止迭代,并与达到最大迭代次数的情况一样返回攻击失败的提示.

4 实验设置

4.1 数据集和攻击环境

为了探究基于单“音频像素”扰动的攻击性能,我们使用 LibriSpeech[16] 公开语音数据库作为语料数据集.LibriSpeech 是一个在语音识别领域被广泛使用的语料库,包含有大约 1 000 h 的英语语音.每段语音经过去除环境噪音处理,且分割为 10s 左右的语音片段.我们将数据库分为 3 个部分,分别对说话人识别系统进行训练、注册和攻击.其中,训练集由 train-clean-100 数据包构成,注册集由 train-clean-100 数据包中随机选择的 60 个人构成,攻击集由 test-clean 数据包中随机选择的 40 个人构成.

我们搭建了一个基于 d-vector 的说话人身份认证系统作为攻击的目标系统.采用由百度提出的 Deep Speaker[12] 作为说话人特制提取器,并将余弦相似性作为说话人识别的决策模块.我们在配备 Ubuntu 16.04 和 Intel® Xeon® CPU E5-2678 v3 2.50 GHz(12 核)的服务器上进行了实验,这台服务器还配有 8 块 NVIDIA GeForce GTX 1070(8 GB)的显卡.

4.2 评价指标

我们通过欺骗说话人识别的成功率(success rate of spoofing speaker recognition, *SRoSSR*)来量化单“音频像素”扰动的攻击的性能,下文简称成功率,具体计算为:对于已注册的说话人,假如存在一个攻击集中的音频,在注入对抗扰动后能使目标声纹识别系统判断说话人为该已注册的说话人,则认为针对这一已注册的说话人的攻击成功,我们对所有注册集中说话人进行攻击,成功率为被攻击成功的说话人占所有目标说话人的比例.定义为

$$SRoSSR=\frac{S}{T},\tag{10}$$

其中,S 是被攻击成功的说话人数量,T 是所有目标说话人的数量.该评价方法也被应用于 Chen 等人[48] 的工作中,能够有效评价针对说话人识别系统的性能.

此外,为了评价说话人识别系统的性能,我们使用错误接受率(fake accept rate, *FAR*)、错误拒绝率(fake reject rate, *FRR*)和精确度(accuracy, *Acc*).其定义分别为:

$$FAR=\frac{FP}{FP+TN},\tag{11}$$

$$FRR=\frac{FN}{TP+FN},\tag{12}$$

$$Acc=\frac{TP+TN}{TP+FP+TN+FN},\tag{13}$$

其中,TP 是正确分类阳性样本的数量,TN 是正确分类阴性样本的数量,FP 是错误分类阳性样本的

数量,FN 是错误分类阴性样本的数量.在这 3 个指标的基础上,我们使用等错误率(equal error rate,EER)对说话人识别系统的整体性能进行客观的评价.等错误率定义为:当错误接受率和错误拒绝率相等时,错误接受率和错误拒绝率的值.

为了评估用户调查中语音的隐蔽性,首先考虑定量的评价指标,正如引言中介绍的增加扰动会引入宽频噪声,因此我们使用攻击注入前后时频谱的失真率作为评价隐蔽性的指标.其中,直方图相似度被用于评价时频谱的失真率,它是一种能够评价时频谱图失真率的指标,直方图相似度定义为[49]

$$S(x,y)=\frac{\boldsymbol{H}_x\cdot\boldsymbol{H}_y}{\|\boldsymbol{H}_x\|\|\boldsymbol{H}_y\|},\tag{14}$$

其中,x 和 y 分别为攻击注入前后时频谱,$\boldsymbol{H}_x$ 和 $\boldsymbol{H}_y$ 分别为 x 和 y 归一化直方图的向量.此外,$S(x,y)\in[-1,1]$,该值越小则失真率越大,反之失真率越小.除此之外,我们还设置评价指标扰动数据占比对于用户调查结果进行评价,该数值越低则攻击隐蔽性越高,定义为

$$DR=\frac{distorted\_number}{total\_number},\tag{15}$$

其中,DR 是扰动数据占比,$distorted\_number$ 是被认为注入过扰动的音频数量,$total\_number$ 是该类音频的总量.

5 实验及评估

5.1 说话人识别系统性能评估

我们首先评估搭建的说话人识别系统在 LibriSpeech 数据集上的性能,该性能能够证明第三方音频被识别成受害者是由攻击引起的,而不是因为系统本身的性能不佳导致的.

首先,我们使用 4.1 节中提到的训练集对 Deep Speaker 进行训练,接着我们使用注册集中每个人的一条语料进行注册,最后使用每个人与注册语料不同的一条语料进行测试.在认证阶段,说话人识别系统会计算输入语料的特征和注册人的特征之间的余弦相似度 sim,设定阈值为 θ,当 $sim<\theta$ 时,则认为语料不属于注册人;否则,认为语料属于注册人.由于余弦相似度的特性 $\theta\in[-1,1]$.通过改变阈值 θ,我们绘制了说话人识别系统的接收者操作特征曲线,如图 4 所示.图 4 说明说话人识别系统的等错误率为 0.05($EER=0.05$).此时设定的阈值 $\theta=0.58$,说话人识别的精确度为 98.5%.图 4 说明,我们攻击的说话人识别系统具有良好的性能,可以有效地识别说话人的身份,用来评估我们攻击的性能.

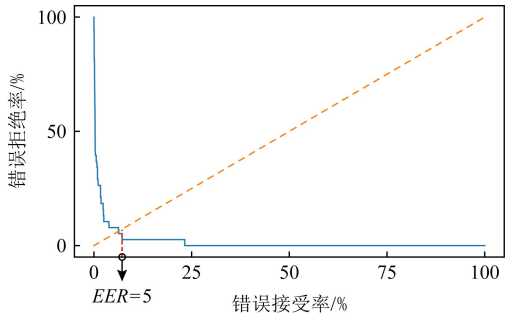


Fig. 4 The receiver operating characteristic of the speaker recognition system

图 4 说话人识别系统的接收者操作特征曲线

5.2 攻击性能和隐蔽性评估

在最优情况下,攻击者可以通过注入单“音频像素”的扰动来实现攻击.我们使用 4.1 节攻击集中的 40 个人,对已注册的 60 个人进行攻击.攻击的成功率能够到达 100%.这一结果说明,基于单“音频像素”扰动的攻击可以有效地攻击说话人识别系统.我们还对基于单“音频像素”扰动的攻击隐蔽性进行了评估并与 Chen 等人[6]的工作 FakeBob 进行了比较,FakeBob 是目前攻击说话人识别的工作中最先进的方法.结果显示我们攻击平均直方图相似度 $S=0.99$ 高于 FakeBob 的 0.94,这一结果表明基于单“音频像素”扰动的攻击在隐蔽性方面具有优越性.除此之外,下文我们还对基于单“音频像素”扰动的攻击的隐蔽性进行了用户调查进一步说明了这一优越性.

5.3 不同攻击集中人数的影响

攻击集人数对于攻击的性能也存在着影响,我们使用 4.1 节中的攻击集进行测试,把整个攻击集的人数等分成了 4 组,每组 10 个人.随机选择其中的一组作为攻击集,对所有注册的说话人进行攻击.接着,在攻击集中加入一组人,重复上述步骤,直到攻击集中人数到达 40 人.

图 5 展示了测试结果,结果表明随着攻击集中人数的上升,攻击的成功率也呈现上升趋势.当人数到达 40 人时,成功率到达了 100%.这说明攻击者可以通过加攻击集中人数来提高攻击的性能,并且由于开源的音频数据库和其他音频数据量十分巨大,攻击者可以借助这些音频数据实现高性能的攻击.由于基于单“音频像素”扰动的攻击在攻击集中人数到达 40 时成功率已经到达了 100%,使得在 40 人

的条件下探究 5.4～5.6 节的实验无法得出有意义的实验结论,因此在 5.4～5.6 节的实验中我们使用本节中划分的 30 人作为攻击集进行攻击,下文称为 30 人攻击集.攻击的基准成功率为 91.7%.

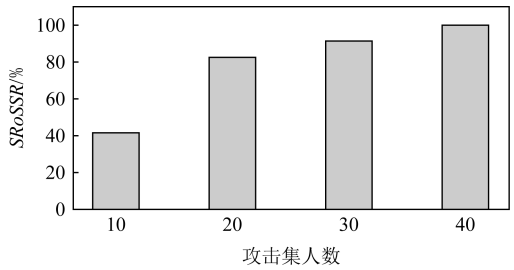


Fig. 5 Experiment of different numbers of individuals

图 5 不同人数的实验

5.4 不同扰动点分布时间宽度的影响

在实际攻击中,通过改变少量"音频像素"扰动的攻击只会产生短时的宽频噪声,同样能够保证攻击的隐蔽性,因此值得深入探究.首先考虑"音频像素"分布的时间宽度对于攻击性能的影响.我们使用 4.1 节的注册集进行注册,并使用 30 人攻击集进行攻击.在设置扰动中改动的"音频像素"数量为 3 的条件下,探究"音频像素"分布时间的宽度分别为 1 ms,10 ms,100 ms,1 000 ms 时攻击的成功率.

图 6 展示了 4 种不同的"音频像素"分布时间的宽度下攻击的成功率.如图 6 所示,当"音频像素"分布时间的宽度从 10 ms 扩大到 100 ms 时,性能有小幅度提高,并在 10 ms 之前和 100 ms 之后保持稳定状态.这一现象说明"音频像素"分布时间的宽度对于攻击性能存在小幅度的影响,且当"音频像素"分布时间的宽度到达 100 ms 后性能达到最优且趋于稳定.

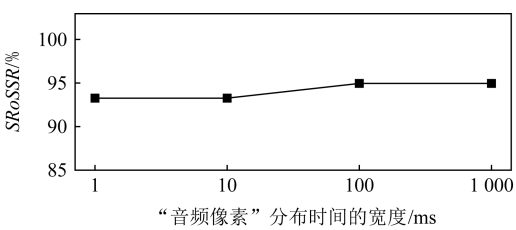


Fig. 6 Experiment of different ranges of distribution of "audio pixel"

图 6 不同"音频像素"分布时间宽度的实验

5.5 不同数量的"音频像素"的影响

我们探究了"音频像素"数量对于攻击成功率的影响.我们使用 4.1 节中的注册集进行注册,并使用 30 人攻击集进行攻击.除此之外,我们固定"音频像素"分布时间为 100 ms.结果如图 7 所示,结果说明当扰动中修改的"音频像素"数量上升时,攻击的成功率会增加,当"音频像素" 数量到达 10 时,攻击的成功率会到达 96.8%并在随后趋于稳定,然而随着"音频像素"数量的增加,攻击的隐蔽性会相对降低.这一现象说明,"音频像素"的数量和攻击的成功率之间存在权衡的关系.当"音频像素"数量上升时,攻击的成功率会增高,而攻击的隐蔽性则会相应降低.值得注意的是,当"音频像素"数量到达 10 个时,攻击的成功率已经超过了 95%,但此时扰动带来噪声的总时间也远低于已有工作汇报的结果[6].

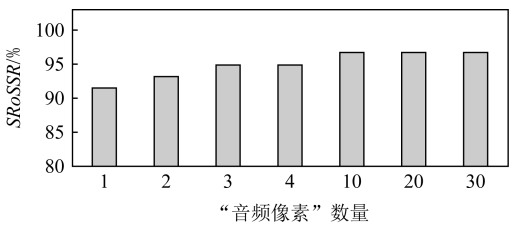


Fig. 7 Experiment of different numbers of "audio pixel"

图 7 不同"音频像素"数量的实验

5.6 不同性别的影响

性别也是影响攻击性能的关键性因素之一.为了探究性别对于攻击性能的影响,我们从 LibriSpeech 数据集中随机选择 20 名女性和 20 名男性的语料组成攻击集(与注册的说话人不重复),并从 4.1 节的注册集中选取 25 名女性和 25 名男性进行注册并作为受害者.在攻击过程中,我们使用攻击者语料库中的每一条语料尝试攻击所有注册人,并记录每一次攻击的结果.最终,对所有攻击成功的语料和受害者的性别配对关系进行了分析.结果如表 1 所示,在所有成功的配对中同性别之间的攻击数量占所有攻击成功配对数的 93%,同时跨性别的攻击数量仅占所有攻击成功配对数的 7%.这一结果表明,性别对于攻击性能有较大的影响.这是因为基于单"音频像素"扰动的攻击只改变音频中的一个或几个"音频像素",微小的改变无法改变男性和女性之间音调和音色的

Table 1 Percentage of Successful Attacks That Were Intergender and Transgender

表 1 成功进行攻击的结果中同性别间的攻击和跨性别的攻击的占比

攻击类型	占比/%
同性别间的攻击	93
跨性别间的攻击	7

天然不同,使得跨性别的攻击难以实现.但是,因为攻击者可以获得大量的语料库进行尝试,所以这并不会成为限制攻击者攻击能力的阻碍.

5.7 不同说话人识别系统的影响

现实中存在不同的说话人识别系统,基于单“音频像素”扰动的攻击能否在不同系统上都获得良好的性能也是值得探究的问题.为了探究这一问题,我们使用 GE2E[11] 和 x-vector[33] 分别搭建了 2 个说话人识别系统,使用 4.1 节中的训练集对它们分别进行了训练.与 5.1 节一样我们使用注册集中的一条语料注册,并使用一条语料进行测试.最后得到 2 个说话人识别系统的 *EER* 分别为 0.05 和 0.06.在此基础上,我们使用 4.1 节中的注册集进行注册,使用 4.1 节中的攻击集进行基于单“音频像素”扰动的攻击,即攻击中改变的“音频像素”数量为 1.结果如表 2 所示,在 2 个系统上的成功率分别为 98.3%和 95%.这一结果说明基于单“音频像素”扰动的攻击在不同说话人识别系统上都能获得良好的性能.

Table 2 *SRoSSR* on Different Speaker Recognition Platforms

表 2 不同说话人识别系统上的成功率

识别系统	*SRoSSR*/%
GE2E[11]	98.3
x-vector[33]	95

5.8 跨数据集的攻击性能

由于不同数据集采集的环境、说话人习惯和采集设备不同.为了说明跨数据集下基于单“音频像素”扰动的攻击仍然有效,我们组织了下面的实验:以 4.1 节中的注册集进行注册,并使用 TIMIT 数据集[46] 和 CMU_ARCTIC 数据集[47] 作为攻击集进行基于单“音频像素”扰动的攻击,即攻击中改变的“音频像素”数量为 1.结果表明,攻击的成功率达到了 98.3%.这一结果说明在跨数据集的情况下,基于单“音频像素”扰动的攻击仍然可以保持良好的性能.

5.9 用户调查

为了体现出我们攻击的隐蔽性,我们进行了用户调查并将基于单“音频像素”扰动的攻击和 Chen 等人[6] 的工作 FakeBob 进行了比较.

1) 用户调查的设置.为了保证志愿者不会被先验知识所影响,我们在进行用户调查时从正常(无攻击)的、FakeBob 攻击后的和我们攻击后的语音数据库中分别随机选择 10 条语音打乱后组成 30 条语音的测试集①.我们总共招募了 10 位志愿者包括 7 名男性和 3 名女性,在测试之前我们会告知志愿者测试集中包含正常(无攻击)的和攻击后的语音.接着,在安静的环境中,每位志愿者单独试听测试集中的所有语音,每条语音志愿者都可以重复试听任意次数,最后志愿者给出当前语音是正常或是被攻击的标记.

2) 用户调查结果和分析.经过用户调查中设置的实验后,我们统计了每种类型的扰动数据占比.结果如表 3 所示,正常(无攻击)的语音中的扰动数据占比为 11%;FakeBob 攻击后的语音中的扰动数据占比为 71%;我们攻击后的语音中的扰动数据占比为 38%.从表 3 可知,在志愿者被告知听到的语音中存在被攻击的语音后,志愿者对于攻击的警惕性增强,因此即使是正常(无攻击)的语音中也有 11%的语音被志愿者认为是攻击后的语音.在这种情况下,基于单“音频像素”扰动的攻击比目前最先进的 FakeBob 扰动数据占比低了 33%,这说明基于单“音频像素”扰动的攻击的隐蔽性更高,更加不易被人所察觉.

Table 3 The Distorted Rate in Different Attack Types

表 3 不同攻击类型下的扰动数据占比

攻击类型	占比/%
正常(无攻击)	11
FakeBob	71
基于单“音频像素”扰动的攻击	38

6 防御方法

由于我们提出的基于单“音频像素”扰动的攻击的成功率高且隐蔽性强,通过用户调查可以说明人工的审查很难检测出攻击;同时,由 Wang 等人[50] 提出的最先进的针对单个元素扰动的攻击的检测方法和候选检测方法的性能只能达到 9.1%和 30.1%.因此本文提出攻击会对说话人识别系统的安全造成显著的危害.为了避免这一危害,我们在本节讨论了 3 种可行的防御方法.

1) 使用去噪器的防御方法.说话人识别系统可以在预处理阶段增加针对基于单“音频像素”扰动的攻击的去噪器,由于注入单“音频像素”的扰动,所以

① 用户调查数据:https://flin.group/attack_demo

注入的扰动幅值大,去噪器将局部变化认为超过阈值的点去除,从而实现去除注入扰动的目的.这一方法在 Chen 等人[51]的工作中已经被证明能够有效地抵抗单像素的攻击,并在图像领域取得了 98.6%的防御成功率.

2) 使用重建算法的防御方法.由于声音是连续的信号,所以在音频信号中时间相近的"音频像素"存在一定相关性,通过这种相关性我们可以通过重建的方式将音频进行重建,重建后的音频将不包含与周围"音频像素"不相关的扰动点.这种方法也被 Liu 等人[52]证明了防御针对基于单像素扰动攻击的有效性.

3) 使用不同压缩方式的防御方法.说话人识别系统可以通过不同的压缩方式将音频进行压缩,实现对于单"音频像素"扰动的攻击的防御.由于不同压缩算法(如 MP3[53])的特性在压缩过程中会将原音频中的"音频像素"压缩,所以扰动点所在位置的信息会被去除,从而使得扰动无法生效.

7 讨 论

7.1 时频信息层面的攻击

在说话人识别领域,由于时频信息相较于时域信息带有更多能够描述说话人特征的信息,因此对于时频信息的分析是目前说话人识别中主流的预处理方式.由于时频信息相较于时域信息的粒度更细,因此如果针对时频信息(如 MFCC 等)进行扰动攻击,实现的说话人识别攻击能够获得更高的性能和隐蔽性.由于说话人识别系统接收的是时域的音频信息并且目前的主流时频分析方法(如 MFCC 等)是不可逆的,所以如何将攻击后的时频信息映射回时域是实现上述攻击的一大挑战.我们设想使用自编码器可以有效地解决这一问题,已有工作[54]已经证明自编码器具有良好的降维能力,这一能力能够有效地帮助完成高维的时频信息映射到低维的时域信息的任务,从而实现时频信息层面的攻击.

7.2 通过空气传播的扰动注入方式

除了本文假设的攻击者通过用户接口注入扰动的方式以外,通过音频在空气中传播的扰动注入方式也是常见的注入方式之一.声波在空气中传播会引入音频的失真(如衰减、环境噪声和多径效应)导致最终的结果无法达到预期.针对这一问题,目前已有工作给出了一些解决方案,如在有先验知识的条件下整合房间脉冲响应附加到扰动上,使得生成的音频能够在空气中传播而不损失扰动信息[55].除此之外,Li 等人的工作[8]提出了通过在对抗学习过程加入符合环境失真条件的随机扰动,使得生成的扰动能够稳定地在空气中传播而避免失真带来的性能损失.因此,在基于单"音频像素"扰动的攻击中可以在扰动生成子模块将符合环境失真条件的随机扰动附加到扰动之上,使得攻击能够生成不受空气传播中失真影响的单"音频像素"扰动,实现通过空气传播的扰动注入方式.

7.3 减少对于说话人识别系统的访问

在攻击过程中,对于说话人识别系统的大量访问会降低攻击的效率并且增加攻击被管理人员发现的可能性.替代模型是解决这一问题最先进的方式.攻击者通过对于目标系统的少量访问,可以在本地建立目标说话人识别系统的替代模型,从而大幅度减少对于目标系统的访问次数.Papernot 等人[56]最早在图像对抗攻击中应用了这种方法并取得了良好的性能.Chen 等人[48]将该方法应用于攻击说话人识别系统中,在不降低攻击性能的前提下,实现了对于访问次数的优化,说明了替代模型在攻击说话人识别领域中的可用性和高效性.因此,单"音频像素"攻击可以借助替代模型,从而实现减少对于说话人识别系统的访问的目的.

8 总 结

本文提出了一种新颖的基于单"音频像素"扰动的说话人识别隐蔽攻击,获得了相较以往攻击更高的隐蔽性.利用差分进化算法不依赖梯度的特性,克服了已有工作中存在局部最优的问题,提出了基于音频段-音频点-扰动的构造模式,解决了针对我们的攻击差分进化算法的候选点难以被描述的问题,实现了具有高性能高隐蔽性的攻击.这种攻击在由百度提出的 Deep Speaker 上获得了 100%的成功率,同时攻击对主流的说话人识别都有良好的攻击性能表现.我们还探究了不同因素对于攻击性能的影响,并且进行了用户调查说明了攻击的隐蔽性.最后,我们提出了几种针对攻击有效的防御手段,进一步增强了说话人识别的安全性.

作者贡献声明:沈轶杰提出基于单"音频像素"扰动的攻击方案,设计总体实验,优化算法,整体文章撰写;李良澄实现基于单"音频像素"扰动的攻击,设计候选点构造模式;刘子威搭建攻击测试平台,收集实验所需数据;刘天天尝试迭代优化中不同参数对于性能的影响;罗浩绘制文章内图片,对文章进行修订;沈汀在修改过程中,对于实验给出指导性建

议,提出使用直方图相似度对失真率进行衡量,从而解决对于攻击性能量化的目标,并对文章整体进行了修订;林峰指导实验的总体设计和文章写作指导;任奎指导文章写作,对于克服文中挑战给出方向性建议.

参 考 文 献

[1] Wang Tao, Wang Guozhong, Zhu Linlin. Design and implementation of a smart door lock system based on voiceprint recognition [J]. Electronic Measurement Technology, 2019, 42(3): 107-111 (in Chinese)
(王涛, 王国中, 朱林林. 一种基于声纹识别的智能门锁系统设计与实现[J]. 电子测量技术, 2019, 42(3): 107-111)

[2] Wei Lianfang. Research on application of “Internet+” based voiceprint recognition technology in criminal case investigation [J]. Modern Electronics Technique, 2020, 43(7): 34-38
(魏莲芳. 基于“互联网+”的声纹识别技术在刑事案件侦破中的应用研究[J]. 现代电子技术, 2020, 43(7): 34-38)

[3] Ava K. Forget About Siri and Alexa —When It Comes to Voice Identification, the “NSA Reigns Supreme”[EB/OL]. (2018-01-19)[2021-06-08]. https://theintercept.com/2018/01/19/voice-recognition-technology-nsa/

[4] Kreuk F, Adi Y, Cisse M, et al. Fooling end-to-end speaker verification with adversarial examples [C] //Proc of IEEE Int Conf on Acoustics, Speech and Signal Processing. Piscataway, NJ: IEEE, 2018: 1962-1966

[5] Gong Yuan, Poellabauer C. Crafting adversarial examples for speech paralinguistics applications [C/OL] //Proc of Int Conf on Learning Reresentation. [2021-06-08]. https://arxiv.org/abs/1711.03280

[6] Chen Guangke, Chen Sen, Fan Lingling, et al. Who is real Bob? adversarial attacks on speaker recognition systems [C/OL] //Proc of IEEE Symp on Security and Privacy. Piscataway, NJ: IEEE, 2021 [2021-06-01]. https://www.ieee-security.org/TC/SP2021/program-papers.html

[7] Li Xu, Zhong Jinghua, Wu Xixin, et al. Adversarial attacks on GMM i-vector based speaker verification systems [C] //Proc of IEEE Int Conf on Acoustics, Speech and Signal Processing. Piscataway, NJ: IEEE, 2020: 6579-6583

[8] Li Zhuohang, Wu Yi, Liu Jian, et al. AdvPulse: Universal, synchronization-free, and targeted audio adversarial attacks via subsecond perturbations [C] //Proc of the ACM SIGSAC Conf on Computer and Communications Security. New York: ACM, 2020: 1121-1134

[9] Azirani A A, Jeannes R L B, Faucon G. Optimizing speech enhancement by exploiting masking properties of the human ear [C] //Proc of IEEE Int Conf on Acoustics, Speech, and Signal Processing. Piscataway, NJ: IEEE, 1995: 6579-6583

[10] Snyder D, Garcia-Romero D, Sell G, et al. X-Vectors: Robust DNN embeddings for speaker recognition [C] //Proc of IEEE Int Conf on Acoustics, Speech and Signal Processing. Piscataway, NJ: IEEE, 2018: 5239-5333

[11] Wan Li, Wang Quan, Papir A, et al. Generalized end-to-end loss for speaker verification [C] //Proc of IEEE Int Conf on Acoustics, Speech and Signal Processing. Piscataway, NJ: IEEE, 2018: 4879-4883

[12] Li Chao, Ma Xiaokong, Jiang Bing, et al. Deep speaker: An end-to-end neural speaker embedding system [J]. arXiv preprint arXiv:1705.02304, 2017, 650

[13] Su Jiawei, Vargas D V, Sakurai K. One pixel attack for fooling deep neural networks [J]. IEEE Transactions on Evolutionary Computation, 2019, 23(5): 828-841

[14] Fawzi A, Moosavi-Dezfooli S M, Frossard P, et al. Classification regions of deep neural networks [J]. arXiv preprint arXiv:1705.09552, 2017

[15] Liu Bo, Wang Ling, Jin Yihui. Advances in differential evolution [J]. Control and Decision, 2007, 22(7): 721-729 (in Chinese)
(刘波, 王凌, 金以慧. 差分进化算法研究进展[J]. 控制与决策, 2007, 22(7): 721-729)

[16] Panayotov V, Chen Guoguo, Povey D, et al. Librispeech: An ASR corpus based on public domain audio books [C] //Proc of IEEE Int Conf on Acoustics, Speech and Signal Processing. Piscataway, NJ: IEEE, 2015: 5206-5210

[17] Povey D, Ghoshal A, Boulianne G, et al. The Kaldi speech recognition toolkit [C/OL] //Proc of IEEE Workshop on Automatic Speech Recognition and Understanding. Piscataway, NJ: IEEE Signal Processing Society, 2011 [2021-07-26]. https://infoscience.epfl.ch/record/192584

[18] Sadjadi S O, Slaney M, Heck L. MSR identity toolbox [EB/OL]. Seattle, WA, USA: Microsoft, 2013 [2020-06-08]. http://research.microsoft.com/

[19] Bonastre J F, Wils F, Meignier S. ALIZE, a free toolkit for speaker recognition [C] //Proc of IEEE Int Conf on Acoustics, Speech, and Signal Processing. Piscataway, NJ: IEEE, 2005: 737-740

[20] Larcher A, Meignier S, Lee K A. SIDEKIT Documentation [M/OL]. (2019-01-22)[2021-06-08]. https://projets-lium.univ-lemans.fr/sidekit/

[21] Li Bo, Sainath T N, Narayanan A, et al. Acoustic modeling for Google home [C] //Proc of the Annual Conf of the Int Speech Communication Association. Grenoble, France: ISCA, 2017: 399-403

[22] Xiamen TalentedSoft Co. Talentedsoft [EB/OL]. [2021-06-08]. http://www.talentedsoft.com

[23] Fujimura H, Ding Ning, Hayakawa D, et al. Simultaneous flexible keyword detection and text-dependent speaker recognition for low-resource devices [C] //Proc of Int Conf on Pattern Recognition Applications and Methods. Setubal, Portugal: Scitepress, 2020: 297-307

[24] Wang Wenchao, Zhang Yike, Xu Ji, et al. Multiple temporal scales based speaker embeddings learning for text-dependent speaker recognition [C] //Proc of IEEE Int Conf on Acoustics, Speech and Signal Processing. Piscataway, NJ: IEEE, 2019: 6311-6315

[25] Jahangir R, Teh Y W, Memon N A, et al. Text-independent speaker identification through feature fusion and deep neural network [J]. IEEE Access, 2020, 8: 32187-32202

[26] Zhao Fei, Li Hao, Zhang Xueliang. A robust text-independent speaker verification method based on speech separation and deep speaker [C] //Proc of IEEE Int Conf on Acoustics, Speech and Signal Processing. Piscataway, NJ: IEEE, 2019: 6101-6105

[27] Sohn J, Kim N S, Sung W. A statistical model-based voice activity detection [J]. IEEE Signal Processing Letters, 1999, 6(1): 1-3

[28] Sigurdsson S, Petersen K B, Lehn-Schiøler T. Mel frequency cepstral coefficients: An evaluation of robustness of MP3 encoded music [C/OL] //Proc of Int Society for Music Information Retrieval. 2006: 286 - 289. [2020-07-26]. https://archives.ismir.net/ismir2006/paper/000080.pdf

[29] McLaughlin J, Reynolds D A, Gleason T. A study of computation speed-ups of the GMM-UBM speaker recognition system [C/OL] //Proc of the European Conf on Speech Communication and Technology. 1999: 1215-1218. [2020-07-26]. https://www.isca-speech.org/archive/eurospeech_1999/e99_1215.html

[30] Dehak N, Dehak R, Kenny P, et al. Support vector machines versus fast scoring in the low-dimensional total variability space for speaker verification [C] //Proc of the Annual Conf of the Int Speech Communication Association. Grenoble, France: ISCA, 2009: 1559-1562

[31] Variani E, Lei Xin, McDermott E, et al. Deep neural networks for small footprint text-dependent speaker verification [C] //Proc of IEEE Int Conf on Acoustics, Speech and Signal Processing. Piscataway, NJ: IEEE, 2014: 4052-4056

[32] Shi Ziqiang, Liu Liu, Lin Huibin, et al. Joint learning of J-vector extractor and joint bayesian model for text dependent speaker verification [C] //Proc of the Annual Conf of the Int Speech Communication Association. Grenoble, France: ISCA, 2018: 1076-1080

[33] Snyder D, Garcia-Romero D, Sell G, et al. X-vectors: Robustdnn embeddings for speaker recognition [C] //Proc of IEEE Int Conf on Acoustics, Speech and Signal Processing. Piscataway, NJ: IEEE, 2018: 5329-5333

[34] rahman C F A R, Wang Quan, Moreno I L, et al. Attention-based models for text-dependent speaker verification [C] // Proc of IEEE Int Conf on Acoustics, Speech and Signal Processing. Piscataway, NJ: IEEE, 2018: 5359-5363

[35] Guo Gongde, Wang Hui, Bell D, et al. KNN model-based approach in classification [C] //Proc of on the Move to Meaningful Internet Systems. Berlin: Springer, 2003: 986-996

[36] Hearst M A, Dumais S T, Osuna E, et al. Support vector machines [J]. IEEE Intelligent Systems and Their Applications, 1998, 13(4): 18-28

[37] Prince S J D, Elder J H. Probabilistic linear discriminant analysis for inferences about identity [C] //Proc of the Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2007: 1-8

[38] Szegedy C, Zaremba W, Sutskever I, et al. Intriguing properties of neural networks [J]. arXiv preprint arXiv:1312.6199, 2013

[39] Cisse M, Adi Y, Neverova N, et al. Houdini: Fooling deep structured prediction models [J]. arXiv preprint arXiv:1707.05373, 2017

[40] Graves A, Fernández S, Gomez F, et al. Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks [C] //Proc of the Int Conf on Machine Learning. New York: ACM, 2006: 369-376

[41] Carlini N, Wagner D. Audio adversarial examples: Targeted attacks on speech-to-text [C] //Proc of IEEE Security and Privacy Workshops. Piscataway, NJ: IEEE, 2018: 1-7

[42] Liu Xiaolei, Wan Kun, Ding Yufei, et al. Weighted-sampling audio adversarial example attack [C] //Proc of the AAAI Conf on Artificial Intelligence. Menlo Park, CA: AAAI, 2020: 4908-4915

[43] Liu Junhong, Lampinen J. A fuzzy adaptive differential evolution algorithm [C] //Proc of the Soft Computing. Berlin: Springer, 2005: 448-462

[44] Qin A K, Suganthan P N. Self-adaptive differential evolution algorithm for numerical optimization [C] //Proc of IEEE Congress on Evolutionary Computation. Piscataway, NJ: IEEE, 2005: 1785-1791

[45] Oord A, Dieleman S, Zen H, et al. Wavenet: A generative model for raw audio [J]. arXiv preprint arXiv:1609.03499, 2016

[46] Bhowmick A, Biswas A, Chandra M. Performance evaluation of psycho-acoustically motivated front-end compensator for TIMIT phone recognition [J]. Pattern Analysis and Applications, 2020, 23(2): 527-529

[47] Kominek J, Black A W, Ver V. CMU_ARCTIC databases for speech synthesis [EB/OL]. Language Technologies Institute at Carnegie Mellon University. Pittsburgh, PA, 2003[2021-06-08]. http://festvox.org/cmu_arctic/index.html

[48] Chen Yuxuan, Yuan Xuejing, Zhang Jiangshan, et al. Devil's whisper: A general approach for physical adversarial attacks against commercial black-box speech recognition devices [C] //Proc of USENIX Security Symp. Berkeley, CA: USENIX Association, 2020: 2667-2684

[49] Ma Yu, Gu Xiaodong, Wang Yuanyuan. Histogram similarity measure using variable bin size distance [J]. Computer Vision and Image Understanding, 2010, 114(8): 981-989

[50] Wang Peng, Cai Zhipeng, Kim D, et al. Detection mechanisms of one-pixel attack [J]. Wireless Communications and Mobile Computing, 2021, 2021: 1-8

[51] Chen Dong, Xu Ruqiao, Han Bo. Patch selection denoiser: An effective approach defending against one-pixel attacks [C] //Proc of Int Conf on Neural Information Processing. Berlin: Springer, 2019: 286-296

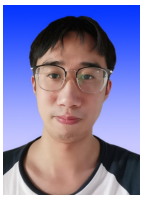
[52] Liu Ziyuan, Wang P S, Hsiao S C, et al. Defense against N-pixel attacks based on image reconstruction [C] //Proc of the Int Workshop on Security in Blockchain and Cloud Computing. New York: ACM, 2020: 3-7

[53] Brandenburg K. MP3 and AAC explained [C/OL] //Proc of AES 17th Int Conf: High-Quality Audio Coding. Audio Engineering Society, 1999 [2020-07-26]. https://www.aes.org/e-lib/browse.cfm?elib=8079

[54] Hinton G E, Salakhutdinov R R. Reducing the dimensionality of data with neural networks [J]. Science, 2006, 313(5786): 504-507

[55] Abdullah H, Garcia W, Peeters C, et al. Practical hidden voice attacks against speech and speaker recognition systems [C/OL] //Proc of the 26th Annual Network and Distributed System Security Symp. Washington, DC: ISOC, 2019[2020-07-26]. https://dx.doi.org/10.14722/ndss.2019.23362

[56] Papernot N, McDaniel P, Goodfellow I, et al. Practical black-box attacks against machine learning [C] //Proc of the 2017 ACM on Asia Conf on Computer and Communications Security. New York: ACM, 2017: 506-519

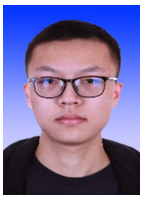


Shen Yijie, born in 1996. Master candidate. His main research interests include IoT security, biometric security, and artificial intelligence.

沈轶杰,1996 年生.硕士研究生.主要研究方向为物联网安全、生物特征安全和人工智能.

Li Liangcheng, born in 1999. Undergraduate student. Her main research interests include IoT security, biometric security.

李良澄,1999 年生.本科生.主要研究方向为物联网安全、生物特征安全.



Liu Ziwei, born in 1998. PhD candidate. His main research interests include IoT security, wireless security.

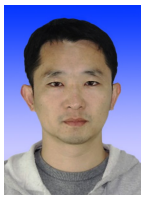
刘子威,1998 年生.博士研究生.主要研究方向为物联网安全、无线安全.

Liu Tiantian, born in 1997. PhD candidate. Her main research interest is IoT security.

刘天天,1997 年生.博士研究生.主要研究方向为物联网安全.

Luo Hao, born in 1995. Master candidate. His main research interests include artificial intelligence and computer vision.

罗　浩,1995 年生.硕士研究生.主要研究方向为人工智能、计算机视觉.



Shen Ting, born in 1984. Received his BSc degree from Naval University of Engineering. Engineer of Zhejiang Dong'an Testing Technology Co., Ltd.. He is a director of IEEE PES Power System Communication and Network Security Technical Committee, Certified Information Security Professional. His main research interests include trusted computing, industrial information security, and security testing of information system.

沈　汀,1984 年生.毕业于海军工程大学.浙江东安检测技术有限公司工程师.IEEE PES 电力系统通信与网络安全技术委员会理事,注册信息安全人员.主要研究方向为可信计算、工业信息安全、信息系统安全检测技术.

Lin Feng, born in 1984. Professor in the Institute of Cyberspace Research and the College of Computer Science and Technology of Zhejiang University. Received his PhD degree in electrical and computer engineering from Tennessee Technological University, USA. IEEE senior member. His main research interests include IoT security, mobile sensing and computing, intelligent sensing, biometric authentication, and artificial intelligence.

林　峰,1984 年生.浙江大学网络空间安全研究中心和计算机科学与技术学院“百人计划”研究员.美国田纳西理工大学电子与计算机工程博士.IEEE 高级会员.主要研究方向为物联网安全、移动传感与计算、智能感知、生物识别身份认证、人工智能.

Ren Kui, born in 1978. University Professor of Zhejiang University. ACM fellow and IEEE fellow. His main research interests include data security, artificial intelligence security, IoT security, authentication and privacy protection.

任　奎,1978 年生.浙江大学求是讲席教授.ACM 会士,IEEE 会士.主要研究方向为数据安全、人工智能安全、物联网安全、认证与隐私保护.