基于鲁棒对抗防御边界的语音伪造方法识别

张 强[1,2],张雄伟[1*],孙 蒙[1],杨吉斌[1]
(1. 陆军工程大学指挥控制工程学院,江苏南京 210007;2. 陆军工程大学研究生院,江苏南京 210007)

摘 要: 深度伪造语音的反欺骗是生成式人工智能安全领域的一项重要技术.除了对真实语音和伪造语音进行二元分类外,语音伪造方法识别正在成为可解释的反欺骗策略的重要组成部分.但为了逃避对语音伪造方法的识别,攻击者很有可能利用对抗样本攻击技术,在伪造语音中加入人耳无法感知的对抗扰动,来降低语音伪造方法识别(Speech Forgery Method Recognition,SFMR)模型的准确性.针对SFMR所面临的对抗样本攻击问题,从防御者的角度出发,提出了对抗防御边界概念.基于此,使用泰勒分析技术,理论分析了网络随机性和决策边界距离对模型对抗鲁棒性的影响,并提出了基于鲁棒对抗防御边界(Robust Adversarial Defense Boundary,RADB)的SFMR算法.该算法采用随机变换(Random Transform, RT)和决策边界距离正则化(Decision Boundary Distance Regularization,DBDR)两个模块实现鲁棒对抗防御.RT模块通过模拟真实世界场景中伪造语音可能受到的干扰,在训练和推理时,均对输入语音进行随机组合变换,利用随机性提高对抗鲁棒性.DBDR模块引入决策边界距离正则化损失函数,鼓励模型提高对抗鲁棒性上限,降低模型的类别预测关于对抗扰动的敏感性.在典型SFMR数据集,即中文伪造音频检测(Chinese Fake Audio Detection,CFAD)数据集和2019年自动说话人验证欺骗与对策挑战赛(2019 Automatic Speaker Verification spoofing and countermeasures challenge,ASVspoof2019)数据集上的实验结果表明,在对抗攻击条件下,与现有先进基线方法相比,所提算法能够将SFMR准确率分别提高5.63%、5.95%,至93.98%、91.71%.

关键词: 语音伪造方法识别;对抗防御边界;随机变换;决策边界距离正则化;对抗样本
基金项目: 国家自然科学基金(No.62371469,No.62071484);江苏省自然科学基金(No.BK20180080)
中图分类号: TP391.4 **文献标识码:** A **文章编号:** 0372-2112(2025)06-2022-16
电子学报URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20241128

Robust Adversarial Defense Boundary-Based Speech Forgery Method Recognition

ZHANG Qiang[1,2], ZHANG Xiong-wei[1*], SUN Meng[1], YANG Ji-bin[1]
(1. *School of Command and Control Engineering, Army Engineering University, Nanjing, Jiangsu* 210007, *China*;
2. *Graduate School, Army Engineering University, Nanjing, Jiangsu* 210007, *China*)

Abstract: Anti-spoofing of deeply forged speech is an important technique in the field of generative artificial intelligence (AI) security. In addition to binary classification of real and forged speech, speech forgery method recognition is becoming an important part of interpretable anti-spoofing strategies. To evade the recognition of the speech forgery method, attackers are likely to utilize the adversarial attack technique to degrade the accuracy of the speech forgery method recognition (SFMR) model by adding adversarial perturbations that are imperceptible to the human ear into the forged speech. To address this problem of adversarial attack faced by SFMR, the concept of adversarial defense boundary is proposed from the defender's point of view. Based on this, the effect of network randomness and decision boundary distance on model adversarial robustness is theoretically analyzed using Taylor analysis techniques, and the robust adversarial defense boundary(RADB)-based SFMR algorithm is proposed. Two modules, random transform (RT) and decision boundary distance regularization (DBDR), are adopted by the algorithm to realize robust adversarial defense. The RT module improves the adversarial robustness by simulating the possible interference of forged speech in real-world scenarios, and randomly transforming the input speech during both training and inference to take advantage of the randomness. The DBDR module introduces the decision boundary distance regularization loss function to encourage the model to increase the upper bound of adversari-

收稿日期:2024-12-16;修回日期:2025-05-10;责任编辑:覃怀银
*通讯作者:张雄伟

al robustness and reduce the sensitivity of the model's class prediction regarding the adversarial perturbation. Experimental results on typical SFMR datasets, i.e., Chinese fake audio detection(CFAD) and 2019 automatic speaker verification spoofing and countermeasures challenge (ASVspoof2019), show that compared with existing state-of-the-art baseline methods, the proposed algorithm is able to improve the SFMR accuracy under adversarial attacks by 5.63% and 5.95% to 93.98% and 91.71%, respectively.

Key words: speech forgery method recognition; adversarial defense boundary; random transform; decision boundary distance regularization; adversarial example

Foundation Item(s): National Natural Science Foundation of China (No.62371469, No.62071484); Natural Science Foundation of Jiangsu Province (No.BK20180080)

1 引言

深度伪造(Deepfakes)[1]的出现和快速发展,增加了社交媒体中的内容安全风险. 针对语音、图像Deepfakes的防御成为信息安全的重要研究方向[2]. 近年来,针对语音Deepfakes的防御已从语音真伪检测发展到伪造语音归因. 伪造语音归因旨在识别伪造语音的属性并判定其出处,其中,属性主要指伪造语音的方法、特性,出处是指伪造语音的来源. 语音伪造方法识别(Speech Forgery Method Recognition,SFMR)是伪造语音归因的关键组成部分,它能够从Deepfakes中识别出相应伪造方法,近年来备受关注[3~7]. 一方面,SFMR是提高语音反欺骗策略可解释性的关键环节,它可以给出判定某个语音是伪造的原因. 另一方面,SFMR能够针对不同类别的伪造语音学习增强表征,学习到的表征可进一步用于提高伪造语音检测的性能[8].

基于深度学习的模式识别技术因其良好的分类性能已逐渐成为现有SFMR工作采用的主流方法. 但现有SFMR工作[3~13]在训练模型时,并未考虑对抗样本攻击等因素的影响. 而攻击者可以利用深度学习模型易受对抗样本影响的特点,在生成的伪造语音中加入人耳无法感知的微小对抗性扰动,生成对抗样本来攻击SFMR模型,使其产生误判,这相当于对伪造方法进行伪装,避免其被正确识别. 虽然提升深度学习模型的对抗鲁棒性一直是人工智能领域的一个关键问题,但初步实验表明,现有SFMR模型的对抗鲁棒性较弱,对经过扰动伪装的伪造语音的伪造方法识别准确率较低. 为此,亟须面向对抗样本攻击条件下的伪造语音,开展SFMR算法研究,提升SFMR模型的对抗鲁棒性.

对抗鲁棒语音伪造方法识别(Adversarially Robust SFMR,AR-SFMR)是指在存在对抗样本攻击的情况下,模型能够正确识别语音伪造方法. 现有对抗防御领域的典型方法可以用于解决AR-SFMR问题,但这些方法大多只考虑了数据增强或模型增强,存在以下缺陷:现有的数据增强技术要么依赖大量对抗样本作为训练数据[14~17],要么对输入采用了不可微分的随机变换[18~20]. 对抗样本加入训练集,会影响算法在干净样本上的性能,并且对抗样本生成会消耗大量的时间和算力. 在输入上引入不可微分的随机变换,会影响神经网络中的梯度反向传播,从而导致梯度混淆,产生虚假的鲁棒现象. 现有的模型增强技术要么集成多个模型共同完成特定任务,要么使用随机神经网络(Stochastic Neural Networks,SNN)将可学习噪声引入模型权重,要么将基于梯度的正则化项作为损失函数的一部分来训练模型. 基于模型集成的方法[21~23]会增加模型大小;引入可学习噪声的方法[24,25]依赖噪声学习策略的精心设计;基于梯度正则化的方法[26~28]缺乏梯度正则化有利于增强对抗鲁棒性的理论分析.

为了提高SFMR模型的对抗鲁棒性,本文首先提出对抗防御边界的概念,然后使用泰勒分析技术,研究影响SFMR模型对抗鲁棒性的因素. 在此基础上,提出一种名为基于鲁棒对抗防御边界(Robust Adversarial Defense Boundary,RADB)的SFMR(RADB-SFMR)算法. 首先,针对模型的输入,构建随机变换(Random Transform,RT)模块. 在该模块中,RADB-SFMR算法通过模拟现实世界中各种场景对语音的影响,设计了不同的随机变换方案,这些变换都是可微分的. 在训练和推理过程中,该模块中的各种变换会经过随机组合并作用于输入,增加输入的随机性. 其次,针对模型本身,设计了决策边界距离正则化(Decision Boundary Distance Regularization,DBDR)模块. 在该模块中,RADB-SFMR算法设计了DBDR损失函数,其可与分类损失函数联合,指导模型学习到更高的对抗鲁棒性上限,从而降低模型预测对扰动的敏感性. 实验结果表明,在AR-SFMR任务中,所提算法优于最近提出的其他方法. 此外,对算法结果的可视化和消融研究,进一步验证了所提算法的优越性. 本文提出的AR-SFMR任务和所提算法可为语音反欺骗领域的研究作出贡献.

本文的贡献可归纳如下. 首先,首次定义了AR-SFMR任务,为后续有关SFMR算法的对抗鲁棒性研究奠定了基础. 其次,本文首次从对抗防御边界的角度分析了SFMR模型的对抗鲁棒性. 在分析的基础上,提出了DBDR模块和RT模块. DBDR模块鼓励模型学习到更高的对抗鲁棒性上限,从而降低模型的类别预测对

输入的敏感性. RT模块在训练和推理过程中,将各种变换经过随机组合并作用于输入,这有助于增加输入的随机性,增加模型的对抗鲁棒性.

2 相关工作

2.1 语音伪造方法识别

近年来,学术界对深度伪造语音检测进行了广泛的研究[29,30]. 在此基础上,SFMR可以进一步提高检测的可解释性. Müller等人[3]认为伪造方法总会在伪造语音中留下独特的痕迹,这些痕迹可以被提取出来作为模型识别伪造方法的特征. Neri等人[4]提出了一个双分支模型,每个分支提取特定的特征来表示伪造方法. Deng等人[5]提出了一种结合对比学习的SFMR算法. 该算法使用ResNet18模型[31]作为神经网络模型,以对数梅尔尺度的短时傅里叶变换谱图块作为输入,使用Softmax损失函数和有监督对比损失函数[32]的组合来监督网络训练. Xie等人[6]将神经音频编解码方法视为一种新型语音伪造方法,并探讨了自监督特征与不同分类器相结合的方法在SFMR任务上的性能. 此外,田野等人[7]提出了一种基于预训练模型的识别算法. 该算法基于预训练模型输出的嵌入特征构建流形空间,通过度量不同伪造方法对应语音数据流形空间的测地线距离,进行语音伪造方法识别. Salvi等人[9]将基于残差网络(Residual Network, ResNet)的检测模型中的真假二分类输出层替换为多类别输出层,将其用于伪造方法识别. Yan等人[10]将ResNet与不同特征相结合,并探究了其在SFMR任务上的性能. Yadav等人[11]和Bartusiak等人[12]利用不同的变换器作为分类器的骨干,并在SFMR任务上对其进行微调,以提高识别性能. Shim等人[8]和Klein等人[13]将伪造方法的不同模块建模为不同属性,并探索了ResNet和图神经网络识别不同属性的性能. 然而,上述研究工作忽略了对抗样本攻击对伪造方法识别性能的影响,而攻击者很可能利用对抗样本攻击技术,通过在样本上加入人耳无法感知的微小扰动,混淆识别模型对伪造方法类别的判断,从而逃避法律制裁.

2.2 对抗样本攻击

对抗样本是将干净样本与精心设计的对抗性扰动δ相加的结果,这种扰动会误导深度学习模型的判断[33]. 对抗性扰动δ通常被限制在以O为中心、ε为半径的p范数超球范围内. 为保证对抗样本不被人类察觉,需要对扰动施加约束ε. 由于在ε相同的情况下,$p=\infty$利于攻击者生成攻击性更强的对抗样本. 因此,本文主要考虑$p=\infty$条件下的攻击. 要找到攻击能力最强的对抗扰动,需要解决以下优化问题:

$$\delta'=\underset{\delta:\|\delta\|_p\leqslant\varepsilon}{\arg\max}\,L\left(g(x+\delta),y\right) \tag{1}$$

其中,$g(\cdot)$是目标模型,L是损失函数,x是输入样本,y是x的真实标签. 在SFMR任务中,x代表伪造语音,x_{adv}代表伪造语音x的对抗样本,$g(\cdot)$是识别模型,即概率分类器. 一般来说,基于梯度的对抗攻击,如快速梯度符号法(Fast Gradient Sign Method, FGSM)[33]、投影梯度下降(Projected Gradient Descent, PGD)[34]、基本迭代法(Basic Iterative Method, BIM)[35]和抖动(Jitter)[36],均通过解决式(1)中定义的优化问题,以获得攻击性能最强的扰动. FGSM是一种高效的一步对抗样本攻击方法. BIM、PGD和Jitter是FGSM的迭代版本.

然而,对SNN来说,其在输入或网络本身引入了随机性. 因此,对SNN求解最优扰动δ'时,使用的输入或目标模型与使用δ'实施攻击时针对的输入或目标模型是不同的. 通过式(1)单次求解得到的最优扰动δ'对于实施攻击时的目标模型来说并不是最优的. 因此,Athalye等人[37]提出了变换期望(Expectation Over Transformation, EOT)算法,该算法可与上述基于梯度的攻击方法结合使用. 具体地,当使用上述攻击(如PGD)生成对抗样本时,在更新对抗样本的每一次迭代中,EOT都会使用蒙特卡罗算法采样多个输入或目标模型权重,并使用基于采样的输入或模型权重计算出的梯度的平均值作为最终梯度. 在这种情况下,EOT采用的损失函数为

$$\mathrm{EOT}(\delta)=\frac{1}{n}\sum_{j=1}^{n}L\left(g_{\phi_j}\left(t_{\theta_j}(x+\delta)\right),y\right) \tag{2}$$

此时,式(1)中的优化问题转换为

$$\begin{aligned}\delta'&=\underset{\delta:\|\delta\|_p\leqslant\varepsilon}{\arg\max}\,\mathrm{EOT}(\delta)\\&=\underset{\delta:\|\delta\|_p\leqslant\varepsilon}{\arg\max}\,\frac{1}{n}\sum_{j=1}^{n}L\left(g_{\phi_j}\left(t_{\theta_j}(x+\delta)\right),y\right)\end{aligned} \tag{3}$$

其中,ϕ_j和θ_j表示第j次采样得到的模型权重和对输入的变换参数,$t(\cdot)$表示关于输入的随机变换,n表示采样次数. 由式(3)可以看出,对SNN的EOT攻击可视为基于替代模型的转移攻击,因为攻击者无法访问所有SNN,而是使用采样得到的n个模型对应损失的平均值替代目标模型的损失,来优化求解对抗扰动,生成对抗样本.

2.3 对抗性防御

对抗性防御的目的是使目标模型在受到带有限制性扰动的对抗样本攻击时,仍能够正确识别对抗样本. 目前,对抗性防御策略主要包括数据增强和模型增强两大类型.

数据增强包括对抗训练和随机变换. 对抗训练[14]是指不仅使用干净样本,还使用对抗样本来训练模型.

比如,Zhang等人[17]通过实验发现基于梯度的对抗攻击方法倾向于在高频添加对抗扰动,并提出了一种频率选择对抗训练(Frequency-Selective Adversarial Training,F-SAT)方法. F-SAT方法在特定高频范围生成对抗扰动,并与干净时频特征相加,得到对抗样本,进行对抗训练. 该类方法将干净样本周围尽可能多的对抗样本映射到正确的标签上,以抵御对抗样本攻击. 然而,该类方法依赖于大量的对抗样本作为训练数据,通常需要耗费大量时间来生成对抗样本,这导致其对干净样本的泛化能力较弱[38],对黑盒攻击[39]和训练期间未见的其他类型攻击[40]的鲁棒性较差. 随机变换是指在训练和推理过程中利用经过随机变换的样本作为输入,这有助于提高对抗鲁棒性. 比如,Xie等人[18]随机对输入图像进行缩放和填充操作. Zhang等人[19]在输入图像中随机添加高斯噪声. Raff等人[20]提出了一个图像随机变换集合,并从中随机选择一系列变换来处理输入,以增加输入的随机性. 然而,所提出的一些变换是不可微分的,这导致他们的方法受到梯度混淆的影响.

模型增强包括噪声注入、模型增广和权重正则化. 噪声注入是指在训练和推理过程中向网络权重或各层输出的特征中添加噪声. 与对输入进行随机变换类似,它通过引入随机性来提高对抗鲁棒性. 参数化噪声注入(Parametric Noise Injection,PNI)[41]将具有固定参数的随机噪声注入模型权重. 此外,自适应噪声注入(Adaptive Noise Injection,AdaNI)[24]根据各层输出特征的重要性,将参数可学习的随机噪声添加到相应特征中. 与上述在全局添加噪声的方法不同,Liu等人[25]通过阈值法将激活值较小的神经元定义为非关键神经元,并在这些神经元的输出中注入可学习的噪声,以破坏扰动,增强防御. 然而,Jeddi等人[42]指出带有可学习参数的噪声会随着训练的进程趋于零,从而削弱防御性能. 因此,他们提出了Learn2Perturb(L2P)方法,为噪声设置了一个较低的阈值. 在训练过程中,低于该阈值的噪声会被重置为该阈值,其防御性能取决于所选的阈值. 权重-协方差对齐(Weight-Covariance Alignment,WCA)[43]从减小模型对干净样本和对抗样本的误分类概率之差的角度出发,通过在模型输出层中注入多元高斯噪声,增强模型的对抗鲁棒性. 模型增广指的是训练多个模型共同完成同一个任务. 例如,通过正交性实现多样性(DIversity via Orthogonality,DIO)[21]设计了一个包含多个分类器的模型,并鼓励分类器拥有相互正交的权重,其中一个分类器的输出用于预测类别标签. 这种方法无疑增加了模型的大小和计算复杂度. 基于逆扰动融合生成对抗网络[22]以对抗样本为输入,通过生成网络构造逆扰动,从而重构原始样本作为分类模型的输入,提高模型的分类性能. 然而,该方法除了需要训练深度神经网络分类器之外,还需要通过生成器与判别器之间的对抗训练,使生成器具备构造逆扰动的能力,这无疑增加了总体模型大小和计算复杂度. Meng等人[23]结合回归学习和对比学习技术额外训练一个深度嵌入提取器,使该提取器能够从经过时频掩蔽的时频特征样本中捕捉到对伪造语音检测任务重要的判别性嵌入表示,并将提取的嵌入表示作为后端检测器的输入,以提高伪造语音检测任务的对抗鲁棒性. 权重正则化是指最小化输出相对于输入的梯度,使该梯度的方向多样化或者趋于攻击性能更弱的通用梯度方向,使模型对输入空间的扰动不敏感. 输入梯度正则化(Input Gradient Regularization,IGR)[26]直接使用梯度的模长作为损失函数的正则化项,以鼓励模型学习较小的梯度模长. 通过不同方向进行防御(Defense Through Diverse Directions,DTDD)[27]将梯度的方差和均值作为正则化项,鼓励模型同时学习具有较小模长和多样化方向的梯度. 统一梯度(Unifying Gradients,UniG)[28]在模型的输出层之前插入一个Hadamard乘积模块,在约束经该模块变换前后的特征的输出逻辑值之差的同时,将不同样本的梯度之差作为正则化项,鼓励该模块在尽量不影响原模型输出的情况下,使得针对不同输入的梯度趋于均匀分布,以弱化求解的对抗扰动的攻击性. Lee等人[44]利用不同的正则化项,比如最小化同一个样本的不同变换版本对应的输出对输入的梯度之间的余弦相似性,来鼓励梯度方向多样化.

3　模型对抗鲁棒性分析

首先开展深度网络模型的对抗鲁棒性分析. 不失一般性,基于同时在模型权重和输入上引入随机性的SNN来分析模型的对抗鲁棒性. 此时,固定神经网络可以看作是一种特殊的SNN,即在网络权重或输入上引入随机性均为零的SNN. 对于类别标签为y的语音样本x和SNN模型$g(\cdot)$,攻击者在求解式(3)获得最优扰动δ'时,使用蒙特卡罗算法采样得到的模型和输入分别为$g_A(\cdot)$和x_A;在使用求解得到的扰动δ'实施攻击时,使用蒙特卡罗算法采样得到的模型和输入分别为$g_I(\cdot)$和x_I. 其中,$x_A=t_A(x)$,$x_I=t_I(x)$,$t_A(\cdot)$和$t_I(\cdot)$表示蒙特卡罗采样得到的对输入的变换参数. 在没有对抗样本攻击的情况下,训练好的模型$g(\cdot)$的任意蒙特卡罗采样模型都能正确地对x的任意蒙特卡罗采样变换版本进行正确分类,即对$\forall k\in[1,C]$,下式成立:

$$\begin{aligned}&\arg\max_k\left(g_A(x_A)_k|k\in[1,C]\right)\\&=\arg\max_k\left(g_I(x_I)_k|k\in[1,C]\right)\\&=y\end{aligned}\tag{4}$$

其中,$g_A(x_A)_k$表示x_A属于第k类的逻辑值,$g_I(x_I)_k$表示

x_1属于第k类的逻辑值，C是类别数.

将求解式(3)得到的最优扰动定义为δ_A. 那么在使用求解得到的扰动δ_A实施攻击时，攻击者会构建一个对抗样本$x+\delta_A$来攻击$g_1(\cdot)$. 该对抗样本在输入模型$g_1(\cdot)$前，会被随机变换为$t_1(x+\delta_A)$. 为了便于分析，将$t_1(x+\delta_A)$改写为

$$t_1(x+\delta_A)=t_1(x)+\delta_1=x_1+\delta_1 \tag{5}$$

如果对$\forall k\in[1,C],k\neq y$，式(6)成立，则说明模型$g(\cdot)$仍然能够正确识别出$x_1+\delta_1$的标签，即$g(\cdot)$是对抗鲁棒的.

$$g_1(x_1+\delta_1)_y-g_1(x_1+\delta_1)_k>0 \tag{6}$$

式(6)给出了模型$g(\cdot)$关于类别y的对抗鲁棒条件. 基于此，本文定义模型$g(\cdot)$的对抗防御边界为

$$\Pi^{\mathrm{Def}}=\{\Pi_{yk}^{\mathrm{Def}}|k=1,2,\cdots,C;\ y=1,2,\cdots,C;\ y\neq k\} \tag{7}$$

其中，Π_{yk}^{Def}为第y类相对于第k类的对抗防御边界，其公式为

$$\begin{aligned}&\Pi_{yk}^{\mathrm{Def}}:\ g_1(x_1+\delta_1)_y-g_1(x_1+\delta_1)_k=0,\\&\forall k\in[1,C],k\neq y\end{aligned} \tag{8}$$

当对于x的任意蒙特卡罗采样变换版本$x_1,x_1+\delta_1$都位于边界Π_{yk}^{Def}大于0的一侧时，模型$g(\cdot)$关于类别y是对抗鲁棒的. 为分析模型的对抗鲁棒性，用泰勒公式在点x_1处展开式(6)的左侧，得到

$$\begin{aligned}&g_1(x_1+\delta_1)_y-g_1(x_1+\delta_1)_k\\&=g_1(x_1)_y-g_1(x_1)_k-\left(\nabla_{x_1}g_1(x_1)_k-\nabla_{x_1}g_1(x_1)_y\right)\cdot\delta_1-o(\delta_1)\\&=g_1(x_1)_y-g_1(x_1)_k-o(\delta_1)\\&\quad-\left\|\nabla_{x_1}\left(g_1(x_1)_k-g_1(x_1)_y\right)\right\|\cdot\|\delta_1\|\cdot\cos\Theta>0\end{aligned} \tag{9}$$

其中，$\|\cdot\|$表示欧几里得范数算子，$o(\delta_1)$是δ_1的高阶无穷小，∇表示梯度，Θ是δ_1与$\nabla x_1\left(g_1(x_1)_k-g_1(x_1)_y\right)$之间的夹角. 式(9)经过恒等变换可得

$$\begin{aligned}&\frac{g_1(x_1)_y-g_1(x_1)_k}{\left\|\nabla_{x_1}\left(g_1(x_1)_k-g_1(x_1)_y\right)\right\|\cdot\|\delta_1\|}\\&-\frac{o(\delta_1)}{\left\|\nabla_{x_1}\left(g_1(x_1)_k-g_1(x_1)_y\right)\right\|\cdot\|\delta_1\|}>\cos\Theta\end{aligned} \tag{10}$$

从式(10)可以看出，Θ是影响模型对抗鲁棒性的因素之一. 由于$\arg\max_k\left(g_1(x_1)_k|k\in[1,C]\right)=y$，式(10)左侧的第一项大于0，左侧第二项的分子$o(\delta_1)$是无穷小项，远小于$\|\delta_1\|$. 在关于对抗样本的实验设计中，$\|\delta_1\|$相对于$\|g_1(x_1)_y-g_1(x_1)_k\|$来说非常小. 因此第二项不影响左侧总体的符号. 可以认为式(10)左侧的符号由左侧第一项的符号决定，因此，式(10)的左侧大于0.

由于$\cos\Theta\in[-1,1]$，下面区分$\Theta=0^\circ$、$0^\circ<\Theta<90^\circ$和$90^\circ\leqslant\Theta\leqslant180^\circ$三种情况，分别分析影响模型对抗鲁棒性的因素.

(1)当$90^\circ\leqslant\Theta\leqslant180^\circ$时，$\cos\Theta\leqslant0$. 由上述分析可知，式(10)的左侧大于0，由于$\cos\Theta\leqslant0$，因此式(10)的左侧$g_1(x_1)_y-g_1(x_1)_k>0$. 可以推得不等式(6)成立，即当$90^\circ\leqslant\Theta\leqslant180^\circ$时，$g_1(\cdot)$可以正确预测对抗样本的类别，模型是对抗鲁棒的.

(2)当$0^\circ<\Theta<90^\circ$时，$1>\cos\Theta>0$. 要保证式(10)成立，可以将式(10)的条件加强为

$$\begin{aligned}&\frac{g_1(x_1)_y-g_1(x_1)_k}{\left\|\nabla_{x_1}\left(g_1(x_1)_k-g_1(x_1)_y\right)\right\|\cdot\|\delta_1\|}\\&-\frac{o(\delta_1)}{\left\|\nabla_{x_1}\left(g_1(x_1)_k-g_1(x_1)_y\right)\right\|\cdot\|\delta_1\|}>1\end{aligned} \tag{11}$$

式(11)经过恒等变换可得

$$\frac{g_1(x_1)_y-g_1(x_1)_k}{\left\|\nabla_{x_1}g_1(x_1)_k-\nabla_{x_1}g_1(x_1)_y\right\|}>\|\delta_1\|+\frac{o(\delta_1)}{\left\|\nabla_{x_1}g_1(x_1)_k-\nabla_{x_1}g_1(x_1)_y\right\|} \tag{12}$$

(3)当$\Theta=0^\circ$时，$\cos\Theta=1$，即$\nabla x_1g_1(x_1)_k-\nabla x_1g_1(x_1)_y$和$\delta_1$的方向是一致的，这包含以下特殊情况：$g(\cdot)$是特殊的SNN，即固定神经网络. 当$g(\cdot)$为固定神经网络时，攻击者在通过式(1)求解最优扰动以及使用该扰动实施攻击时，针对的输入和模型是一致的. 此时，$g_A(\cdot)=g_1(\cdot)=g(\cdot)$，$x_A=x_1=x$，$\delta_A=\delta_1=\delta'$. 求解式(1)得到的最优扰动位于损失函数对于输入的梯度方向上，即$\nabla x_1g_1(x_1)_k-\nabla x_1g_1(x_1)_y$和$\delta_1$的方向是一致的，$\cos\Theta=1$. 此时，式(10)等价于式(11)，经过恒等变换就可以得到式(12).

对于第(2)和第(3)种情况，如果对$\forall k\in[1,C],k\neq y$，式(12)成立，则式(10)成立，同样可以推导出式(6)成立. 将式(12)的左侧定义为Ω_k，即

$$\Omega_k=\frac{g_1(x_1)_y-g_1(x_1)_k}{\|\nabla_{x_1}g_1(x_1)_k-\nabla_{x_1}g_1(x_1)_y\|} \tag{13}$$

观察式(13)，由于$g_1(x_1)_y-g_1(x_1)_k>0$，Ω_k在形式上相当于样本x_1与第k类和第y类的决策边界Π_{ky}：$g_1(x_1)_y-g_1(x_1)_k=0$之间的距离. Ω_k可以看作是模型在第k维上的对抗鲁棒性的上限(Upper Bound on Adversarial Robustness，UBAR). Ω_k越大，意味着式(12)成立的可能性越大，模型在第k维上的对抗鲁棒性越强. 因此，Ω_k是影响模型对抗鲁棒性的一个因素. 综合考虑模型在各个维度的对抗鲁棒性，模型的对抗鲁棒性上限可以定义为

$$\Omega_{\min}=\min\{\Omega_k|k=1,2,\cdots,C\} \tag{14}$$

可以看出，$\Omega_{\min}$是模型在各个维度上的对抗鲁棒性上限的最小值. 因此，增加$\Omega_{\min}$，将有利于增加每一个Ω_k，从而增加式(12)成立的可能性，这将更有利于增强模型的对抗鲁棒性.

值得注意的是，对于固定神经网络，网络对抗鲁棒的条件为式(11)成立，即式(11)左侧大于1. 而对于

SNN,网络对抗鲁棒的条件为式(10)成立,即式(10)左侧大于 $\cos\Theta$. 由于式(10)和式(11)的左侧相同,并且 $\cos\Theta\leqslant 1$,因此,在其他因素相同的条件下,SNN的对抗鲁棒条件更容易得到满足,即SNN具有更强的对抗鲁棒性.

现有的防御方法大多从增加随机性、权重正则化等角度出发增强模型的对抗鲁棒性. 从对抗防御边界的角度出发,可以对这些方法进行统一解释. 例如,部分研究工作通过对输入进行随机变换[18~20]或在网络中加入随机噪声[24,25]来增强对抗鲁棒性,这是由于引入随机性,使网络成为SNN,有助于增强模型的对抗鲁棒性;部分研究工作[26,27]通过输出相对于输入的梯度正则化项来增强对抗鲁棒性,这是由于这有利于减小式(13)中 Ω_k 的分子,即有利于增加 Ω_k,从而增加模型的对抗鲁棒性上限 $\Omega_{\min}$.

4　所提算法

基于上述关于模型对抗鲁棒性的分析,本节从对抗防御边界的角度出发,提出了RADB-SFMR算法. 该算法从数据增强和模型增强两个方面,分别设计了RT模块和关于 $\Omega_{\min}$ 的正则化模块,增强模型的 $\Omega_{\min}$. 本节首先介绍RADB-SFMR算法的总体框架,而后详细介绍该算法中的RT模块和正则化模块. 由于该正则化模块约束了样本与决策类边界间的距离(Decision Boundary Distance,DBD),因此将其命名为DBDR模块. 最后,对所提算法的流程进行了总结梳理.

4.1　总体框架

RADB-SFMR算法总体框架如图1所示. 算法以语音样本 x 为输入,通过RT模块构建语音样本的随机组合变换版本,通过特征提取器 $f(\cdot)$ 输出变换语音的特征表示,通过分类器 $h(\cdot)$ 输出各个语音样本对应的类别逻辑值. 基于输出的类别逻辑值,计算RADB-SFMR算法的总体损失,并通过反向传播算法更新特征提取器 $f(\cdot)$ 和分类器 $h(\cdot)$ 的参数,提升模型的对抗鲁棒性上限.

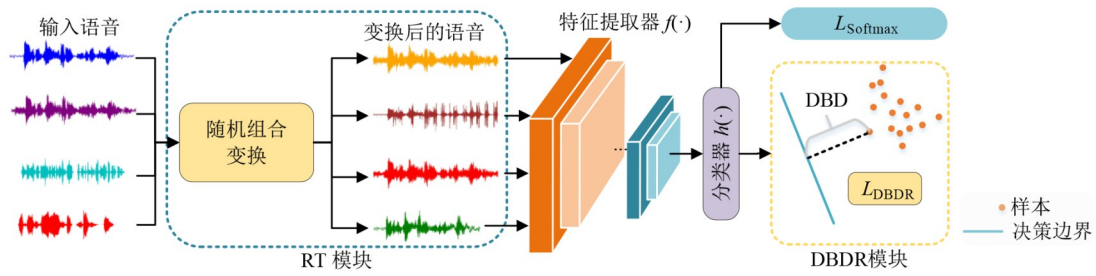


图1　基于RADB的SFMR算法总体框图

作为算法的特征提取器,$f(\cdot)$ 通常采用深度神经网络结构学习输入语音的特征表示. 作为算法的分类器,$h(\cdot)$ 由全连接层构成,输出样本属于各伪造类别对应的逻辑值. 作为RADB-SFMR算法的关键,RADB有三个模块,分别为RT模块、DBDR模块、分类模块. 其中,分类模块是模式识别任务中常用的模块,主要用于计算分类损失. RT模块主要用于模拟真实世界中各种场景对语音的干扰,对输入语音进行随机组合变换. DBDR模块主要用于设计关于模型对抗鲁棒性上限 $\Omega_{\min}$ 的正则化项,鼓励模型学习更大的 $\Omega_{\min}$. 其中,分类模块、DBDR模块对应的损失函数分别为Softmax损失函数 L_{Softmax}、决策边界距离正则化损失函数 L_{DBDR}. 该算法使用的总体损失函数 L_{RADB} 为上述两个损失函数的组合,公式为

$$L_{\text{RADB}}=L_{\text{Softmax}}+\alpha L_{\text{DBDR}} \tag{15}$$

其中,参数 α 用于平衡两种损失函数的重要性.

分类模块的损失函数 L_{Softmax} 主要用于计算分类损失,监督模型学习各个类的分布,实现正确分类,其定义为

$$\begin{aligned}L_{\text{Softmax}}&=-\sum_{k=1}^{C}I\{k=y\}\log p_k=-\log p_y\\&=-\log\frac{\exp\left(g(x)_y\right)}{\sum\limits_{k=1,k\neq y}^{C}\exp\left(g(x)_k\right)+\exp\left(g(x)_y\right)}\\&=\log\left(1+\sum_{k=1,k\neq y}^{C}\exp\left(g(x)_k-g(x)_y\right)\right)\end{aligned} \tag{16}$$

其中,C 表示样本类别数,y 是样本 x 的真实标签,$I\{\}$ 是指示函数,p_k 表示样本属于第 k 类的置信度概率,$g(x)$ 是样本属于各个类别的逻辑值向量,$g(x)_k$ 表示样本属于第 k 类的逻辑值. 从式(16)可以看出,最小化 L_{Softmax} 只鼓励最小化 $g(x)_k-g(x)_y$. 对于被正确分类的样本 x,$g(x)_k-g(x)_y<0$ 成立. 此时,最小化 $g(x)_k-g(x)_y$ 等价于最大化 $g(x)_y-g(x)_k$,但这并不鼓励最小化式(13)中的 $\left\|\nabla x_I\left(g_I(x_I)_k\right)-\nabla x_I\left(g_I(x_I)_y\right)\right\|$,这意味着只最小化 L_{Softmax} 并不是增强对抗鲁棒性的理想选择.

4.2　随机变换模块

现实世界中,作为伪造方法识别对象的语音通常

会受到混响、背景噪声、嘈杂人声等因素干扰. 因此，RT模块模拟上述干扰，实现了以下4种变换，并通过这些变换的随机组合来增加输入的随机性. 这4种变换分别为混响效应、环境噪声叠加、背景音乐叠加、嘈杂人声叠加. 对于语音样本x，从这4种变换中随机选择Q种变换并依次作用于x，得到经随机组合变换的样本. 这一过程可以形式化描述为

$$\begin{aligned} &t(x;\ \psi)=t_{\psi^{(Q)}}\circ t_{\psi^{(Q-1)}}\circ\cdots\circ t_{\psi^{(1)}}(x),\\ &\psi=\{\psi_{(q)}|q=1,2,\cdots,Q\}\in \mathrm{Perm}(4,Q) \end{aligned} \tag{17}$$

其中，$\mathrm{Perm}(4,Q)$是从这4种变换中随机选择Q种变换的所有排列组合的集合，$\psi=\{\psi_{(q)}|q=1,2,\cdots,Q\}\in\mathrm{Perm}(4,Q)$表示这$Q$种变换的一种排列组合方案. 每种变换$\psi_{(q)}=\{\alpha_{(q)},\beta_{(q)},\gamma_{(q)}\}$是一个三元随机参数组，由$\alpha_{(q)}$、$\beta_{(q)}$、$\gamma_{(q)}$这3个随机变量来描述. $\alpha_{(q)}$表示变换的类型，从这4种变换类型中采用无放回的方式随机选择. $\beta_{(q)}$表示是否使用$\alpha_{(q)}$，它遵循参数为u的伯努利分布，即使用$\alpha_{(q)}$的概率为u. 依据熵最大化原理[45]，u设为0.5. 此时，伯努利分布的方差最大，随机性最强，最有可能使模型见到种类最为丰富的随机变换组合. 此外，这种按概率决定是否使用变换的方式，不仅能使模型见过经不同变换的语音，也能使模型见过未经变换的语音. $\gamma_{(q)}$表示变换的强度，例如添加噪声的信噪比(Signal-to-Noise Ratio，SNR)，它遵循给定参数的均匀分布. 这4种变换的具体实现如下：

(1)混响效应模拟伪造语音在房间内传播的情景，通过伪造语音与房间脉冲响应的卷积来实现. 对于伪造语音样本x，从房间脉冲响应数据集[46]中随机选取一个房间脉冲响应音频，与x进行卷积运算，得到混响变换后的伪造语音.

(2)环境噪声叠加模拟伪造语音受到环境噪声干扰的情景. 对于x，从MUSAN数据集[47]的NOISE子集中随机提取一段环境噪声音频，然后对其进行裁剪或零填充，使其长度与x相同，并根据随机信噪比添加到x中. 信噪比在0~15 dB之间随机选取.

(3)嘈杂人声叠加模拟其他说话人的真实声音对伪造语音的干扰. 对于x，从MUSAN数据集的SPEECH子集中随机选取z个语音样本. z在[3, 4, 5, 6, 7, 8]中随机取值. 每个语音样本均通过裁剪或零填充将其调整为与x相同的长度，信噪比在13~20 dB之间随机选取.

(4)背景音乐叠加模拟伪造语音受音乐干扰的情景. 对于x，从MUSAN数据集的MUSIC子集中随机选取一段音频，通过裁剪或零填充将其调整为与x相同的长度，然后添加到x中，其信噪比在5~15 dB之间随机选取.

可以看出，混响效应相当于对语音进行卷积操作，其余三种相当于在语音上叠加噪声. 这些变换均是可微分的.

4.3 决策边界距离正则化模块

假设语音伪造方法识别数据集有C个类别. 依据本文第3节中的分析，对于攻击者来说，要成功攻击分类器，则至少存在某个维度k，扰动δ的模长大于Ω_k. 从防御者的角度来看，假设攻击者的最大扰动幅度是ε，防御者希望在每个维度上的Ω_k均大于ε，即模型的对抗鲁棒性上限大于ε，可形式化描述为$\Omega_{\min}=\min_k\{\Omega_k|k=1,2,\cdots,C\}>\varepsilon$，从而使模型能够防御对抗样本攻击. 因此，DBDR模块设计了关于$\Omega_{\min}$的正则化损失函数，其公式为

$$L_{\mathrm{DBDR}}=\max\{0,\zeta-\Omega_{\min}\}=\max\left\{0,\zeta-\min_k\left\{\frac{g_I(x_I)_y-g_I(x_I)_k}{\left\|\nabla_{x_I}g_I(x_I)_k-\nabla_{x_I}g_I(x_I)_y\right\|}\middle|k\in[1,C],k\neq y\right\}\right\} \tag{18}$$

其中，ζ是阈值参数. 最小化式(18)鼓励模型学习到的$\Omega_{\min}$满足$\Omega_{\min}>\zeta$.

4.4 算法流程

所提算法的流程如算法1所示. 该算法的具体步骤为首先初始化特征提取器$f(\cdot)$和分类器$h(\cdot)$的参数，设在验证集上的最佳准确率ACC_b为0. 将一批数量为$|B|$的语音样本送入RT模块，输出$|B|$个经随机组合变换的语音样本. 提取这$|B|$个语音样本的对数梅尔谱(Logarithmic Mel Spectrogram，LMS)特征. 以这些特征为输入，依次通过特征提取器$f(\cdot)$和分类器$h(\cdot)$，输出样本属于各个伪造方法类别的逻辑值. 依据式(15)，计算总体损失. 接着使用优化器Opt，基于梯度反向传播算法更新$f(\cdot)$和$h(\cdot)$的参数. 如果学习率更新策略Sch被触发，则调整学习率lr. 每轮训练结束后，计算所提算法在验证集上的准确率ACC，如果$\mathrm{ACC}>\mathrm{ACC}_b$，则将$\mathrm{ACC}_b$的值更新为ACC，并保存特征提取器$f(\cdot)$和分类器$h(\cdot)$的参数. 当达到最大训练轮数Epoch时，则返回特征提取器$f(\cdot)$和分类器$h(\cdot)$的参数，停止训练.

由于所提算法在网络的输入上引入了随机性，因此在评估算法性能时，所提算法将依据一个样本的多个变换版本，来预测样本对应的类别. 具体地，对于一个待测样本，通过RT模块，输出n_1个经过转换的语音样本，提取这n_1个语音样本的LMS特征，记为$\mathrm{LMS}_i, i\in[1,n_1]$. 使用训练得到的特征提取器$f(\cdot)$和分类器$h(\cdot)$输出这$n_1$个样本属于各个类的Softmax概率向量$\boldsymbol{p}_i$，其表达式为$\boldsymbol{p}_i=\mathrm{Softmax}\left(h\left(f(\mathrm{LMS}_i)\right)\right), i=1,2,\cdots,n_1$. 基于该样本对应的这$n_1$个概率向量的均值向量，预测该样本的伪造方法类别. 如果没有特殊强调，实验中设置

算法1　基于鲁棒对抗防御边界的SFMR算法

输入：训练集D_{tr}、验证集D_{dev}、最大训练轮数Epoch、批大小$|B|$、优化器Opt、学习率lr、学习率更新策略Sch

输出：特征提取器$f(\cdot)$和分类器$h(\cdot)$

初始化$f(\cdot)$和$h(\cdot)$的参数

FOR epoch = 1:Epoch DO

　FOR B in D_{tr} DO

　　使用RT模块得到$|B|$个语音样本的随机组合变换样本

　　提取已生成的随机组合变换语音样本的LMS特征

　　利用特征提取器$f(\cdot)$和分类器$h(\cdot)$,计算各语音样本属于各个类别的逻辑值

　　依据式(18),计算DBDR模块损失L_{DBDR}

　　依据式(15),计算所提算法损失L_{RADB}

　　基于L_{RADB},使用Opt,依据梯度反向传播算法更新$f(\cdot)$和$h(\cdot)$的参数

　END FOR

　计算所提算法在验证集上的准确率ACC

　IF ACC > ACC_b THEN

　　$\mathrm{ACC}_b \leftarrow$ ACC

　　保存$f(\cdot)$和$h(\cdot)$的参数

　END IF

　IF 触发Sch THEN

　　按照Sch更新lr

　END IF

END FOR

返回$f(\cdot)$和$h(\cdot)$的参数

$n_1 = 10$. 基于对每个样本的预测结果,就可以计算该算法在评估集上的各种评估指标,比如准确率等.

5　仿真实验及结果分析

本节首先介绍实验设置,然后比较所提算法与其他基线方法的性能. 此外,通过直观的可视化和消融实验,解释所提算法的优势.

5.1　实验设置

5.1.1　数据预处理

采用中文伪造音频检测(Chinese Fake Audio Detection,CFAD)数据集[48]和2019年自动说话人验证欺骗与对策挑战赛(2019 Automatic Speaker Verification spoofing and countermeasures challenge, ASVspoof2019)数据集[49]作为实验数据集.

CFAD数据集是第一个公开可得并可用于语音伪造方法识别的中文数据集,共有训练集、验证集和评估集3个子集,包含9种伪造方法,定义为F00~F08,其中真实语音(用F00表示)被视为特殊类别. 其余8个类别是基于传统声码器或神经声码器的语音合成方法. 训练集、验证集和评估集均包含这9个类别. 训练集中,每个类别包含3 200个语音样本. 验证集中,每个类别包含1 200个语音样本. 评估集中,每个类别包含3 500个语音样本.

ASVspoof2019数据集是伪造语音反欺骗领域广泛使用的数据集,共涵盖20种伪造方法,其中包括常用的伪造方法Tacotron2[50]和WaveNet[51]. 实验中,将其划分为训练集、验证集和评估集3个子集,每个子集均包含上述20种伪造方法,对应20个类别,定义为A00~A19. 其中真实语音(用A00表示)作为一个特殊类别. 训练集中,每个类别包含2 948个语音样本. 验证集中,每个类别包含983个语音样本. 评估集中,每个类别包含983个语音样本.

所有语音信号均重采样为16 kHz,单通道. 在RT模块中,对每个语音样本,依次进行Q次随机变换,得到经变换后的输入样本. 然后通过剪切或零填充将其调整为5.12 s. 如果没有特殊强调,$Q = 3$. 然后,使用25 ms的汉宁窗和10 ms的步长将其切成帧. 使用短时傅里叶变换、梅尔滤波和对数操作,将每个样本转换为LMS特征. 用于梅尔滤波操作的梅尔频段数为128. 因此,LMS的高度(即梅尔频段数)为128,宽度(即帧数)为512. 语音变换函数都基于Pytorch 2.1.2的Torchaudio库实现,并集成到网络模型中. 由于RT模块中的所有变换函数均是可微分的,并且这些变换函数的代码实现也确保了语音变换函数在模型层面的可微分性,使得损失函数关于输入的梯度可以通过变换函数作用于输入,从而避免了梯度掩蔽现象.

5.1.2　目标模型训练设置和算法评估

考虑到ResNet系列和Inception系列模型在学习特征表示方面的良好性能,在CFAD数据集、ASVspoof2019数据集上,作为被攻击和被防御对象的SFMR模型分别使用在ImageNet[52]数据集上预训练的ResNet-50模型[31]、Inception-v3模型[53]作为特征提取器,并以全连接层作为分类器. 对于所提算法和所有基线方法,SFMR模型均训练50轮,批大小$|B|$为128,并使用Adam优化器[54]. 初始学习率为2×10^{-4},每10轮下降为原来的20%. 权重衰减为1×10^{-5}. 如果没有特殊强调,DBDR模块中的参数$\zeta = 100$,$\alpha = 0.1$. 所有实验均在配备型号为英伟达4090的图形处理器(Graphics Processing Unit,GPU)和Pytorch深度学习平台的工作站上进行. 在验证集上获得最高准确率的模型作为最终模型,用于评估各种方法的性能. 所有实验使用准确率衡量模型的识别性能.

5.1.3　防御方法设置

为了全面评估算法的有效性,本文区分对抗训练、噪声注入、模型集成、权重正则化4种防御策略,选择9种对抗防御方法作为对比基线方法. 除了使用相同

的SFMR模型外,基线方法的其他设置参照其对应的原始工作,具体设置如下.

对抗训练,采用F-SAT[17]以及对应的无频率选择对抗训练(None Frequency-Selective Adversarial Training,NF-SAT)作为对比基线.其中,NF-SAT采用由PGD生成的对抗样本进行对抗训练.PGD生成对抗样本时,最大扰动约束设为0.03,迭代次数设为30,单次扰动步长设为0.003.与NF-SAT唯一不同的是,F-SAT在生成对抗扰动时,会通过掩码操作仅保留4~8 kHz频率范围内的扰动.对抗训练时,采用Softmax损失函数计算干净样本、对抗样本对应的损失,对模型进行训练.

噪声注入,采用PNI[41]、L2P[42]、WCA[43]作为对比基线方法.其中,PNI为每一层权重添加高斯噪声的均值为零,方差设为该层权重的方差,随网络权重一同更新;PNI中添加噪声的比例因子初始化为0.25,并通过Softmax损失函数计算损失,更新网络权重和比例因子.在PNI的基础上,L2P[42]为添加的噪声设置了下限阈值0.01,小于该阈值的噪声均会被置为该阈值.WCA[43]为最后一层全连接层添加多元高斯噪声.多元高斯噪声的均值固定为0,协方差矩阵初始化为元素均为1的下三角阵;使用Softmax损失函数和基于协方差矩阵的正则化项共同训练网络,更新网络权重和协方差矩阵;其中,正则化项的系数为–1.

模型集成,采用DIO[21]作为对比基线.DIO为模型添加的分类头数量设为10,每个分类头均由一层全连接层组成,全连接层的节点数设为类别数;采用Softmax损失函数、正交约束以及距离约束正则化项的组合作为最终损失函数,训练神经网络;其中,两个正则化项的系数均为0.1.

权重正则化,采用IGR[26]、DTDD[27]、UniG[28]作为对比基线方法.这三种方法均使用Softmax损失函数作为分类损失.不同的是,IGR采用Softmax损失函数对输入的梯度的模长作为正则化项.正则化项的系数设为1 000.DTDD采用输出对输入梯度的均值和方差作为正则化项.均值正则化项和方差正则化项的系数分别设为10、20.UniG插入的Hardarmad乘积模块的形状与目标模型输出的特征形状保持一致,其权重参数均初始化为1;约束模型关于原始特征、经该模块变换的特征的输出逻辑值之差的阈值参数设置为0.1;该模块参数的学习率设置为0.5.

5.1.4 攻击方法设置

在白盒攻击和黑盒攻击两种情况下,对所提算法的防御性能进行了评估.在白盒攻击场景中,采用五种攻击方法:FGSM、BIM、Jitter、PGD和CW[55].在黑盒攻击场景中,采用两种攻击方法:单像素(One-pixel)攻击[56]、Square攻击[57]方法.对于FGSM、BIM、Jitter和PGD,最大扰动约束设为0.03.对于BIM、Jitter和PGD,迭代次数设为30,单次扰动步长设为0.003.对于Jitter,缩放因子设为2,添加的高斯噪声方差设为0.01.对于CW,学习率设置为0.001,迭代次数为100,置信度参数设置为3.One-pixel攻击通过向输入中的单个数据点添加扰动,来生成对抗样本实施攻击;其最大扰动约束设为0.03,种群大小设置为400,最大迭代次数设置为75,交叉变异概率设为0.7,变异常数在0.5~1.0之间均匀选择.基于查询的攻击方法Square,通过向局部方形区域添加扰动来实施攻击;其最大扰动约束设为0.03,查询次数设为100.

由于所提算法和部分基线方法[41~43]在模型输入或权重中引入了随机性,被防御的目标模型属于随机神经网络,因此,实验中还采用FGSM、BIM、PGD和Jitter的EOT版本,即EOT-FGSM、EOT-BIM、EOT-PGD、EOT-Jitter来评估模型的防御性能.依据实验观察,在基于EOT的攻击方法中,蒙特卡罗采样次数达到15次时,攻击方法的性能增益开始减小,达到20次时,攻击方法的性能逐渐收敛.因此,在基于EOT的攻击方法中,蒙特卡罗采样次数设置为30次.实验中,所有攻击方法都是基于Torchattack库实现的.

5.2 与基线方法的对比实验结果

表1和表2分别显示了在CFAD数据集和AS-Vspoof2019数据集上,在各种攻击情况下,经不同基线方法防御的目标模型的性能.其中,“干净”表示没有攻击.此外,在标为“Vanilla”的一栏中还列出了在没有任何防御措施的情况下,仅使用Softmax损失函数训练的模型对应的结果.对于不引入随机性的基线方法,即Vanilla、NF-SAT、F-SAT[17]、DIO[21]、IGR[26]、DTDD[27]和UniG[28],基于梯度的攻击与其相应的EOT版本是等价的.因此,经这些基线方法防御的模型在基于梯度的攻击及其对应的EOT版本攻击条件下具有一致的性能.每行中的最优结果以粗体显示.

从表1和表2中可以看出,与所有基线方法相比,所提算法在干净样本和对抗样本上均获得了更高的准确率.如表1所示,当不使用EOT时,在最强白盒攻击下,与最有效的基线方法WCA相比,所提算法能将模型的准确率从55.78%提升至65.13%.在黑盒攻击One-Pixel和Square下,与最有效的基线方法WCA相比,所提算法能将模型的准确率分别从88.35%提升至93.98%、从86.98%提升至91.23%.当使用EOT时,在最强白盒攻击下,所提算法对应的准确率仍然比WCA高7.55%.如表2所示,当不使用EOT时,在最强白盒攻击下,与最有效的基线方法WCA相比,所提算法能将模型的准确率从52.92%提升至63.77%.在黑盒攻击One-Pixel和Square下,与最有效的基线方法WCA相比,所提算法能将模型的准确率分别从85.76%提升至91.71%、

从82.58%提升至88.86%. 当使用EOT时,在最强白盒攻击下,所提算法对应的准确率仍然比WCA高7.61%.

表1　在CFAD数据集上所提算法和基线方法的对比实验结果　　单位:%

	准确率										
	Vanilla	NF-SAT	F-SAT[17]	DIO[21]	IGR[26]	DTDD[27]	UniG[28]	PNI[41]	L2P[42]	WCA[43]	RADB-SFMR
干净	98.60	97.89	98.14	98.56	98.17	98.24	98.60	93.71	93.62	88.27	**99.71**
FGSM	45.64	62.35	62.46	58.53	57.21	60.67	46.57	46.38	55.36	70.23	**82.64**
BIM	20.33	39.51	40.03	26.60	25.14	25.29	22.22	21.86	41.67	56.03	**66.12**
PGD	17.31	38.49	38.88	26.36	24.76	25.17	18.23	23.03	40.24	55.78	**65.13**
Jitter	37.02	56.34	56.75	57.93	51.17	51.55	39.02	42.19	52.34	59.87	**81.60**
CW	6.02	9.19	9.78	21.21	24.36	32.71	6.47	17.71	20.35	40.76	**65.79**
EOT-FGSM	45.64	62.35	62.46	58.53	57.21	60.67	46.57	44.29	53.68	69.59	**78.43**
EOT-BIM	20.33	39.51	40.03	26.60	25.14	25.29	22.22	17.55	37.43	53.75	**63.78**
EOT-PGD	17.31	38.49	38.88	26.36	24.76	25.17	18.23	18.19	36.48	52.48	**60.03**
EOT-Jitter	37.02	56.34	56.75	57.93	51.17	51.55	39.02	32.83	48.37	58.48	**75.67**
One-Pixel	68.64	77.64	78.12	78.43	79.54	81.50	70.39	82.86	84.31	88.35	**93.98**
Square	66.31	75.86	76.13	76.25	76.68	79.56	82.62	80.51	81.56	86.98	**91.23**

表2　在ASVspoof2019数据集上所提算法和基线方法的对比实验结果　　单位:%

	准确率										
	Vanilla	NF-SAT	F-SAT[17]	DIO[21]	IGR[26]	DTDD[27]	UniG[28]	PNI[41]	L2P[42]	WCA[43]	RADB-SFMR
干净	98.38	97.97	98.12	98.36	97.86	97.98	98.38	95.90	94.96	89.49	**99.93**
FGSM	43.54	60.27	60.78	58.67	56.51	59.35	44.55	44.61	56.20	72.11	**82.72**
BIM	18.30	37.79	38.23	26.25	27.03	24.95	20.04	20.00	38.77	54.12	**68.71**
PGD	19.26	39.65	40.24	28.14	27.60	27.53	21.81	22.16	38.57	52.92	**63.77**
Jitter	38.55	57.22	57.69	57.82	52.59	52.62	40.12	43.93	50.16	60.18	**78.92**
CW	5.59	7.26	7.34	20.80	22.96	35.38	7.73	16.86	23.01	41.34	**64.72**
EOT-FGSM	43.54	60.27	60.78	58.67	56.51	59.35	44.55	41.52	52.72	67.65	**78.69**
EOT-BIM	18.30	37.79	38.23	26.25	27.03	24.95	20.04	17.04	36.64	51.62	**61.67**
EOT-PGD	19.26	39.65	40.24	28.14	27.60	27.53	21.81	19.39	37.37	50.77	**58.38**
EOT-Jitter	38.55	57.22	57.69	57.82	52.59	52.62	40.12	31.94	47.89	59.69	**77.74**
One-Pixel	69.42	77.98	78.19	81.15	81.59	80.05	70.65	80.63	84.77	85.76	**91.71**
Square	64.23	74.87	75.27	77.23	78.06	77.56	81.62	79.20	80.34	82.58	**88.86**

上述实验结果表明,所提算法在干净样本和对抗样本上均能够取得更好的性能表现,这可以归因于:

(1)所提算法中的RT模块通过模拟真实场景中的干扰(如混响、噪声叠加),对输入语音进行随机组合变换. 这种操作不仅增强了模型对对抗扰动的鲁棒性,还通过数据增强提升了模型在干净样本上的泛化能力.

(2)所提算法中的DBDR模块通过最大化对抗鲁棒性上限,鼓励模型学习判别性更强的决策边界. 这种优化使得模型在干净样本上的类内特征更紧凑、类间特征更分离,从而提升分类准确率. 与Vanilla方法相比,其他防御方法虽然增强了模型的防御能力,但是却削弱了模型在干净样本上的性能,这是因为WCA、PNI、L2P等基于噪声注入的防御方法向模型权重或特征中注入噪声,干扰了模型对干净样本的判别性特征提取,导致性能下降. 部分基于权重正则化的防御方法(如IGR、DTDD)通过约束梯度模长,鼓励模型学习到模长接近零的梯度,过度限制了模型的学习空间,削弱了模型对复杂特征的拟合能力. 基于模型集成的防御方法(DIO)需要额外的模块辅助识别,增加了模型的复杂度,在相同规模数据集上的过拟合风险更大,容易造成在干净样本上的性能下降. 基于对抗训练的防御方法(NF-SAT、F-SAT)会对模型在干净样本上的性能产生负面影响,造成准确率下降. 值得注意的是,虽然UniG也是基于权重正则化的防御方法,但是它通过约束模型输出特征的变化,尽量减小模型对干净样本预测逻辑值的影响. 因此,该方法几乎不改变模型在干净样本上的性能表现,取得了与Vanilla方法一致的准确率.

此外,图2展示了在CFAD数据集上,在不同扰动阈值下,经所提算法与各种基线方法防御的目标模型

在EOT-PGD攻击下的准确率曲线. 实验结果表明,随着扰动阈值的增加,即攻击强度的增加,各种方法对应的性能均呈现下降趋势. 同时,在不同扰动阈值下,所提算法对应的准确率均高于其他基线方法对应的准确率. 这进一步说明了所提算法的有效性.

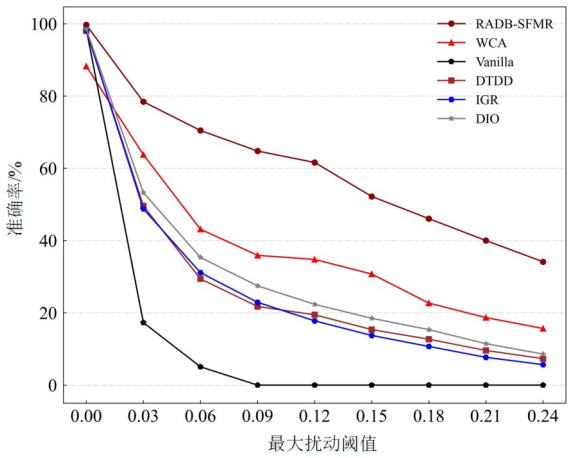


图2 各种防御方法在EOT-PGD攻击下的性能表现

为了直观地验证所提算法的有效性,图3显示了在CFAD数据集上,所提算法与各种基线方法学习到的对抗鲁棒性上限$\Omega_{\min}$随训练进程的变化曲线. 图中曲线上的每个点表示训练过程中的每个批次对应的$\Omega_{\min}$. 每个训练批次对应的$\Omega_{\min}$用这批样本对应的$\Omega_{\min}$的平均值来估计. 结果显示,所提算法学习到的$\Omega_{\min}$大于各种对比基线的相应结果. 这表明,与基线方法防御的模型相比,要想成功攻击所提算法防御的模型,攻击者需要更大的扰动强度. 因此,所提算法被攻击成功的可能性更小,这也是所提算法的对抗鲁棒性优于基线方法的原因.

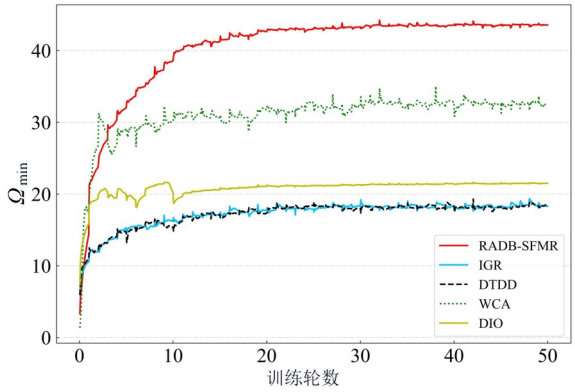
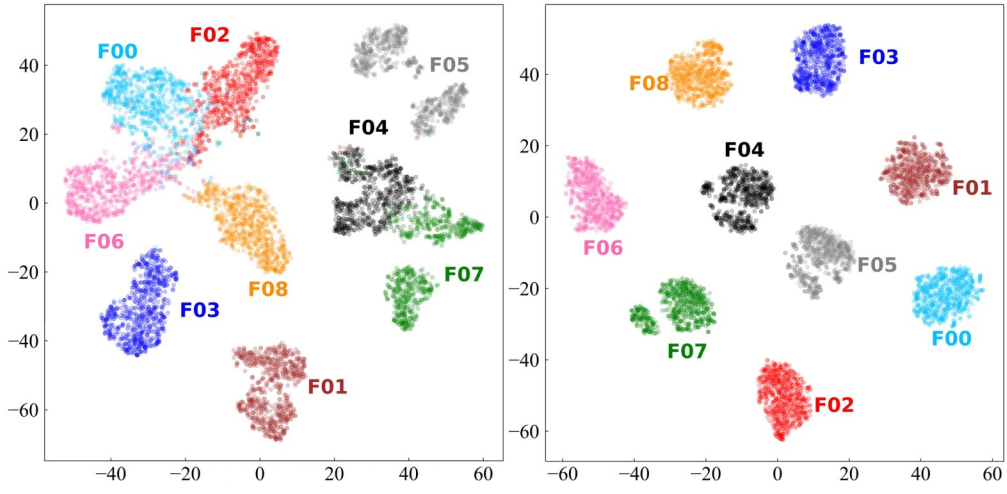


图3 各种方法学习到的$\Omega_{\min}$随训练进程的变化曲线

此外,图4展示了在CFAD数据集上,使用t分布随机邻域嵌入(t-distributed Stochastic Neighbor Embedding,t-SNE)[58]技术,对所提算法和最佳基线方法WCA学习到的特征表示所代表的语音样本的可视化结果. 从图4(*a*)中可以看出,用WCA学习到的特征表示来表示语音样本时,有些类别彼此接近(如F06和F08),甚至重叠(如F00和F02、F04和F07). 与WCA学习到的结果相比,所提算法大大提高了特征表示的性能,不同类别的语音样本可以很容易地在特征表示空间中区分开来. 这说明所提算法学习到的伪造方法的特征在类与类之间的区分度更高,在类内部更紧凑. 这意味着,与WCA相比,要混淆模型对所提算法学习到的特征表示的判断,将模型学习到的一个类别的特征表示推入另一个类别的聚类中,攻击者一般需要使用幅度更大的扰动,因此所提算法具有更强的对抗鲁棒性.



(*a*) WCA (*b*) RADB-SFMR

图4 所提算法和最佳基线方法学习到的特征表示的t-SNE可视化

5.3 消融实验结果

本节在CFAD数据集上,对所提算法各个模块进行了消融实验.

5.3.1 对所提算法各个模块的消融实验

表3显示了对所提算法的两个关键模块进行消融实验的结果,其中最优结果以加粗表示. 第一行表示不

使用RT模块和DBDR模块,仅使用Softmax损失函数训练模型,因此其结果等价于Vanilla方法的结果. 从前3行可以看出,RT模块和DBDR模块都有利于提高SFMR在对抗样本上的性能;并且,单独使用任何一个模块,所提算法的性能表现均优于表1中表现最佳的基线方法WCA的性能. 同时,后3行的结果表明,与使用单一模块相比,RT模块和DBDR模块的组合能获得更好的SFMR性能. 这表明,对输入进行随机组合变换和对模型进行决策边界距离正则化都有利于算法学习到鲁棒性更强的对抗防御边界,从而增强对抗鲁棒性.

表3　对所提算法的两个模块的消融实验结果　　单位:%

模块		准确率						
RT	DBDR	干净	EOT-FGSM	EOT-BIM	EOT-PGD	EOT-Jitter	CW	One-Pixel
×	×	98.60	45.64	20.33	17.31	37.02	6.02	68.64
√	×	99.29	75.81	58.76	57.46	75.21	55.48	92.37
×	√	99.33	70.30	55.73	54.70	67.16	53.16	90.81
√	√	**99.71**	**78.43**	**63.78**	**60.03**	**75.67**	**65.79**	**93.98**

注:√表示使用该模块,×表示不使用该模块.

5.3.2　对决策边界距离正则化模块的消融实验

表4显示了对DBDR模块使用的损失函数超参数进行消融实验的结果,其中最优结果以加粗表示. 可以看出,当$\zeta=100$、$\alpha=0.1$时,所提算法在无攻击和各种对抗样本攻击条件下都取得了最佳性能. 从前3行可以看出,在ζ不变时,在无攻击条件下,准确率随着α的增加而轻微下降,这是因为随着α的增加,模型更加注重通过决策边界距离正则化损失函数增加已经被正确分类样本对应的$\Omega_{\min}$,而忽视了使用Softmax损失函数对错误分类样本的优化. 在对抗攻击条件下,准确率随着α的增加,出现先增加后略微下降的情况,这是因为α从0.01增加至0.1时,正确分类样本对应的$\Omega_{\min}$逐步达到最优,模型的对抗鲁棒性得到进一步增强;而当α从0.1增加至1时,由于对正确分类样本对应的$\Omega_{\min}$的优化效果已达到饱和状态,这时模型在错误分类样本上的性能下降占据了主导,导致模型在对抗攻击条件下的准确率出现了轻微下降. 从倒数3行可以发现,在α不变的情况下,ζ越大,所提算法在无攻击和各种对抗攻击条件下,准确率均随着ζ的增加而逐步增加. 这是因为阈值参数ζ越大,模型学习到的$\Omega_{\min}$越大,即模型学习到的对抗鲁棒性上限就越大,因此模型在无攻击和各类攻击条件下的准确率就越高. 同时,当ζ从50增加至100时,准确率增加幅度较小,这说明对模型的优化达到饱和状态,继续增加ζ并不能带来明显的性能提升.

5.3.3　随机变换模块的消融实验

表5显示了对RT模块中的参数Q进行消融实验的结果. 可以看出,无论Q如何取值,所提算法在无攻击和各种攻击条件下均能取得较好的性能表现. 这是因为RT模块通过对输入进行随机组合变换,有利于增加模型的对抗鲁棒性上限. 同时,随着Q值的增加,模型在攻击条件下的准确率呈现上升趋势,但在无攻击条件下的准确率呈现下降趋势. 相比于$Q=3$,设置更大的Q,比如$Q=4$,模型在对抗攻击条件下的准确率提升并不明显,但在无攻击条件下的准确率下降较为明显. 这是因为设置更大的Q值,相当于对干净的伪造语音样本进行了更多次的随机变换,增加了模型从中学习伪造方法特征的难度,因此在无攻击条件下的准确率下降较为明显. 这也是在本文中设置$Q=3$的原因.

表4　决策边界距离正则化模块的消融实验结果

参数		准确率/%						
ζ	α	干净	EOT-FGSM	EOT-BIM	EOT-PGD	EOT-Jitter	CW	One-Pixel
100	1	99.17	77.74	63.75	59.14	74.69	65.09	93.03
100	0.01	**99.75**	65.89	61.45	58.54	73.92	65.25	93.35
100	0.1	99.71	**78.43**	**63.78**	**60.03**	**75.67**	**65.79**	**93.98**
50	0.1	99.65	77.81	63.16	59.49	75.53	65.51	93.25
10	0.1	99.52	69.92	62.27	58.21	69.64	63.38	92.28

表5　对随机变换模块的消融实验结果

Q	准确率/%						
	干净	EOT-FGSM	EOT-BIM	EOT-PGD	EOT-Jitter	CW	One-Pixel
1	99.89	70.56	62.19	58.60	69.80	50.27	79.70
3	99.71	78.43	63.78	60.03	75.67	65.79	93.98
4	88.84	78.55	64.13	60.25	75.72	66.06	84.24

5.4　梯度混淆现象分析

Athalye等人[59]提出了一系列条件,用于检查防御方法是否依赖于混淆梯度. 如果其中任何一个条件成立,则认为该防御方法不可靠. 下面的分析验证了所提算法带来的对抗鲁棒性增强不是由于梯度混淆造成的.

条件1. 一步攻击优于迭代攻击.

反驳. 从表1和表2可以看出,当攻击所提算法防御的模型时,在FGSM攻击下,模型的识别准确率更高,因此单步攻击方法FGSM的攻击性能明显低于其对应的迭代攻击版本,如PGD和BIM. 因此,这一条件不成立.

条件2. 黑盒攻击优于白盒攻击.

反驳. 从表1和表2可以看出,与黑盒攻击(One-Pixel)相比,在白盒攻击(如FGSM、BIM和PGD)下,模型的准确率更低. 因此,白盒攻击的攻击性能明显优于黑盒攻击. 因此,这一条件不成立.

条件3. 无界攻击条件下,模型的准确率依然大于0.

反驳. 当将EOT-PGD攻击的最大扰动幅度阈值设

置为1,迭代次数设置为100,蒙特卡罗采样次数设为30时,所提算法在CFAD和ASVspoof2019数据集上训练的模型对应的准确率均下降至0. 因此,该条件不成立.

条件4. 随机采样能够发现对抗样本.

反驳. 从CFAD数据集的评估集中随机抽取100个语音样本. 在没有攻击的情况下,经所提算法训练的模型可以正确地识别这些样本对应的伪造方法. 但在FGSM攻击下,这些语音样本都被误判,其中FGSM使用的扰动幅度阈值为$\varepsilon=0.03$. 对于每一个语音样本,本节从一个以该语音样本为球心、以0.03为半径的超球中随机抽取1 000个语音样本. 因此,本节创建了一个更大的、由随机采样得到的评估集,其包含10^5个语音样本. 在随机采样的评估集上,所提算法的准确率仍能达到98.55%. 因此,该条件不成立.

条件5. 增加扰动幅度阈值并不会降低模型的准确率.

反驳. 从图2可以看出,增加扰动幅度阈值会降低模型的准确率. 因此,该条件不成立.

上述实验结果说明,所提算法带来的对抗鲁棒性增强不是由于梯度混淆造成的.

5.5 噪声鲁棒性分析

本节在CFAD数据集的噪声评估子集上对所提算法与基线方法的噪声鲁棒性进行了比较分析. 该噪声评估子集是通过在CFAD数据集的干净评估子集的样本上添加不同信噪比的噪声音频获得的. 噪声音频来自PNL100非语音声音数据集[60]、NOISEX-92数据集[61]和城市声学场景数据集[62]共3个不同的噪声数据集,其中包含环境噪声、嘈杂人声等类型的噪声. 值得注意的是,这些噪声与所提算法对输入添加的噪声来自不同的数据集. 添加噪声的信噪比包括0、5、10、15和20 dB共5种. 这5种信噪比对应的评估样本数量分别为5 249、5 782、5 780、5 784、8 905. 实验结果如表6所示,其中最优结果以加粗表示. 所有模型都来自表1,它们都是在不带噪声的训练集上训练的. 可以看出,在噪声条件下,所提算法依然能够取得最高的准确率. 与所有基线方法相比,所提算法带来的准确率提升均在30%左右. 这表明所提算法不仅能有效增强对抗鲁棒性,还能有效增强噪声鲁棒性. 这是因为所提算法采用的RT模块是模拟各种声学环境构建的,比如混响、环境噪声、嘈杂人声、音乐声,因此,经所提算法训练的模型在带噪声语音样本上也表现出了稳健的性能.

表6 不同方法训练的模型在噪声评估集上的性能对比

信噪比	准确率/%						
	Vanilla	NF-SAT	DIO[21]	IGR[26]	DTDD[27]	WCA[43]	RADB-SFMR
0	54.84	61.14	58.79	56.61	56.58	51.87	**86.69**
5	57.34	63.75	61.54	59.53	59.71	55.15	**89.98**
10	60.15	66.60	64.47	62.42	62.51	57.99	**92.92**
15	62.23	68.71	66.58	64.53	64.72	60.21	**95.14**
20	65.80	72.21	70.01	67.94	68.05	63.48	**98.36**
平均	60.73	67.14	64.94	62.87	62.98	58.41	**93.29**

6 结论

本文针对经对抗样本伪装的伪造语音的伪造方法识别难题,开展了AR-SFMR任务建模. 通过使用泰勒分析技术,深入研究了影响SFMR模型对抗鲁棒性的因素,提出了对抗防御边界概念. 该概念为解释基于数据增强或模型增强的防御策略提供了统一视角. 在此基础上,进一步提出了一种基于鲁棒对抗防御边界的SFMR算法,通过对输入的随机变换和对输出的决策边界距离正则化,增强了输入的随机性,降低了模型输出的类别逻辑值对输入的敏感性,从而增强了模型的对抗鲁棒性. 在典型语音伪造方法识别数据集CFAD、ASVspoof2019上的实验结果表明,所提算法在AR-SFMR任务上的性能优于先进的基线方法,并且能够显著提升模型的噪声鲁棒性. 下一步,我们将在更多的语音伪造方法识别数据集和更多的目标模型上,探索并提高所提方法的泛化性.

参考文献

[1] MASOOD M, NAWAZ M, MALIK K M, et al. Deepfakes generation and detection: State-of-the-art, open challenges, countermeasures, and way forward[J]. Applied Intelligence, 2023, 53(4): 3974-4026.

[2] 李旭嵘, 纪守领, 吴春明, 等. 深度伪造与检测技术综述[J]. 软件学报, 2021, 32(2): 496-518.
LI X R, JI S L, WU C M, et al. Survey on deepfakes and detection techniques[J]. Journal of Software, 2021, 32(2): 496-518. (in Chinese)

[3] MÜLLER N, DIEKMANN F, WILLIAMS J. Attacker attribution of audio deepfakes[C]//Interspeech 2022. Singapore: ISCA, 2022: 2788-2792.

[4] NERI M, FERRAROTTI A, DE LUISA L, et al. ParalMGC: Multiple audio representations for synthetic human speech attribution[C]//2022 10th European Workshop on Visual Information Processing (EUVIP). Piscataway: IEEE, 2022: 1-6.

[5] DENG J L, REN Y Z, ZHANG T, et al. VFD-net: Vocoder fingerprints detection for fake audio[C]//2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE, 2024: 12151-12155.

[6] XIE Y K, LU Y, FU R B, et al. The codecfake dataset and

countermeasures for the universally detection of deepfake audio[J]. IEEE Transactions on Audio, Speech and Language Processing, 2025, 33: 386-400.

[7] 田野, 罗曦, 许斌, 等. 一种基于预训练模型的语音深度伪造算法识别方法[J]. 电声技术, 2024, 48(2): 28-31, 35.
TIAN Y, LUO X, XU B, et al. A pre-trained model based recognition method for speech deepfake algorithms[J]. Audio Engineering, 2024, 48(2): 28-31, 35. (in Chinese)

[8] ZHU T L, WANG X M, QIN X Y, et al. Source tracing: Detecting voice spoofing[C]//2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC). Piscataway: IEEE, 2022: 216-220.

[9] SALVI D, BESTAGINI P, TUBARO S. Exploring the synthetic speech attribution problem through data-driven detectors[C]//2022 IEEE International Workshop on Information Forensics and Security (WIFS). Piscataway: IEEE, 2022: 1-6.

[10] YAN X R, YI J Y, TAO J H, et al. An initial investigation for detecting vocoder fingerprints of fake audio[C]// Proceedings of the 1st International Workshop on Deepfake Detection for Audio Multimedia. New York: ACM, 2022: 61-68.

[11] YADAV A K S, BARTUSIAK E R, BHAGTANI K, et al. Synthetic speech attribution using self supervised audio spectrogram transformer[J]. Electronic Imaging, 2023, 35(4): 1-11.

[12] BARTUSIAK E R, DELP E J. Transformer-based speech synthesizer attribution in an open set scenario[C]//2022 21st IEEE International Conference on Machine Learning and Applications (ICMLA). Piscataway: IEEE, 2022: 329-336.

[13] KLEIN N, CHEN T X, TAK H, et al. Source tracing of audio deepfake systems[C]//Interspeech 2024. Singapore: ISCA, 2024: 1-5.

[14] QIAN Z, HUANG K Z, WANG Q F, et al. A survey of robust adversarial training in pattern recognition: Fundamental, theory, and methodologies[J]. Pattern Recognition, 2022, 131: 108889.

[15] 刘延华, 李嘉琪, 欧振贵, 等. 对抗训练驱动的恶意代码检测增强方法[J]. 通信学报, 2022, 43(9): 169-180.
LIU Y H, LI J Q, OU Z G, et al. Adversarial training driven malicious code detection enhancement method[J]. Journal on Communications, 2022, 43(9): 169-180. (in Chinese)

[16] 曹刘娟, 匡华峰, 刘弘, 等. 双标签监督的几何约束对抗训练[J]. 软件学报, 2022, 33(4): 1218-1230.
CAO L J, KUANG H F, LIU H, et al. Towards robust adversarial training via dual-label supervised and geometry constraint[J]. Journal of Software, 2022, 33(4): 1218-1230. (in Chinese)

[17] ZHANG Z R, HAO W, SANKOH A, et al. I can hear you: Selective robust training for deepfake audio detection[EB/OL]. (2024-10-31)[2025-05-12]. https://arxiv.org/abs/2411.00121v1.

[18] XIE C, ZHANG Z, YUILLE A L, et al. Mitigating adversarial effects through randomization[C]//International Conference on Learning Representations. Washington DC: ICLR, 2018: 10-25.

[19] ZHANG Y C, LIANG P. Defending against whitebox adversarial attacks via randomized discretization[EB/OL]. (2019-05-25)[2025-05-12]. https://arxiv.org/abs/1903.10586v1.

[20] RAFF E, SYLVESTER J, FORSYTH S, et al. Barrage of random transforms for adversarially robust defense[C]// 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2019: 6521-6530.

[21] FANG K, TAO Q H, WU Y W, et al. Towards robust neural networks via orthogonal diversity[J]. Pattern Recognition, 2024, 149: 110281.

[22] 张世辉, 张晓微, 宋丹丹, 等. 基于逆扰动融合生成对抗网络的对抗样本防御方法[J]. 电子学报, 2023, 51(4): 879-884.
ZHANG S H, ZHANG X W, SONG D D, et al. Adversarial example defense method based on inverse perturbation fusing generative adversarial network[J]. Acta Electronica Sinica, 2023, 51(4): 879-884. (in Chinese)

[23] MENG H, OU W, HUANG J, et al. A robust unified spoofing audio detection scheme[J]. Computers and Electrical Engineering, 2025, 122: 109974.

[24] LI Y Z, ZHANG C, QI H G, et al. AdaNI: Adaptive noise injection to improve adversarial robustness[J]. Computer Vision and Image Understanding, 2024, 238: 103855.

[25] LIU Z Y, GAGNON G, VENKATARAMANI S, et al. Enhance DNN adversarial robustness and efficiency via injecting noise to non-essential neurons[EB/OL]. (2024-12-06)[2025-05-12]. https://arxiv.org/abs/2402.04325.

[26] ROSS A, DOSHI-VELEZ F. Improving the adversarial robustness and interpretability of deep neural networks by regularizing their input gradients[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2018, 32(1): 1660-1669.

[27] BENDER C M, LI Y, SHI Y F, et al. Defense through

diverse directions[EB/OL]. (2020-05-24) [2025-05-12]. https://arxiv.org/abs/2003.10602v1.

[28] WU Y W, CHEN S Z, FANG K, et al. Unifying gradients to improve real-world robustness for deep networks[J]. ACM Transactions on Intelligent Systems and Technology, 2023, 14(6): 1-16.

[29] LIU X C, WANG X, SAHIDULLAH M, et al. ASVspoof 2021: Towards spoofed and deepfake speech detection in the wild[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2023, 31: 2507-2522.

[30] YI J, TAO J, FU R, et al. ADD 2023: The second audio deepfake detection challenge[C]//International Joint Conference on Artificial Intelligence. Freiburg: IJCAI, 2023: 125-130.

[31] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2016: 770-778.

[32] KHOSLA P, TETERWAK P, WANG C, et al. Supervised contrastive learning[J]. Advances in Neural Information Processing Systems, 2020, 33: 18661-18667.

[33] GOODFELLOW I J, SHLENS J, SZEGEDY C. Explaining and harnessing adversarial examples[C]//International Conference on Learning Representations. San Diego: ICLR, 2015: 10-21.

[34] MADRY A, MAKELOV A, SCHMIDT L, et al. Towards deep learning models resistant to adversarial attacks[C]// International Conference on Learning Representations. San Diego: ICLR, 2018: 10-37.

[35] KURAKIN A, GOODFELLOW I J, BENGIO S. Adversarial Examples in the Physical World[M]//Artificial Intelligence Safety and Security. First edition. | Boca Raton, FL: CRC Press/Taylor & Francis Group, 2018: Chapman and Hall/CRC, 2018: 99-112.

[36] SCHWINN L, RAAB R, NGUYEN A, et al. Exploring misclassifications of robust neural networks to enhance adversarial attacks[J]. Applied Intelligence, 2023, 53(17): 19843-19859.

[37] ATHALYE A, ENGSTROM L, ILYAS A, et al. Synthesizing robust adversarial examples[C]//International Conference on Machine Learning. New York: ICML, 2018: 449-468.

[38] SITAWARIN C, GOLAN-STRIEB Z, WAGNER D. Demystifying the adversarial robustness of random transformation defenses[EB/OL]. (2020-07-15) [2025-05-12]. https://arxiv.org/abs/2207.03574v2.

[39] XING Y, SONG Q, CHENG G. On the generalization properties of adversarial training[J]. Proceedings of Machine Learning Research, 2021, 130: 505-513.

[40] ZHANG H, CHEN H G, SONG Z, et al. The limitations of adversarial training and the blind-spot attack[EB/OL]. (2019-01-15)[2025-05-12]. https://arxiv.org/abs/1901.04684v1.

[41] HE Z Z, RAKIN A S, FAN D L. Parametric noise injection: Trainable randomness to improve deep neural network robustness against adversarial attack[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2019: 588-597.

[42] JEDDI A, SHAFIEE M J, KARG M, et al. Learn2Perturb: An end-to-end feature perturbation learning to improve adversarial robustness[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2020: 1238-1247.

[43] EUSTRATIADIS P, GOUK H, LI D, et al. Weight-covariance alignment for adversarially robust neural networks[C]//International Conference on Machine Learning. New York: ICML, 2021: 3047-3056.

[44] LEE S, KIM H, LEE J. GradDiv: Adversarial robustness of randomized neural networks via gradient diversity regularization[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(2): 2645-2651.

[45] KESAVAN H K, KAPUR J N. The generalized maximum entropy principle[J]. IEEE Transactions on Systems, Man & Cybernetics, 1989, 19(5): 1042-1052.

[46] KO T, PEDDINTI V, POVEY D, et al. A study on data augmentation of reverberant speech for robust speech recognition[C]//2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE, 2017: 5220-5224.

[47] SNYDER D, CHEN G G, POVEY D. MUSAN: A music, speech, and noise corpus[EB/OL]. (2015-10-28)[2025-05-12]. https://arxiv.org/abs/1510.08484.

[48] MA H X, YI J Y, WANG C L, et al. CFAD: A Chinese dataset for fake audio detection[J]. Speech Communication, 2024, 164: 103122.

[49] WANG X, YAMAGISHI J, TODISCO M, et al. ASVspoof2019: A large-scale public database of synthesized, converted and replayed speech[J]. Computer Speech & Language, 2020, 64: 101114.

[50] SHEN J, PANG R M, WEISS R J, et al. Natural TTS synthesis by conditioning wavenet on MEL spectrogram predictions[C]//2018 IEEE International Conference on

Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE, 2018: 4779-4783.
[51] VAN DEN OORD A, DIELEMAN S, ZEN H G, et al. WaveNet: A generative model for raw audio[EB/OL]. (2016-09-19)[2025-05-12]. https://arxiv.org/abs/1609.03499v2.
[52] DENG J, DONG W, SOCHER R, et al. ImageNet: A large-scale hierarchical image database[C]//2009 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2009: 248-255.
[53] SZEGEDY C, VANHOUCKE V, IOFFE S, et al. Rethinking the inception architecture for computer vision[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2016: 2818-2826.
[54] KINGMA D P, BA J L. Adam: A method for stochastic optimization[C]//3rd International Conference on Learning Representations. Vienna: ICLR, 2015: 1-15.
[55] CARLINI N, WAGNER D. Towards evaluating the robustness of neural networks[C]//2017 IEEE Symposium on Security and Privacy (SP). Piscataway: IEEE, 2017: 39-57.
[56] SU J W, VARGAS D V, SAKURAI K. One pixel attack for fooling deep neural networks[J]. IEEE Transactions on Evolutionary Computation, 2019, 23(5): 828-841.
[57] ANDRIUSHCHENKO M, CROCE F, FLAMMARION N, et al. Square Attack: A query-efficient black-box adversarial attack via random search[EB/OL]. (2019-11-29)[2025-05-12]. https://arxiv.org/abs/1912.00049.
[58] VAN DER MAATEN L, HINTON G. Visualizing data using t-SNE[J]. Journal of Machine Learning Research, 2008, 9(11): 2579-2605.
[59] ATHALYE A, CARLINI N, WAGNER D. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples[C]//International Conference on Machine Learning. New York: ICML, 2018: 274-283.
[60] HU G N, WANG D L. A tandem algorithm for pitch estimation and voiced speech segregation[J]. IEEE Transactions on Audio, Speech, and Language Processing, 2010, 18(8): 2067-2079.
[61] VARGA A, STEENEKEN H J M. Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems[J]. Speech Communication, 1993, 12(3): 247-251.
[62] MESAROS A, HEITTOLA T, VIRTANEN T. A multi-device dataset for urban acoustic scene classification[EB/OL]. (2018-10-11)[2025-05-12]. https://arxiv.org/abs/1807.09840v2.

作者简介

张　强　男,1991年2月出生于江苏省宿迁市. 2024年获得陆军工程大学博士学位. 现为陆军工程大学博士后研究员. 主要研究方向为机器学习、模式识别、语音信号处理.
E-mail: zq308297543@126.com

张雄伟　男,1965年11月出生于浙江省嘉兴市. 1992年获得南京解放军通信工程学院通信与电子系统学科博士学位. 现为陆军工程大学教授. 主要研究方向为语音信号处理、智能信息处理、模式识别.
E-mail: xwzhang9898@163.com

孙　蒙　男,1984年12月出生于山东省德州市. 2012年获得鲁汶大学电子工程系博士学位. 现为陆军工程大学教授. 主要研究方向为语音处理、无监督/半监督机器学习、序列模式识别.
E-mail: sunmeng@aeu.edu.cn

杨吉斌　男,1978年8月出生于安徽省明光市. 2006年获得中国人民解放军理工大学博士学位. 现为陆军工程大学副教授. 主要研究方向为语音信号处理、机器学习、模式识别.
E-mail: yangjibin@aeu.edu.cn